Computational Social Sciences

Shu-Heng Chen *Editor*

# Big Data in Computational Social Science and Humanities

Springer

# Computational Social Sciences

# Computational Social Sciences

A series of authored and edited monographs that utilize quantitative and computational methods to model, analyze and interpret large-scale social phenomena. Titles within the series contain methods and practices that test and develop theories of complex social processes through bottom-up modeling of social interactions. Of particular interest is the study of the co-evolution of modern communication technology and social behavior and norms, in connection with emerging issues such as trust, risk, security and privacy in novel socio-technical environments.

Computational Social Sciences is explicitly transdisciplinary: quantitative methods from fields such as dynamical systems, artificial intelligence, network theory, agentbased modeling, and statistical mechanics are invoked and combined with state-of-theart mining and analysis of large data sets to help us understand social agents, their interactions on and offline, and the effect of these interactions at the macro level. Topics include, but are not limited to social networks and media, dynamics of opinions, cultures and conflicts, socio-technical co-evolution and social psychology. Computational Social Sciences will also publish monographs and selected edited contributions from specialized conferences and workshops specifically aimed at communicating new findings to a large transdisciplinary audience. A fundamental goal of the series is to provide a single forum within which commonalities and differences in the workings of this field may be discerned, hence leading to deeper insight and understanding.

Shu-Heng Chen
Editor

# Big Data in Computational Social Science and Humanities

Springer

*Editor*
Shu-Heng Chen
AI-ECON Research Center
Department of Economics
National Chengchi University
Taipei, Taiwan

ISSN 2509-9574 ISSN 2509-9582 (electronic)
Computational Social Sciences
ISBN 978-3-319-95464-6 ISBN 978-3-319-95465-3 (eBook)
https://doi.org/10.1007/978-3-319-95465-3

Library of Congress Control Number: 2018956726



This Springer imprint is published by the registered company Springer Nature Switzerland AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

# Preface

Computational social science was once known as the computer simulation of social phenomena. This perception is very clear from the first book entitled *Computational Social Science* that was edited by Nigel Gilbert in 2010 in his 4-volume collections of 66 articles published from the years 1963 to 2008, a range of almost half a century. The idea of blending social sciences with computer simulation is, therefore, both new and anticipated; as Robert Axelrod has remarked, social interaction by its very nature is highly computational. It involves multi-threading and parallel information retrieval, processing, decision-making, and information spreading in evolving social networks. By tracing how information is generated, used, and spread, various disciplines in the social sciences can find their deep-grounded connections. Computational social science, therefore, provides social scientists with a platform to get overarched or integrated.

Big data, nonetheless, was not mentioned in any of those 66 articles; obviously, this neologism did not exist in 2008. However, big data was mentioned 11 times in the 320-page book, *Introduction to Computational Social Science: Principles and Applications*, authored by Claudio Cioffi-Revilla in 2014. Hence, within such a short span of 5 years, i.e., 2009–2014, big data has already become an essential part of computational social science (CSS).

Over this very short history, what big data adds to the 50-year-old CSS is twofold. First, as already mentioned, CSS can be viewed as the computer simulation of various domains of social interactions, from the individual (micro) level to the aggregate (macro) level. Before the era of big data, not all the details of the corresponding social interactions were available; effectively speaking, many of them were simply not archivable. With the advancement of ICT technology, Web 2.0, social media, ubiquitous computing, wearable devices, and the Internet of Everything, the advent of the big data era makes these details become increasingly available. These data availability results have a substantial impact on the evolution of CSS: it enables one to calibrate, validate, and test simulated social interactions with the ultrahigh-frequency micro details. Not only can we see and simulate the ducks swimming on the water, but we can also see and simulate their feet paddling

underwater. Taking financial markets as an example, it is not just the dynamics of stock prices, or the decisions of myriads of traders, but we are now also endowed with the opportunity to advance further into traders' decision-making processes. In this sense, big data consolidates the micro-macro links, as constantly pursued by CSS.

Second, and probably more remarkably, is that big data availability facilitates the dialogues and cooperation between social scientists and humanists. This is so because big data brings in a new "geometry" of data, in the form of texts, images, audios, and videos, which has made narratives, the essence of the humanities, a rather substantial or an indispensable part of the social sciences. Needless to say, social scientists and humanists share some common interests: human nature and their social embeddedness. Classical economics is filled with such kinds of writings, namely, Adam Smith's *Wealth of Nations* (1776) and Karl Marx's *Das Kapital* (1867, Vol. 1), to name just two. This narrative style was gradually disappearing when economics became increasingly "pure" (mathematized), from Leon Walras's *Elements of Pure Economics* (1874), to John von Neumann and Oskar Morgenstern's *Theory of Games and Economic Behavior* (1944), further to Gérard Debreu's *Theory of Value* (1959). However, to get immersed in the deplorable conditions of workers under an industrial capitalist society, one can probably learn more from Charles Dickens, say, in his *Great Expectations*, than from axiomatic or mathematical analysis alone. Similarly, just ask from whom one can learn more about human nature: is it Sigmund Freud or Leo Tolstoy?

The new kind (geometry) of data in the era of big data also promotes the use of a number of data analytics, such as text mining, corpus linguistics, sentiment analysis, social network analysis, geographic information systems, and co-word network analysis, which have now been used by both social scientists and humanists. The sharing of these toolkits further enhances the dialogues and cooperation between the social sciences and humanities. This in turn narrows the gap between CSS and the humanities. For example, Leo Tolstoy's magnum opus, *War and Peace* (1969), has "simulated" more than 500 actors in some fine detail in his "model." Can a machine or CSS do this, or is a human writer armed with CSS able to come up with something closer? We do not know, but the issue itself motivates us and is the vision behind this book.

This book is unique in the sense that it treats big data as a key driver to actively engage social scientists and humanists together in order to at least prepare their dialogues in the future. In this vein, the book can be related to the recent book *Cents and Sensibility* (2017), authored by Gary Saul Morson and Morton Schapiro. They considered that social scientists can learn from the humanities in their inherent wisdom. Throughout their book, Morson and Schapiro have employed Isaiah Berlin's famous caricature "Hedgehogs and Foxes" to shed light on the difference between the social sciences and humanities. We appreciate their viewpoints. As for us, we believe that computational social scientists can benefit greatly from their conversations with humanists, but we also believe that such conversations can be much facilitated if the humanities can also be studied in a computational format.

The latter is well illustrated by what Franco Moretti has demonstrated in his *Graphs, Maps, Trees: Abstract Models for a Literary History* (2005), which in effect coined the term "computational criticism."

In 2013, National Chengchi University (NCCU) initiated a research circle, known as the digital humanities consortium. The constituent faculty members are from both the social sciences and the humanities and are both local and international. After 2 years of conversations, we found that it would be desirable to have a special edition to provide an overview of the current state of big data in the social sciences and humanities so that our ongoing dialogues can be landmarked and extended. In this intended project, big data is the common language. While allowing for different dialects, in a very similar way to there being different branches of a mighty river, this river which is depicted by big data then traverses through a great landscape.

At the 2015 Conference on Complex Systems, held at Arizona State University, Phoenix, the editor of this book had the chance to meet the Springer editor, Christopher Coughlin. At that time, Mr. Coughlin was promoting the Springer series on the computational social sciences. As the result of his kind invitation, encouragement, and subsequent assistance, we can finally put our aforementioned vision into action. In addition to Mr. Coughlin, our gratitude is also extended to Jeffrey Taub, the project coordinator, who has helped us with various copyediting details and logistics. Finally, the support from National Chengchi University's Digital Humanities Project, which is in turn sponsored by the Ministry of Education, Taiwan, under the "Top Universities Project," is also highly appreciated.

Hope that with all these internal efforts and external supports, we are ready to begin a new page in the dialogue between social sciences and humanities.

Taipei, Taiwan
May 20, 2018

Shu-Heng Chen

# Contents

# Contributors

**David Blundell** Asia-Pacific SpatioTemporal Institute (ApSTi), National Chengchi University, Taipei, Taiwan

Electronic Cultural Atlas Initiative (ECAI), University of California, Berkeley, USA

**Romain Boulet** Université de Lyon, Jean Moulin, Institut d'Administration des Entreprises de Lyon, Centre de Recherche Magellan, Lyon, France

**Pang-Ru Chang** Department of Risk Management and Insurance, Shih-Chien University, Taipei, Taiwan

**Don-Yun Chen** Department of Public Administration & Taiwan E-Governance Research Center, National Chengchi University, Taipei, Taiwan

**Kung Chen** Department of Computer Science, National Chengchi University, Taipei, Taiwan

**Pai-Lin Chen** College of Communication, National Chengchi University, Taipei, Taiwan

**Shu-Heng Chen** AI-ECON Research Center, Department of Economics, National Chengchi University, Taipei, Taiwan

**Yu-Chung Cheng** Hsuan Chuang University, Hsinchu, Taiwan

**Siew Ann Cheong** School of Physical and Mathematical Sciences, Nanyang Technological University, Singapore, Republic of Singapore

**Ming-Te Chi** Department of Computer Science, National Chengchi University, Taipei, Taiwan

**Yao-Min Chiang** Department of Finance, National Taiwan University, Taipei, Taiwan

**Kawai Chui** Department of English, National Chengchi University, Taipei, Taiwan

**Siaw-Fong Chung** Department of English, National Chengchi University, Taipei, Taiwan

**Carol R. Ember** Human Relations Area Files at Yale University, New Haven, CT, USA

**Michael D. Fischer** University of Kent, Canterbury, UK

Human Relations Area Files at Yale University, New Haven, CT, USA

**Michael J. Gallagher** Department of Finance, St. Bonaventure University, St. Bonaventure, NY, USA

**Ann Heylen** National Taiwan Normal University, Taipei, Taiwan

**Naiyi Hsiao** Department of Public Administration & Taiwan E-Governance Research Center, National Chengchi University, Taipei, Taiwan

**Kuo-Wei Hsu** Department of Computer Science, National Chengchi University, Taipei, Taiwan

**Jihn-Fa Jan** Department of Land Economics, Taiwan Institute for Governance and Communication Research, National Chengchi University, Taipei, Taiwan

**Brian Kokensparger** Journalism, Media & Computing Department, Creighton University, Omaha, NE, USA

**Huei-Ling Lai** Department of English, National Chengchi University, Taipei, Taiwan

**Claire Lajaunie** INSERM/CERIC, UMR DICE 7318, CNRS et Aix-Marseille Université, Aix-en-Provence Cedex 1, France

**Zhoupeng Liao** Department of Public Administration & Taiwan E-Governance Research Center, National Open University, New Taipei City, Taiwan

**Ching-Chih Lin** Graduate Institute of Religious Studies, National Chengchi University, Taipei, Taiwan

**I-Ying Lin** College of Communication, National Chengchi University, Taipei, Taiwan

**Chao-Lin Liu** Department of Computer Science, National Chengchi University, Taipei, Taiwan

**Hui-Wen Liu** College of Communication, National Chengchi University, Taipei, Taiwan

**Pierre Mazzega** GET Géosciences Environnement Toulouse UMR5563, CNRS/IRD/Université de Toulouse, Toulouse, France

**James X. Morris** International Doctoral Program in Asia Pacific Studies, National Chengchi University, Taipei, Taiwan

**Andrea Nanetti** School of Art, Design and Media, Nanyang Technological University, Singapore, Republic of Singapore

**Wen-Hui Sah** Department of English, National Chengchi University, Taipei, Taiwan

**Jia-Lang Seng** Department of Accounting, National Chengchi University, Taipei, Taiwan

**Dehua Shen** College of Management and Economics, Tianjin University, Tianjin, China

**Oliver Streiter** National University of Kaohsiung, Kaohsiung City, Taiwan

**Tzu-Chieh Tsai** Department of Computer Science, National Chengchi University, Taipei, Taiwan

**Chia-Rong Tsao** Department of Social Psychology, Shih Hsin University, Taipei, Taiwan

**Feng-Shang Wu** Graduate Institute of Technology, Innovation and Intellectual Property Management, National Chengchi University, Taipei, Taiwan

**Lee-Xieng Yang** Department of Psychology, National Chengchi University, Taipei, Taiwan

**Yung-Shen Yen** Department of Computer Science and Information Management, Providence University, Taichung, Taiwan

**Tina Yu** AI-ECON Research Center, Department of Economics, National Chengchi University, Taipei, Taiwan

**Yawen Zou** Center for Biology and Society, Arizona State University, Tempe, AZ, USA

The Chinese University of Hong Kong, Shenzhen, China

# Chapter 1
# Big Data in Computational Social Sciences and Humanities: An Introduction


**Shu-Heng Chen and Tina Yu**

With the advance of information and digital technologies, specifically in relation to Web 2.0, ubiquitous computing, wearable devices, social media, and the Internet of things, a massive amount of information has been generated in the modern digital society. While the "big data gold rush" is under way in the business world (Peters 2012), government and social scientists are also interested in reaping the economic and scientific benefits from harnessing the power of big data. For example, the ESRC's Big Data Network in the UK and the US NSF's Big Data Research and Development Initiative are pouring fortunes into innovative projects to develop new methods and tools for capturing, managing, and exploiting enormous volumes of information. These major initiatives indicate that the governance of big data is essential for the advancement of human knowledge to accelerate economic growth and to provide a better quality of life for the people.

This edited book is about the potentials and challenges of big data for computational social sciences and the humanities.[1] It is prepared not only for computational

[1]Computational social sciences, as the title of this book series demonstrates, require little explanation. The term, computational humanities, however, is less popular. Gerhard Heyer distinguishes digital humanities from computational humanities as follows. The former is the creation, dissemination, and use of digital repositories, and the latter is the computer-based analysis of digital repositories using advanced computational and algorithmic methods (Biemann et al. 2014). Alternatively, "[c]omputational humanities is an emerging field that bridges the sciences and humanities with the goal of creating accurate computer simulations of historical, social, cultural, and religious events (Cruz-Neira 2003, p. 10)." See Gavin (2014) for a demonstration of the above two descriptions of computational humanities.

S.-H. Chen (✉) · T. Yu
AI-ECON Research Center, Department of Economics, National Chengchi University, Taipei, Taiwan



S.-H. Chen (ed.), *Big Data in Computational Social Science and Humanities*, Computational Social Sciences, https://doi.org/10.1007/978-3-319-95465-3_1

social scientists and humanists who are interested in using big data in their areas of research, but also for big data technologists who are interested in the governance of big data for social sciences- and humanities-related projects.

Big data means different things to different people. Regardless of the sources of the digital data, such as books, social media, databases, audio, and video, big data exhibit the characteristics of high-volume, high-velocity (speed of data in and out), and high variety (range of data types and sources). This new type of data enriches research prospects and has potential to advance research in the social sciences and humanities in the following ways:

- Advanced big data collection tools, such as web scraping, and innovative analytic techniques, such as machine learning, may help establish new research methodologies;
- New types of data may reveal new patterns and insights into human society, politics, and economics;
- New types of data may lead to new kinds of research questions that are beyond the perspectives of the established theories.

It is no wonder that many social scientists and humanists are turning to big data in their research.

With its great potential to revolutionize social science research, to what extent would big data inevitably replace more costly and time-consuming traditional methods of gathering information, e.g., surveys or in-depth interviews (Savage and Burrows 2007)? In addition, to what extent would big data fundamentally challenge traditional scientific practices? As a consequence, will the current standard research paradigms in the social sciences shift toward a new one? As one may learn from this volume, we consider that the change or the shift will gradually happen, but before that some words of caution have to be given. For example, Kleiner et al. (2015) believe that at this time big data should supplement, but not replace, traditional methods and data sources in the social sciences. "An important principle within the social sciences is that a study design should be optimal in producing valid and reliable data that fit and address the research questions of interest. Big data are usually not generated following a design intended to address specific research questions, and so generally do not easily lend themselves to use by social scientists" (Kleiner et al. 2015, p. 24). Hence, a related point is that the soundness of big data can never be established unless we can have a scientifically sound theory (model) for the generation process of big data (Chen and Venkatachalam 2017).

To promote the appropriate usage of big data in a scientific context, it would be beneficial to establish a research environment where the data utility, quality, and their accessibility are governed. The environment facilitates the sharing of big data best practices, such as suitable analytical strategies for a particular data type, and the early identification of error data sources. Furthermore, the ethical issues of using personal data can be addressed there.

In this volume, we are thrilled to have many distinguished scholars from diverse disciplines who have contributed to this book. The 18 chapters collected in this volume are organized into two parts, to be further elaborated in the first and second sections of this introductory chapter. In the first part of the book (Sect. 1.1), we present works that incorporate three major different kinds of big data, namely *geographic data*, *text corpus data*, and *social media data*, to conduct research on the social sciences in a wide range of fields, including anthropology, economics, finance, psychology, history, linguistics, political science, and mass communications. In the second part of the book (Sect. 1.2), we include two types of contribution: surveys that review published papers using big data to conduct research in various social science disciplines (Sect. 1.2.1), and articles that discuss challenges of using big data in social science research (Sect. 1.2.2). This is a book with many fascinating works. We hope that readers can enjoy it as much as we do.

## 1.1 Big Data for Computational Social Sciences and Humanities in Practice

There are 11 chapters that together constitute the first part of the book. These 10 chapters (Chaps. 2, 3, 4, 5, 6, 7, 8, 9, 10, and 11) are further classified into three groups based on *the type of big data* employed. The three types of big data reviewed in this section are geographic data (Sect. 1.1.1), text corpus data (Sect. 1.1.2), and social media data (Sect. 1.1.3). While this is not an exhaustive list, these three give us the three most frequently seen types of big data used in the social sciences and humanities.

### *1.1.1 Geographic Data*

Geographic data constitute information that has an implicit or explicit association with a location relative to the Earth. The data can be captured in many different ways, such as satellite remote sensing. Chapter 2, "Application of Citizen Science and Volunteered Geographic Information (VGI): Tourism Development for Rural Communities," authored by Jihn-Fa Jan, reports a successful application of using geographic data for tourism development in a rural community. In this case, the geo-referenced data are obtained through the global positioning system (GPS).

Chi-Shi is an agricultural-based community in the southern part of Taiwan, which is currently engaging in various activities to protect the natural environment and preserve the valuable cultural heritage. With additional interest in promoting

environmental education and cultural tourism, the community collaborated with the author to develop a web-based geographic information system (GIS) for community resources management and for tourism planning.[2] In particular, residents and visitors carried GPS loggers and digital cameras to record GPS coordinates and images while they traveled along the trails. These data were processed using various software tools and stored in a web-based GIS database. With the web interface, tour planners can dynamically query existing maps or create new richly annotated tour maps in real time using Google Maps applications. In addition to the geographic data, this chapter also provides a concrete illustration of how advances in information and communications technology (ICT), specifically, ubiquitous computing and wearable devices, have facilitated the development of citizen science, which has used crowdsourcing to gather information originally only sparsely disseminated, and then to share and process the pooled information so as to enhance decision-making and planning.[3]

Are cultural practices inventions or the propagation of existing ones? To answer that question, particularly the cultural practice of tombstone inscriptions in Taiwan, Oliver Streiter analyzed the tombstone inscriptions and their geographic information on more than 600 burial sites in Taiwan and in Penghu. Chapter 3, "Telling Stories through R: Geo-temporal Mappings of Epigraphic Practices on Penghu," presents this work. One particular place-name style on the tombstones, called *datanghao*, is most popular in recent Taiwan (after the Republic of China) and in Penghu (after the Japanese occupation). Given the localization of tombstones in time and space and their relatively large number, the *datanghao* style offers a unique opportunity to trace back the development of cultural practices under the influence of global political and economic changes.

The primary data collected are digital geo-referenced tombs photos, which contain in their meta-data the geolocation, the altitude, and the cardinal direction of the photo. Additionally, manual annotations and transcriptions values are assigned as attributes to the primary data. Using data collected in Penghu (3304 tombs), the Monte Carlo sampling method is applied to model the propagation of *datanghao* among the Penghu archipelago. In this case, a model is a directed graph where each node represents a burial site and the edges between the burial sites are directed from the early to the later to indicate the propagation direction. To make the models interpretable, in terms of their properties, some assumptions were made during the sampling. Among the sampled models, the one with the shortest average spatial

[2]For the related applications of GIS to the humanities, also see Chaps. 3, 4, and 14. In fact, these four chapters can together be read as part of the spatial humanities.

[3]For a general understanding of citizen science, also known as crowd science, and its recent development, the interested reader is referred to Cooper (2017) and Franzoni and Sauermann (2014).

distance calculated over the entire graph is selected as the final model. The model identifies Xiyu as the geographic origin of the *datanghao* in Penghu.

After interviewing tombstone carvers in Penghu and in Taiwan, Streiter hypothesized that the migration of carvers from Penghu to Taiwan might have caused the spread of *datanghao* to Taiwan. By expanding the Penghu model with data on 67,945 tombs collected in Taiwan, the new model supports the hypothesis. The research concludes that the *datanghao* place-name style on tombstones was invented most probably by a tombstone carver in Xiyu, Penghu to express loyalty toward the Qing dynasty without offending the Japanese during their occupation. Through the migration of carvers from Penghu to Taiwan, the *datanghao* spread and became an epigraphic practice in Taiwan.

### *1.1.2 Text Corpus Data*

Text corpus data are digital data obtained from various sources (e.g., news, publications, and books), where the focus is on the text itself, and the texts are usually relatively long (big). Six chapters in this volume demonstrate the use of big data in this manner. The application domains covered here include history (Chap. 4), economics (Chap. 5), finance (Chap. 6), health (Chap. 7), literature (Chap. 8), and linguistics (Chap. 9). Not only do these six chapters show us how the use of the text corpus data can allow us to address some questions which are beyond the reachability of the conventional social sciences and humanities, but they also extend our generally neglected computational aspects of the social sciences and humanities. They, therefore, represent a new frontier in the social sciences and humanities.

As the title of this section suggests, corpus linguistics plays a threading role for these six chapters, but it serves only as the rudimental element (raw data). On top of it, different techniques, activities, and functions can be further added, such as map construction (Chap. 4), time series analysis (Chap. 5), sentiment analysis (Chap. 6), network modeling (Chap. 7), complexity analysis (Chap. 8), and event analysis (Chap. 9). In sum, these six chapters together show that the information hidden in the text can be very valuable for progress-making in the social sciences and humanities. In the following, we shall briefly give a sketch of each chapter.

#### 1.1.2.1 History in Light of Dynamic Maps

In Chap. 4, "Expressing Dynamic Maps through 17th-Century Taiwan Dutch Manuscripts," Ann Heylen digitized a 17th-century Dutch handwritten manuscript (Church Minutes), which documented the presence of the Dutch community in Taiwan at that time. Based on the information, visualization GIS is developed to provide a better understanding of Taiwan's history in a global setting. The Dutch

United East India Company (VOC) established its presence in Taiwan in 1624 and continued until 1662. Since the VOC possessed quasi-governmental powers, tracing the mobility of their personnel, e.g., relocation and displacement, offers new insights into the social and economic changes taking place on the island during that period of time. Incidentally, the Church Minutes written by clergy contain such information, and hence these minutes are chosen for this research.

Initially, Heylen digitized the manuscripts (with Dutch and English translations). Next, names of persons and places in the text were extracted and standardized through the search/replace command using regular expressions. After that, place names were converted into geo-coding to produce the first blueprint map, where a "Personality Icon" was added to a place name if the person was mentioned at that location. To visualize the mobility of VOC personnel, users can select the person's "Personality Icon" at a location and a pop-up window will appear displaying the next location to which the individual moved. At the new location, one can select the person's "Personality Icon" there and the person's next location will appear. By continuing this process, one can obtain information on the spatial mobility of the individual throughout the time period.

#### 1.1.2.2 Economics Paradigm Shift in Light of Corpus Linguistics

In Chap. 5, "Has Homo economicus Evolved into Homo sapiens from 1992 to 2014?: What Does Corpus Linguistics Say?" Yawen Zou and Shu-Heng Chen present an innovative big data application that studies the paradigm shift in economic research. In 2000, Richard Thaler predicted that economists would shift their research approach from the Homo economicus paradigm to the paradigm of *Homo sapiens*.[4] To check whether this prediction has been fulfilled or not, they adopted a corpus linguistic method to analyze their data.

Initially, they built a corpus using the abstracts of 51,285 economics research articles published from 1992 to 2014 in 42 mainstream economics journals. Next, they identified and selected two sets of keywords that are associated with each of the two paradigms. The keywords are in regular expression form, which supports the don't-care symbol *. For example, "cognit*" matches "cognitive," "cognition," "cognitively," "cognitivity," and so on. Since the two paradigms are opposite to each other, i.e., Homo economicus formulates the rationality of economic behavior in an ideal mathematical optimization framework while *Homo sapiens* emphasizes the consideration of the psychological, cultural, and social factors that constrain a human's rationality, the two keyword-sets are disjointed with little overlapping.

They then counted the frequencies of each keyword in the corpus and regressed these frequencies time series to see which keywords have an upper trend (increased usage) and which keywords have a downward trend (decreased usage) over time.

[4]Richard Thaler is the 2017 Nobel Laureate in Economics.

The results show that 81.4% of the *Homo sapiens* keywords have an upper trend and 65.2% of the Homo economicus keywords have a downward trend. From this observation, they concluded that Thaler's prediction is largely correct. Moreover, since the period studied is not long (22 years), they believe the paradigm shift is still happening and will continue for some time.

Using corpus linguistics to analyze economic texts has become a new research trend in economics. This trend can be considered to represent a small amount of progress in the less-explored interdisciplinary areas between economics and the humanities, the area that Deirdre McCloskey has long promoted, which is partially known as the rhetoric of economics (McCloskey 1983, 1998), and the Nobel Laureate in Economics, Robert Shiller, recently also advocated and coined the term *Narrative Economics* (Shiller 2017). How economists can learn from the humanities is also well illustrated in general by Morson and Schapiro (2017) and, specifically, in behavioral economics by Roy and Zeckhauser (2016). Probably the closest related study to this chapter is the recent application of computational linguistics to the study of the open market operations of Federal Reserve Banks (Hansen et al. 2018).

#### 1.1.2.3 Financial Prediction in Light of Sentiment Analysis

One way to see the close connection between the humanities and the social sciences is through the sentiments or emotions derived from or extracted from the text or, as alternatively put, the psychology of reading the text. In the era of big data, humans will not read or handle the overwhelmingly generated texts themselves; instead, intelligent machines will take over the reading of the massive amount of texts. Machines and the embodied algorithms will simulate the emotions that humans may put into the text as an author or get from the text as a reader.[5] The replacement of humans with machines in reading or even in authoring is currently known as *sentiment analysis*.[6] What does this text mean? Does it imply something positive or negative? What are its implications for investors' emotions? This whole business overarching the humanities, psychology, computer science, and finance has defined a hot spot for the "big data gold rush" in the sense that if you know what the text means, you may know how to invest.

The financial industry has been greatly impacted by the big volume of data and by the technologies that generated these data. It is quite likely in the future that "quants" on Wall Street will no longer solely rely on numerical data to trace the ups and downs of prices; instead, they will increasingly rely on using data analytics, text

[5]There is a philosophical issue as to whether machines will evolve to have their own interpretations of the text and hence develop their own emotions which are different from those of general human beings under the governance of their own culture. More *positively*, would machines surpass humans by demonstrating the features of *positive psychology,* as advocated by Martin Seligman (Seligman 2004), more successfully than humans?

[6]There are already quite a few good references giving a panoramic guide to this fast growing field. The interested reader is referred to Liu (2015), Pozzi et al. (2016), and Cambria et al. (2017).

analytics, and the digital humanities to discover the underlying market sentiments as leading indicators for prices. In fact, a stream of the literature documenting these studies has already been piling up (see Chaps. 6 and 12).[7] In Chap. 6, "Big Data and FinTech," Jia-Lang Seng, Yao-Min Chiang, Pang-Ru Chang, Feng-Shang Wu, Yung-Shen Yen, and Tzu-Chieh Tsai first presented their works using online news to conduct sentiment analysis for the Taiwan stock market. They then discussed the strategies and the computer framework that they have developed to better serve financial institutes using real-time mobile/cloud computing.

The chapter presents two studies on the use of news sentiments to predict Taiwan stock market performance. The first study was based on an asset-pricing model, which incorporated news sentiments related to investment, macroeconomics, and politics. The authors reported that there was a positive correlation between the investment news sentiment and the Taiwan 50 (TW50) index returns. There was also a negative correlation between the political news sentiment and the returns of some stocks listed within the TW50 index. The second study on news sentiment analysis was performed using a software tool that graded the view of a news article as being positive or negative on a scale from $+5$ to $-5$. The kind of news articles that they were concerned with were also related to investment, macroeconomics, and politics. The authors reported that the news sentiments were correlated with the stock prices, but this result may have been affected by the subjective judgment of the software tool. In addition to sentiment analysis, the authors also discussed the strategies that financial business security and technology providers are using to better serve their customers. They proposed a cloud-based mobile computing framework that can handle complex data structures and large amounts of data within a short period of time.

#### 1.1.2.4 Network Modeling of Health-Related Concepts and Institutions

As already seen in Chap. 6, from the perspective of the humanities, each text can be represented by its sentimental and emotional ingredients that present readers with a "mood" of the text. On the other hand, in a more concrete manner, a text can be represented by graphs (networks, word clouds, and maps), which present readers with *a geometry of the text*. Several chapters in this volume do work on the graphical representation of texts. Chapter 5 is an illustration of the *co-word network*, which is basically a single network, a network with one type of link. Chapter 7 extends this graphical representation of texts to *multiplex networks*, or networks with more than one type of link (relation), to which we now turn.

[7] While there are only two chapters collected in this volume, the interested reader may find more useful references in Peterson (2016) and the excellent collections edited by Mitra and Xiang (2016). However, sentiment analysis may go further, beyond what the current literature delineates, and can be further incorporated into agent-based computational finance and give new impetus to behavioral finance (Chen and Venkatachalam 2017).

The World Health Organization (WHO) and the Secretariat of the Convention on Biological Diversity recently decided to strengthen their collaboration, most notably in raising awareness of the complex linkages between biological diversity, ecosystems, and human health, acknowledging the strong connection between biodiversity and health. In a joint report, they highlight the fact that biodiversity loss constitutes a fundamental risk to the healthy and stable ecosystems that sustain all aspects of our societies (WHO-CBD 2015). Themes related to Health are increasingly cited by the COPs (Conferences of the Parties) of the CBD, the Convention on the Conservation of Migratory Species (CMS), and the Convention on International Trade in Endangered Species (CITES), encompassing dimensions of human health, animal health (domestic and wild fauna), and ecosystem health. Other ecological or environmental concepts, such as biodiversity, an ecosystemic approach, and risk assessment, favor the emergence of Health issues and their integration into the CBD.

In Chap. 7, "Health in Biodiversity-Related Conventions: Analysis of a Multiplex Terminological Network (1973–2016)," Claire Lajaunie, Pierre Mazzega, and Romain Boulet investigated which health themes have emerged from these three biodiversity-related conventions. They analyzed how concepts are used in a complete or partial form in each COP and how they are transmitted between COPs through a multiplex network, with each type of link in the network corresponding to a concept. They then identified the most central COPs and their gathering into communities in the process of emerging Health issues.

Their aim is to study the dynamic of health within the biodiversity-related conventions and to capture simultaneously the dynamic of the importance of COPs in the diffusion of health issues and the dynamic of health themes within their decisions and resolutions. The common dynamic shown, thanks to the use of multiplex analysis combined with text mining used in a big data perspective, facilitates in understanding how each concept contributes to the building of an integrative and multi-dimensional approach of Health issues in international environmental law.

In their approach, they first collected the textual corpus from the CBD, CMS, and CITES conventions (agreements) and from the decisions and resolutions published by their respective COPs from 1973 to 2016. Next, they extracted 213 terms that are related to health issues and divided them into 13 concept groups. After that, they counted the number of occurrences of the 13 concepts (occurrences of any of the constituent terms) in the textual corpus. They then analyzed these frequency data using network modeling. The network model consisted of two different kinds of nodes, namely 13 concepts and 27 COPs held from 1973 to 2016. A concept was linked to a COP if the concept was mentioned in the COP's publication. Moreover, two COPs were linked if the same concept was mentioned in both of their publications. By measuring the degree and betweenness centrality of the network, they identified the five most central COPs that played the most important role in disseminating health-related themes. Moreover, the temporal evolution shows that the three most frequently mentioned health-related concepts over time are risk and threat, health and security.

#### 1.1.2.5 Linguistic Complexity of Shakespeare Plays

We have seen the "mood" and the geometry of a text. It is then natural to ask an even more fundamental question: What is the complexity of a text, considering that some texts are simple and some are not? One approach to this issue is to directly provide a complexity measure suitable for sentiment and another measure suitable for landscape, as one can easily imagine that the "mood" can be simple or complex, and the same for its shape. In fact, studies along these lines already exist.[8] However, there is a third approach that looks at this issue, which is presented in Chap. 8.

Although Shakespearean plays have been regarded as the finest works in the English language, they are not always easy to enjoy, because the language can be unfamiliar, and hence intimidating to the general public. Does the linguistic complexity of a Shakespearean drama play a role in its audience's acceptance and its commercial success? In Chap. 8, "How Does Linguistic Complexity in Shakespeare's Plays Relate to the Production History of a Commercial American Theatre?" Brian Kokensparger applied a computational method to the Shakespeare corpus to answer this question.

Kokensparger designed four measures to quantify the linguistic complexity of Shakespeare's 38 plays: average syllables per word, average words per sentence, percentage of complex words, and percentage of words not found in a standard dictionary. The plays were ranked from the lowest linguistic complexity score to the highest. Then these rankings were compared with the ranked production frequency of a commercial Shakespearean theater. The results indicated that the plays offering the highest frequency over the theater's history were also among the least complex of Shakespeare's plays. Therefore, there appears to be a relationship between linguistic complexity in the text of Shakespeare's plays and the commercial viability of offering those plays to a paying audience. As the linguistic complexity of a performed play affects the cognitive load on the audiences, it is reasonable to suggest that plays with the lowest linguistic complexity will be more frequently chosen for production than their counterparts with higher linguistic complexity for a theater that seeks to successfully entertain patrons and keep them coming back.

#### 1.1.2.6 Evolution of "Language" in Light of Corpus Linguistics

Corpus linguistics is the study of languages as expressed in the corpora of real-world texts. In Chap. 9, "Language Communities, Corpora, and Cognition", Huei-Ling Lai, Kawai Chui, Wen-Hui Sah, Siaw-Fong Chung, and Chao-Lin Liu presented three case studies using corpus-based methods to investigate the linguistic patterns and the cognition of different community groups.

[8] For example, for the complexity measure for sentiments, see Joshi et al. (2014); for the complexity measure for networks, see Morzy et al. (2017).

The first study investigated how the lexicalized term <nganggiang stiff neck 硬頸>, a metonymy-based metaphor, is used in the news media. This body-oriented metaphor characterizes a person as being stubborn and tough by describing his/her body expression of making the neck stiff to show an unyielding attitude. To understand how this metaphorical expression has become entrenched and conventionalized to carry such a meaning, they collected online news data from four major newspapers in Taiwan, and counted the usage of the term in the corpus. Quite interestingly, they found that the frequency of the usage is correlated with the major election years in Taiwan. Moreover, while originally carrying a negative connotation, the term is now a positive phrase used to characterize Hakka-related matters. In addition, through the mechanisms of denotation extension, metonymy, and metaphor, its usage has increased over the years.

The second study analyzed the usage of gestures in Taiwan Mandarin conversations. In a collection of 15 recorded conversational excerpts, 2012 gestures were found across male-speaker conversations, female-speaker conversations, and mixed-gender conversations. They counted the frequency of five different kinds of gestures across the three types of interaction and found that each one differed from the others. This indicates that gender affects the usage of gestures.

The third study compared the oral narrative abilities of Mandarin-speaking children with and without the *autism spectrum disorder* (ASD). Among various indices of narrative abilities, referential choice is regarded as an important window to show a speaker's sensitivity to listeners' needs. The authors therefore used referential forms and pragmatic functions data to conduct their study. They found that both groups of children were comparable in using nominal forms such as introducing and reintroducing referents. However, null forms, rather than pronominal forms that are normally used to maintain reference as reported by other researchers, appeared to be the dominant device for both groups of children to maintain reference.

### *1.1.3 Social Media Data*

While digital archives and the resultant text corpus data constitute the essential body of the big data, they are regarded as the "classical type" of big data in the sense that the original forms of texts are not digital and conversions are needed before placing them into the arrays of big data. Nowadays, thanks to the advances in ICT technology, the modern forms of texts are "born" digital; the online news data of Chap. 6 provide a case in point. Apart from that, social media networks built upon and further facilitated by the ICT technology have fundamentally revolutionized the way in which texts can be generated and also archived. Basically, social media and the Internet of everything can technically map and archive the entire world into its cyber counterpart, and thus new historiographies may be introduced. The conventional *discrete-in-time* historiographies will be challenged by the future *continuous-in-time* ones. With this irreversible trend, social media data will eventually monopolize the whole of big data.

Citizen science (crowd science or volunteer science), as we have already seen in Chap. 2, demonstrates how social media data can fundamentally change the way in which we do science by allowing people from all over the world to contribute to groundbreaking scientific discoveries. In addition to science, social media data can also exert great influences on the way in which the democratic system is operated; nevertheless, in addition to golden opportunities and promises presented to public administrators, challenges are also prevalent. Basically, it is still not clear whether we shall have more "wisdom of crowds" (information aggregation) or more "stupidities of herds" (noise amplification). The accumulated discussions are very long.[9] To some extent, all are concerned with the possibilities of building a good or better society using advanced digital technology, artificial intelligence, and big data. The following two chapters of this volume exhibit the typical flavor of this so-called *social media dilemma*; Chap. 10 is concerned with digital governance, whereas Chap. 11 is concerned with the grassroots politics.

#### 1.1.3.1 Digital Governance Using Public Opinion Mining

Social media have become a popular channel for citizens to express their opinions and complaints regarding public policies. While the data volume is large, the analysis and interpretation of the data to produce meaningful insights is not an easy task. In Chap. 10, "From Naive Expectation to Realistic Progress – Government Applications of Big Data to Public Opinions Mining," Naiyi Hsiao, Zhoupeng Liao, and Don-Yun Chen presented their work on Internet public opinion analysis regarding the *Free Economic Pilot Zone* (FEPZ) policy in Taiwan.

In March 2014, the FEPZ bill was submitted to the Legislative Yuan for review and approval. The bill raised much controversy due to its *Free Trade Agreement* (FTA) with Mainland China, which has become a cause for concern due to fears of losing political independence under the increased economic dependency on Mainland China. To understand public opinion in relation to the bill, the Taiwan National Development Council commissioned the authors and a private technology company to collect and analyze unstructured public opinion data from various Internet media, including news websites, forums, blogs, social media (Facebook and Twitter), and PTT.[10] The analyzed results were then presented to the public officials for their feedback.

During the study conducted from May to November of 2014, many governmental officials participated in the project, for example, by providing keywords, key events, and the names of policy-relevant stakeholders for the team to query media data for

[9]Interested readers are referred to Bauerlein (2008), Sunstein (2008), Ceron et al. (2016), Thompson (2016), Helbing et al. (2017), O'Neil (2017), and Stephens-Davidowitz and Pabon (2017).

[10]The PTT Bulletin Board System is the largest terminal-based bulletin board system (BBS) based in Taiwan. For more information, see https://en.wikipedia.org/wiki/PTT_Bulletin_Board_System.

various kinds of analysis: sentiments (positive/negative), volumes, popular media channels, and the public opinion leaders. However, feedback in regard to the analysis was mixed. Some argued that most netizens are not experts on economic policies; hence their opinions appear relatively useless. Others believed that the statistics and computer algorithms used to conduct the analysis were not adequate. Nevertheless, much has been learned from this study and it provides a baseline for future improvement.

#### 1.1.3.2 Grassroots Politics

Social media have also become a critical platform for social activists to promote their movement. A case in point is the *Sunflower Student Movement* in Taiwan, where social media were the major vehicles used to disseminate information and to organize activities. Unlike the keyword mining approach used in the previous chapter, the study in Chap. 11 conducted by Hui-Wen Liu, I-Ying Lin, Ming-Te Chi, and Kuo-Wei Hsu analyzed the fan pages centered around the Sunflower Student Movement on Facebook, and is entitled *Understanding "the User-Generated": The Construction of the "ABC model" and the Imagination of "Digital Humanities."*

The Sunflower Student Movement evolved as a protest movement driven by a coalition of students and civic groups to protest the passing of the Cross-Strait Service Trade Agreement (CSSTA) by the then ruling party Kuomintang (KMT) at the legislature without a clause-by-clause review. The movement started on March 18, 2014 when a group of students and activists occupied the Legislative Yuan. Over the next 23 days, various Facebook fan pages were created, including that of the leading organization, *Black Island Youth Front* (BIYF), and two others that concentrated on this movement during the occupation (*New e-Forum* and *Anti-Media Monsters Youths*, AMMY). In addition, at least 13 other fan pages were created to support this movement.

The authors chose 16 fan pages for analysis. For each fan page, they first collected its activities (*fan*, *like*, *share*, and *comment*) during the movement from March 18 to April 11 in 2014. Next, they computed their proposed *ABC indices*. A represented "activity," summing up the activities of each user within a fan page. B represented "broadness," totaling the amounts of posts in which each user had participated. C referred to "continuity," calculating the duration time of the various users, from the first time they participated to the last time they left their digital footprints on one fan page. Using the ABC indices, they found diversified functions contributed by the fan pages during the Sunflower Student Movement.

All in all, through the social media data, this chapter is able to provide a lot of details regarding the university students' participation process in politics via social media networks. The findings have become valuable additions to the growing literature on this research stream, see, for example, Conover et al. (2013) and Loader et al. (2015).

## 1.2 Survey and Challenges

The first part of the volume focuses on the use of big data and the kinds and the flavor of the studies facilitated by the use of big data. Each of the ten chapters there works with one or more specific kinds of big data and addresses a set of intriguing questions, covering subjects belonging to anthropology, art and theater, citizen science, economics, finance, history, linguistics, literature, political science, public health, and sociology. Differing from the first part, the second part of the volume does not address a specific set of questions with a particular set of big data; instead, it provides a panoramic view of the development of big data in the computational social sciences and humanities (CSS&H), including its trendy directions and the evoked challenges. Hence, the two parts of the book are connected by employing the *specific-to-general* scheme. With this description, Part 2 is further divided into two sub-parts. We begin with a sketch of the general development of big data in CSS&H (Sect. 1.2.1). Four survey articles are written for this purpose, and they together cover some representative cases of the timely development of big data in business (Sect. 1.2.1.1), the social sciences (Sect. 1.2.1.2), the humanities (Sect. 1.2.1.3), and technology (Sect. 1.2.1.4). They are big data finance (Chap. 12), big data in psychology (Chap. 13), spatial humanities (Chap. 14), and cloud computing (Chap. 15). The second sub-part (Sect. 1.2.2) then presents an overview of some of the challenges associated with big data. Four challenges are presented in this book, namely the complexity of big data (Sect. 1.2.2.1) or the ontology and epistemology of big data, big data search (Sect. 1.2.2.2), big data simulation (Sect. 1.2.2.3), and big data risks (Sect. 1.2.2.4). While these challenges may have also been addressed in the literature, in this book these challenges are uniquely presented by social scientists from their work on mass communication (Chap. 16), anthropology (Chap. 17), history (Chap. 18), and sociology (Chap. 19); hence, their shared views are domain-specific and more focused.

### *1.2.1 Survey of Published Research*

#### 1.2.1.1 Big Data Finance

From the perspective of business, probably one the most astonishing examples of progress observed in the last decade is in the area known as *big data finance*. FinTech, as we have seen in Chap. 6, is a case in point, but, from a macrocosmic viewpoint, what can interest us is the following: with the avalanche of financial market data coming out every minute, in the form of Yahoo Finance, TRMI sentiment data, and other social media, have these financial big data changed financial market behavior? In Chap. 12, "Big Data Finance and Financial Markets," Dehua Shen and Shu-Heng Chen have surveyed those works which help define what we now know as big data finance.

There are two fundamental cornerstones in finance. The first one is the efficient markets hypothesis, basically, the unpredictability of financial returns. The second one is various characterizations of market activities, such as volatility, trading volumes, duration between trades, spreads, and depths; sometimes, their steady patterns are known as "stylized facts" (Chen 2008). With big data, Shen and Chen ask whether more light will be shed on studies of these issues. The first question has been addressed many times in similar forms in the history of economics, and no easy conclusion will be reached.[11] The second one is key because of the nature of big data.

Big data, by its very nature, is the archive of what people said, what they did, and what they thought (Chen and Venkatachalam 2017). It can inform us of lots of human intentions, interactions, emotions, attitudes, and hence decisions (see also Chap. 13 of this volume); to some extent, it functions as a mirror of the world, enabling us to see who we are and how we got here. In other words, we can now see the physical world around us through the lens of the cyber world, thanks to various measures derived from text mining, content analysis, sentiment analysis, and Google trends. Hence, one may reasonably expect to be exposed to an unprecedentedly high level of transparency, and not just be recipients of more information. The theory of general reflexivity (Soros 2013), level-k reasoning, or similar forms of reasoning (Chen 2013) may then suggest that this can further lead to some fundamental changes from microcosmic decision-making to macrocosmic market dynamics; for example, patterns of volatility can be distorted, as with many other familiar financial patterns, and therefore be destroyed. Hence, the survey provided by Shen and Chen is best viewed as the beginning of big data finance, and one still needs to figure out how big data finance can be fundamentally different from "small data" finance, by incorporating human cognition, emotions, social interactions and, above all, human psychology, a subject to which we now turn.

#### 1.2.1.2 Big Data in Psychology

Psychology is the science of behavior and mind. It seeks to understand individuals and groups by establishing general principles and by studying specific cases. While many recognized research methods, such as controlled laboratory experiments, and electroencephalogram (EEG) and animal studies, have contributed to the advancement of the field, big data are providing new alternatives to study psychology. As already mentioned, the very nature of big data is the mapping of people's behavior in the physical world to its cyber counterpart. Therefore, it is anticipated that some human behavior which is not directly observable from the physical world may

[11] The conundrum has been well illustrated by the so-called *adaptive market hypothesis*, which endowed the efficient markets hypothesis with a dynamic and evolutionary interpretation (Lo 2004). In the vein of the agent-based fashion, the adaptive market hypothesis has been further studied in the form of the *market fraction hypothesis* (Chen et al. 2010).

be easier to observe in the cyber world. Hence, some forms of big data can be very useful for psychology, in particular, social media data, health tracker data, geolocation data, dynamic public records, travel route data, behavioral and genetic data, etc.

Currently, maybe the most ambitious project launched to explore this research opportunity is the Kavli HUMAN Project.[12] This project aims to gather a detailed array of measurements from 10,000 New York City residents over a 20-year span, allowing a team of scientists to monitor in intimate detail how these New Yorkers lead their lives over the course of 20 years, including where they go, what they eat, who they talk to, what they buy, and how their bodies grow, change, and deteriorate. Needless to say, this is a massive data-collection endeavor with the main pursuit being to learn how everything, from biology to behavior and the environment, affects the human condition; for example, how biological, medical, and social factors interact and impact the risks of cognitive decline from birth through to older age. We, as decision makers, make a large number of decisions (choices) in each typical day, some of which can substantially impact our well-being (Clark et al. 2018). Hence, if we can be better informed of how these choices are made, we may further improve our quality of life and increase our "happiness index."

In fact, much before the advent of the big data era, many man-made efforts to develop "big data" already existed. For example, it is known that the longest study of adult life that has ever been conducted is probably the Harvard Study of Adult Development (Vaillant 2008). In this unprecedented series of studies, Harvard Medical School has followed 824 subjects associated with different genders and different economic and social status, from their teens to old age. This "big data" study, containing subjects' individual histories, work, home lives, and health, which may be the most complete ever done anywhere in the world, has been used to illustrate the factors involved in reaching a happy, healthy old age. This pioneering study indicates that the demand for big data in psychology research has been much ahead of the times.

While psychologists have been aware of the indispensability of big data in their research, the responses made and the actions taken have by no means been sluggish. In Chap. 13, "Applications of Internet Methods in Psychology," Lee-Xieng Yang reviewed works that leverage big data to conduct research in modern psychology. In his survey, Yang grouped the existing works into three categories.[13]

First, crowdsourcing that has been used to conduct psychological surveys and experiments. Various works have reported that using Internet websites, such as

[12]This project is carried out within a collaboration between the Kavli Foundation, the Institute for the Interdisciplinary Study of Decision Making at New York University (NYU), and the NYU Center for Urban Science and Progress. For more details, the interested reader is referred to Azmak et al. (2015).

[13]The current use of big data in psychology is not just exhausted by the survey presented in this chapter. The journal *Psychological Methods* has published a special issue on this frontier (Harlow and Oswald 2016). For other developments, the interested reader is also referred to Cheung and Jak (2016) and Jones (2016).

Amazon's Mechanical Turk (MTurk), to recruit participants for psychological experiments has generated results that are consistent with those produced in a laboratory setting. Moreover, personality measures that survey data collected using Internet websites are more representative than traditional samples with respect to gender, socioeconomic status, geographic location, and age, and are about as representative as traditional samples with respect to race. However, there are also concerns about data quality due to duplication of participants and the cost caused by the large amount of post hoc data exclusion. Yang suggested that researchers consider their psychological study factors, such as the type of dependent variables and prior knowledge interference, to decide whether to conduct a survey or an experiment on Internet websites.

Second, Google news archives and the Wikipedia database that have been used to study psychology at the population level. Yang exemplified this direction of research with one work that used Google news archives to measure the amount of media attention a humanitarian crisis receives. They found that the more deaths that are involved in a crisis, the more media attention it receives. This result is consistent with previous studies reporting that humans generally have a diminishing sensitivity to the number of human fatalities. Another study used the entire English language Wikipedia corpus to train a computational model that represents human heuristics. The model was then used to answer a series of multiple choice trivia questions. One interesting observation is that the trained model mimics human behavior in making the probabilistic fallacies associated with the representativeness heuristic.[14]

Third, the influences of social networks on individuals. Yang referred to a study that has used Facebook data to study the emotional contagion of human beings. The results indicate that the emotion expressed in a user's Facebook friends' posts is a valid predictor of the emotional expression of the user's own posts. This, to some extent, is consistent with the so-called three degrees of influence in the literature on social networks (Pinheiro et al. 2014).

#### 1.2.1.3 Spatial Humanities

Big data form the digital archive of what people did, what they said, and what they thought. What is implicit in this definition are the surroundings or the embeddedness, i.e., the places, the spaces within which these actions, narratives, perceptions, beliefs, and more extended social settings and social interactions are operated. The surroundings do not just refer to the physical settings, but, more importantly, to the information or the meanings associated with these settings, which in Chinese culture is broadly known as Feng Shui (Rossbach 1983; Webster 2012).

[14]The representativeness heuristic is one of the heuristics that has been carefully studied by psychologists and behavioral economists, regarding how human decisions or judgments are made under uncertainty (Kahneman and Tversky 1972).

Over the last few decades, the awareness of these surroundings, either known as the *spatial awareness* or the *network awareness* (the spatial thinking or the network thinking), has been widely received in parallel in many disciplines, covering both the humanities and social sciences. "Space," as it may literally suggest, becomes the engine to integrate subjects originally studied in isolation in different disciplines, which include people, time, events, history, beliefs, cultures, religion, politics, etc., as already exemplified in Chaps. 3 and 4. This integration forcefully shows the dynamic nature of space, motivates us to adopt a spatial approach to historiography, and promotes *spatio-temporal thinking* in the humanities and social sciences.

On the other hand, this "spatial turn" has also occurred within the discipline of geography itself. Beginning in the 1970s, many geographers were seeking alternative paradigms to rigorous geographical analysis that were not reducible to merely geometries. As Edward Soja (1940–2015) observed, "rather than being seen only as a physical backdrop, container, or stage to human life, space is more insightfully viewed as complex social formation, part of a dynamic process (Soja 2001)." This desire to make maps "deeper" with many time-layers of features associated with the same location further gained momentum with the technology innovation in geography around the same time, especially geographical information systems (GIS). These GIS refer to software that captures, stores, manages, displays, and analyzes information linked to a location on earth. GIS can relate different types of data—quantitative, textual, image, and audio—to each other based on their shared location. It also allows a visualization of these relationships on a map of the geographical space in which they all occur. The availability of GIS as well as other related technologies plus the spatial turn in geography have facilitated the spread of the idea of space, place, and place-making in the humanities, and have helped grow the interdisciplinary field, *spatial humanities*.

With the funding of the National Chengchi University President's Office and other sponsorships, the Asia-Pacific Spatio Temporal Institute (APSTI) was established in 2014. The Institute is a home for innovative GIS-based research on humanities-related subjects.[15] In Chap. 14, "Spatial Humanities: An Integrated Approach to Spatiotemporal Research," David Blundell, Ching-Chih Lin, and James Morris have summarized four thematic forms of research that are ongoing in APSTI and mark the nascent field of spatial humanities.[16]

First, there is the attempt to develop an interactive 3D visualization website to enhance the understanding of the diffusion of culture and oceanic navigation in Monsoon Asia. Second, there is the development of a GIS database that provides a heritage inventory and its management, with a specific application to the Nouli community in Taiwan. Third, there is the documenting and mapping of earth god shrines to establish the patterns of settlement, communal organization, and historical trade networks of communities in Taiwan, southern maritime China, Hong Kong,

[15]The interested reader is welcome to visit its home page: http://apsti.nccu.edu.tw/.

[16]For a general background of this fast-growing field, the interested reader is referred to Bodenhamer et al. (2010).

Macau, and outlying islands. Fourth, there is the development of methods and resources, such as GPS devices and visualization, to support the study of Chinese religion.

#### 1.2.1.4 Cloud Computing

Cloud computing is an Internet-based form of computing that provides shared computer processing resources and data storage to other computers and devices on demand. Advocates claim that cloud computing allows organizations to avoid up-front infrastructure costs (e.g., the purchase of hardware servers). Moreover, individual researchers can enjoy the high-performance computer power from their own desktops and laptops. With the increased number of viable cloud computing providers on the market, adopting the cloud for big data research has become a very attractive option for computational social scientists.

In Chap. 15, "Cloud Computing in the Social Sciences and Humanities," Michael Gallagher has reviewed four major cloud-computing platforms, namely Amazon Web Services, Google Cloud Platform, Microsoft Azure, and Hewlett Packard Enterprises, from the perspective of a social science researcher. To address the deep learning curve and user-friendliness issues of using cloud computing, he has also provided step-by-step instructions to create an Amazon Web Services account and to set up the computing environment.

### *1.2.2 Challenges of Using Big Data for Research*

In transitioning from the small-data research paradigm to the big-data research platform, social scientists and humanists face various challenges. Sect. 1.2.2.1 discusses about how to process/transform social media big data to be used to answer mass communication research questions. In Sect. 1.2.2.2, big database query/search issues, such as the ranking of relevancy to the query among a large number of possible matches, are addressed. Sect. 1.2.2.3 presents an approach to conduct big data simulation using agent-based models. Sect. 1.2.2.4 argues the risk of cyborgs, hybrids of humans and technologies, developed under the big data technologies era. In other words, our thinking will be a hybrid of biological and non-biological thinking. Would that be a curse or a blessing for the humanity in the long run? History will be the judge of that.

#### 1.2.2.1 Big Data Complexity

Mass communications research is chiefly concerned with how the content of mass communications affects the attitudes, opinions, emotions, and ultimately behaviors of the people who receive the messages. Since social media have become important

vehicles to disseminate information, it is natural for mass communications scholars to become interested in analyzing their content. However, the data analytics developed so far is still inadequate to address the complexity of social media. Consequently, diving in the social media big data to search for the answers to their research questions is not an easy task for mass communications researchers. In Chap. 16, "Analysis of Social Media Data: An Introduction to the Characteristics and Chronological Process," Pai-Lin Chen, Yu-Chung Cheng, and Kung Chen discuss the challenges of social media data analytics.

Two challenges are addressed in this chapter. The primary challenge resides in the characteristics of social media data. First, social media data are enormous in scale and diverse in structure, which make the traditional research methods unworkable. Second, social media data are generated by humans, and hence are made complicated by the language usages and by the diverse human interaction patterns. Third, data integrity is not reliable because not all data sources are accessible. In addition, data transparency is insufficient, due to the black-box data algorithms. The authors suggest how these problems can be ameliorated or managed.

The second challenge concerns how to connect a research question to the data and to discover a problem-solving approach. To overcome this problem, the authors have proposed establishing a team of trans-disciplinary experts that consists of not only mass communications researchers but also data scientists who are familiar with the data processing tools. Together, the team can extract related data from the big data to address the research questions posted by researchers.

#### 1.2.2.2 Big Data Search

There exist many social science and humanities databases that can support the sharing and reuse of information. One example is the HRAF, an abbreviation for Human Relations Area Files databases (eHRAF World Cultures and eHRAF Archaeology). Over the years, the databases have grown very rapidly, which has created a database search challenge: there are too many complex entries retrieved from each query. In Chap. 17, "Big Data and Research Opportunities Using HRAF Databases," Michael Fischer and Carol Ember discuss some potential and partial solutions to address this big data search problem. In addition, they have highlighted the challenges of integrating disparate ethnographic materials into HRAF databases to improve their reusability.

The HRAF databases contain texts that describe social and cultural life in past and present societies around the world. As of the spring of 2018, the two eHRAF databases contained almost three million "paragraph" units from over 8000 documents describing over 400 societies and archeological traditions. A typical keyword search normally returns around 20,000 paragraphs. Although there are simple strategies to reduce the number of returns, such as narrowing the search to a focal community and time period, or combining subject categories or adding keywords, these strategies are not effective due to the complexity of the

databases. They therefore proposed using big data technologies to develop a range of post-processing tools and methods to be applied to the returned text to help the users refine their search. An example of one method is search by example, where the user has selected entries from a search return that has been used as a basis for identifying similar entries from across the databases.

The HRAF has another challenge: how to achieve the interoperability across various kinds of ethnographic sources so that more documents can be shared. They address this problem at three levels: syntactic interoperability, semantic interoperability, and pragmatic interoperability. Currently, they are exploring research opportunities to resolve these issues.

#### 1.2.2.3 Big Data Simulation

The third challenge for the use of big data is the frequent lack of a scientific or theoretical foundation for big data. As already seen in Chap. 16, big data are often naturally occurring (as archives of social processes); they are not obtained by being designed for the purpose of scientific inquiries. Hence, it will be desirable to place them in plausible contexts in which these big data can be generated. As Chen and Venkatachalam (2017) have argued, to the best of current knowledge, agent-based modeling is the only methodology that can help simulate big data, given that big data are often continuous in time and are individual-based (actor-based, agent-based). Despite this being so, it is still not often seen in the literature how an agent-based model is actually applied to simulate a real set of big data. In Chap. 18, "Computational History: From Big Data to Big Simulations," Andrea Nanetti and Siew Ann Cheong have echoed many of the big database challenges mentioned in the previous chapter. In addition, they have proposed new methodologies to take advantage of the big volume of historical data and have used agent-based model simulation to deepen our understanding of why our history has developed the way it has.

Their proposed method consists of two stages. It begins with restructuring the historical data, so that the narratives of events can be extracted and their relationship can be identified. In the second stage, the extracted narratives and identified relationship are employed to build an agent-based model and the agent-based simulation is performed to identify events as tipping points in a society's natural nonlinear life.

They used their research work on the Engineering Historical Memory (EHM) project to demonstrate their methodologies. The benefit of this method is that, based on the large number of simulated histories, we can identify the tipping point where the histories diverge. By comparing the key factors that led to the two different outcomes, and by understanding how they are different, we can have a better understanding of why our history developed the way it did.

#### 1.2.2.4 Big Data Risks

As already mentioned, big data do not constitute a panacea, and their dark side should never be ignored. Not only shouldn't big data be regarded as a solution for everything, but they could further be a trigger for many problems, regardless of being new or existing data. First, we have not been assured that big data will enhance our decision-making capability and quality. Behavioral economists may tell us more on this. There is no clear picture as to how the stupidities of herds can be avoided. The floods of fake news may continue to characterize this year, next year, and the following years. So, despite the efforts being made by computer scientists (Conroy et al. 2015), we have not seen powerful remedies (algorithms) to take away these new forms of "virus." In fact, as we have learned from history, each technology-triggered problem can have human behavior as its catalyst; hence, psychology or behavioral economics may play an even more important role in the era of big data. The second one is more familiar. The very nature of big data implies little privacy and can threaten cybersecurity unless some cautions and actions are taken to protect against privacy rights (Lane et al. 2014).[17]

Nevertheless, the two fundamental challenges above do not exhaust the list of big data threats. Here comes another philosophical question: Who are we? Can we safeguard our identity? How do we know that we are loyal to ourselves? This series of "self" questions is not about cybersecurity, but about cyborgs.[18] In Chap. 19, "A Posthumanist Reflection on the Digital Humanities and Social Sciences," Chia-Rong Tsao has discussed a challenging reality: we are post-human cyborgs, hybrids of humans and technologies. Moreover, the symbiotic relationship between humans and technologies has transformed our cognition with the technologies that constitute the "extended cognitive system." In the big data era where technologies are playing an even more important role in the knowledge development process, Tsao has advocated that we should not ignore the dangers that the development of the digital humanities and computational social sciences may bring.

In conducting research on the digital humanities and social sciences, Tsao has argued that the relationship between researchers and digital tools can be examined from two perspectives. First, the researcher and the digital tool are regarded as collaborators within a hybrid network, which produces knowledge. Second, they not only collaborate with but also co-constitute each other; thus the digital tool changes the researcher when it magnifies and reduces the different dimensions of the world. In other words, the digital tool is not only a passive object manipulated by the researcher, but also a delegated actor that extends the agency of its user and

[17] This feature can be coined as the big data paradox, namely too big to be "small."

[18] In the development of the computational social sciences and humanities, the role of cyborgs is often ignored. For example, in social simulation or agent-based simulation, there is a clear distinction between human agents and software agents, but their possible hybridizations are left out. See Chen et al. (2018).

is capable of betraying users because of neglected rules or incorrect codes. To what extent would the danger be increased by such betraying behavior? Tsao seeks to investigate this question in future work.

## 1.3 Conclusion and Outlook

With massive amounts of digitized and born-digital data accessible to the public, many social scientists and humanists have explored such data to advance their research. In this book we present works that have incorporated geographic data (Chaps. 2 and 3), text corpus data (Chaps. 4, 5, 6, 7, 8, and 9), and social media data (Chaps. 10 and 11). Meanwhile, surveys of research leveraging big data in finance (Chap. 12), in psychology (Chap. 13), and in digital humanities (Chap. 14) are provided. Moreover, the transition of computational social sciences and digital humanities research from the small-data to the big-data scale presents many challenges Kleiner et al. (2015). We have also discussed some of them in this book (Chaps. 16, 17, 18, and 19).

Although big data research is not going to replace the traditional methods used in studies in the social sciences and humanities in the near future, its importance is undeniably growing. Currently, the number of social scientists embracing big data for new research opportunities is still relatively small. This might be due to the research questions that most social scientists are investigating being traditional ones, which can be solved using small data sets on their own laptops. To inspire more researchers to explore the advantages of big data, we need to provide more demonstrator projects, which can serve as examples to the rest of the community of what big data can provide for scientific development.

Another path to motivate more big data research involves providing assistance on big data tools and methods. For example, Chap. 15 provides a survey and instructions on using cloud computing. However, a more formal approach would be to establish an institution, which provides the research environment mentioned at the beginning of the chapter. In addition to data governance, the institution can offer assistance to social scientists and humanists who are interested in becoming part of the exciting big data research in their fields.

## References

Azmak, O., Bayer, H., Caplin, A., Chun, M., Glimcher, P., Koonin, S., & Patrinos, A. (2015). Using Big data to understand the human condition: The Kavli HUMAN project. *Big Data, 3*(3), 173–188.

Bauerlein, M. (2008). *The dumbest generation: How the digital age stupefies young Americans and jeopardizes our future (or, don't trust anyone under 30)*. London: Penguin.

Biemann, C., Crane, G. R., Fellbaum, C. D., & Mehler, A. (2014). Computational humanities-bridging the gap between computer science and digital humanities (Dagstuhl Seminar 14301). In *Dagstuhl reports* (Vol. 4, No. 7). Dagstuhl: Schloss Dagstuhl-Leibniz-Zentrum für Informatik.

Bodenhamer, D. J., Corrigan, J., & Harris, T. M. (Eds.). (2010). *The spatial humanities: GIS and the future of humanities scholarship*. Bloomington: Indiana University Press.

Cambria, E., Das, D., Bandyopadhyay, S., & Feraco, A. (Eds.). (2017). *A practical guide to sentiment analysis* (Vol. 5). Heidelberg: Springer.

Ceron, A., Curini, L., & Iacus, S. M. (2016). *Politics and Big data: Nowcasting and forecasting elections with social media*. Didcot: Taylor & Francis.

Chen, S.-H. (2008). Financial applications: Stock markets. In B. Wang (Ed.), *Wiley encyclopedia of computer science and engineering* (pp. 1227–1244). Hoboken: Wiley.

Chen, S.-H. (2013). Reasoning-based artificial agents in agent-based computational economics. In K. Nakamatsu & L. Jain (Eds.), *The handbook on reasoning-based intelligent systems* (pp. 575–602). Singapore: World Scientific.

Chen, S.-H., & Venkatachalam, R. (2017). Agent-based modelling as a foundation for big data. *Journal of Economic Methodology, 24*(4), 362–383.

Chen, S. H., Kaboudan, M., & Du, Y. R. (2018). Computational economics in the era of natural computationalism. In S. H. Chen, M. Kaboudan, & Y. R. Du (Eds.), *The Oxford handbook of computational economics and finance*. New York: Oxford.

Chen, S.-H., Kampouridis, M., & Tsang, E. (2010). Microstructure dynamics and agent-based financial markets. In *International workshop on multi-agent systems and agent-based simulation* (pp. 121–135). Berlin: Springer.

Cheung, M. W. L., & Jak, S. (2016). Analyzing big data in psychology: A split/analyze/meta-analyze approach. *Frontiers in Psychology, 7*, 738 https://doi.org/10.3389/fpsyg.2016.00738.

Clark, A. E., Flèche, S., Layard, R., Powdthavee, N., & Ward, G. (2018). *The origins of happiness: The science of Well-being over the life course*. Princeton: Princeton University Press.

Conover, M. D., Ferrara, E., Menczer, F., & Flammini, A. (2013). The digital evolution of occupy Wall Street. *PLoS One, 8*(5), e64679.

Conroy, N. J., Rubin, V. L., & Chen, Y. (2015). Automatic deception detection: Methods for finding fake news. *Proceedings of the Association for Information Science and Technology, 52*(1), 1–4.

Cooper, C. (2017). *Citizen science: How ordinary people are changing the face of discovery*. London: Gerald Duckworth & Co.

Cruz-Neira, C. (2003). Computational humanities: The new challenge for VR. *IEEE Computer Graphics and Applications, 23*(3), 10–13.

Franzoni, C., & Sauermann, H. (2014). Crowd science: The organization of scientific research in open collaborative projects. *Research Policy, 43*(1), 1–20.

Gavin, M. (2014). Agent-based modeling and historical simulation. *DHQ: Digital Humanities Quarterly, 8*(4). Retrieved January 12, 2015, from http://www.digitalhumanities.org/dhq/vol/8/4/000195/000195.html

Hansen, S., McMahon, M., & Prat, A. (2018). Transparency and deliberation within the FOMC: A computational linguistics approach. *The Quarterly Journal of Economics, 1*, 70. https://doi.org/10.1093/qje/qjx045.

Harlow, L. L., & Oswald, F. L. (2016). Big data in psychology: Introduction to the special issue. *Psychological Methods, 21*(4), 447.

Helbing, D., Frey, B. S., Gigerenzer, G., Hafen, E., Hagner, M., Hofstetter, Y., et al. (2017). Will democracy survive big data and artificial intelligence? *Scientific American, 25*. Retrieved February 27, 2017, from https://www.scientificamerican.com/article/will-democracy-survive-big-data-and-artificial-intelligence/ (accessed 27 Feb, 2017)

Jones, M. N. (Ed.). (2016). *Big data in cognitive science*. Hove: Psychology Press.

Joshi, A., Mishra, A., Senthamilselvan, N., & Bhattacharyya, P. (2014). Measuring sentiment annotation complexity of text. In *Proceedings of the 52nd annual meeting of the Association for Computational Linguistics (volume 2: Short papers)* (Vol. 2, pp. 36–41).

Kahneman, D., & Tversky, A. (1972). Subjective probability: A judgment of representativeness. *Cognitive Psychology, 3*(3), 430–454.

Kleiner, B., Stam, A., & Pekari, A. (2015). *Big data for the social sciences* (FORS Working Papers, 2015-2).

Lane, J., Stodden, V., Bender, S., & Nissenbaum, H. (Eds.). (2014). *Privacy, big data, and the public good: Frameworks for engagement*. Cambridge: Cambridge University Press.

Liu, B. (2015). *Sentiment analysis: Mining opinions, sentiments, and emotions*. Cambridge: Cambridge University Press.

Loader, B. D., Vromen, A., Xenos, M. A., Steel, H., & Burgum, S. (2015). Campus politics, student societies and social media. *The Sociological Review, 63*(4), 820–839.

Lo, A. W. (2004). The adaptive markets hypothesis: Market efficiency from an evolutionary perspective. *Journal of Portfolio Management, 30*, 15–29.

McCloskey, D. N. (1983). The rhetoric of economics. *Journal of Economic Literature, 21*(2), 481–517.

McCloskey, D. N. (1998). *The rhetoric of economics*. Madison: University of Wisconsin Press.

Mitra, G., & Xiang, Y. (2016). *Handbook of sentiment analysis in finance*. New York: Albury Books.

Morson, G. S., & Schapiro, M. (2017). *Cents and sensibility: What economics can learn from the humanities*. Princeton: Princeton University Press.

Morzy, M., Kajdanowicz, T., & Kazienko, P. (2017). On measuring the complexity of networks: Kolmogorov complexity versus entropy. *Complexity, 2017*, 3250301.

O'Neil, C. (2017). *Weapons of math destruction: How big data increases inequality and threatens democracy*. New York: Broadway Books.

Peters, B. (2012). *The big data gold rush*. New York: Forbes Magazine.

Peterson, R. L. (2016). *Trading on sentiment: The power of minds over markets*. Hoboken: Wiley.

Pinheiro, F. L., Santos, M. D., Santos, F. C., & Pacheco, J. M. (2014). Origin of peer influence in social networks. *Physical Review Letters, 112*(9), 098702.

Pozzi, F. A., Fersini, E., Messina, E., & Liu, B. (2016). *Sentiment analysis in social networks*. Burlington: Morgan Kaufmann.

Rossbach, S. (1983). *Feng Shui, the Chinese art of placement*. New York: EP Dutton. Inc.

Roy, D., & Zeckhauser, R. (2016). Literary light on decision's dark corner. In R. Frantz, S. H. Chen, K. Dopfer, F. Heukelom, & S. Mousavi (Eds.), *Routledge handbook of behavioral economics* (pp. 230–249). Abingdon: Routledge.

Savage, M., & Burrows, R. (2007). The coming crisis of empirical sociology. *Sociology, 41*(5), 885–899.

Seligman, M. E. (2004). *Authentic happiness: Using the new positive psychology to realize your potential for lasting fulfillment*. New York: Simon and Schuster.

Shiller, R. J. (2017). Narrative economics. *American Economic Review, 107*(4), 967–1004.

Soja, E. (2001). In different spaces: Interpreting the spatial organization of societies. In *Proceedings, 3rd international space syntax symposium* (p. 1-s1).

Soros, G. (2013). Fallibility, reflexivity, and the human uncertainty principle. *Journal of Economic Methodology, 20*(4), 309–329.

Stephens-Davidowitz, S., & Pabon, A. (2017). *Everybody lies: Big data, new data, and what the internet can tell us about who we really are*. New York: HarperLuxe.

Sunstein, C. R. (2008). Neither Hayek nor Habermas. *Public Choice, 134*(1–2), 87–95.

Thompson, A. (2016). *Journalists and Trump voters live in separate online bubbles, MIT analysis shows*. New York: Vice News.

Vaillant, G. E. (2008). *Aging well: Surprising guideposts to a happier life from the landmark study of adult development*. Boston: Little, Brown.

Webster, R. (2012). *Feng Shui for beginners: Successful living by design*. Woodbury: Llewellyn Worldwide.

WHO-CBD. (2015). Connecting global priorities: biodiversity and human health: a state of knowledge review, p. 344.

# Part I
# Practice

# Chapter 2
# Application of Citizen Science and Volunteered Geographic Information (VGI): Tourism Development for Rural Communities


**Jihn-Fa Jan**

## 2.1 Introduction

GIS (geographic information system) relies on digital data, which are categorized as: (1) primary data captured using surveying techniques such as satellite imaging, aerial photography, SONAR (sound navigation and ranging), RADAR (radio detection and ranging), LIDAR (light detection and ranging), GPS (global positioning system), and electronic total station; and (2) secondary data obtained by processing or analyzing primary data, e.g., scanning maps, heads-up digitizing, data conversion, and photogrammetry.

Geographic information is required by a wealth of scientific research for various disciplines. Due to much progress of geospatial technologies in recent years, acquisition of high-quality spatial and temporal information has become much more efficient and cost effective than past few decades. Remote sensing provides massive high-resolution imageries about Earth surface, which can be analyzed by image processing tools to automatically derive valuable information for various applications such as climate change, resources inventory, environmental monitoring, and urban sprawl. The advent of GPS has revolutionized the process of surveying and allows users to directly measure coordinates of location very accurately and rapidly. Integrated with IMU (inertial measurement unit) and other sensors, GPS can also be used for mobile mapping when mounted on mobile platforms such as bicycles, cars, boats, aircrafts, and UAS (unmanned aerial systems) [7].

The emergence of Internet and Web provides an exceptional base for incubating new technologies for disseminating geographic information. Rapid development of ICT (information and communications technology) in recent years has enabled Web

J.-F. Jan (✉)
Department of Land Economics, Taiwan Institute for Governance and Communication Research, National Chengchi University, Taipei, Taiwan
e-mail: jfjan@nccu.edu.tw



S.-H. Chen (ed.), *Big Data in Computational Social Science and Humanities*, Computational Social Sciences, https://doi.org/10.1007/978-3-319-95465-3_2

users with all sorts of tools to access information stored on a server, and even construct a website to let obscure users create, assemble, edit, and disseminate information with little or no restrictions on the content. Using GPS enabled mobile devices such as smart phones, tablet PC, digital cameras, and vehicles mounted with GPS, or sensors for capturing environmental data that are carried on body, almost anyone can be a mobile sensor for collecting geographic information, whether a young child or a field scientist with highly developed skills. Combining Web tools, this type of geographic data can be disseminated voluntarily by individuals. Goodchild [6, 7, 8] coined this as volunteered geographic information (VGI). Wikimapia and OpenStreetMap both are compelling examples of VGI. They are collaborative mapping projects that encourage general public to participate in describing geographical objects in the world, and provide free geospatial data for anyone to use and share [6, 7].

Traditionally, scientific research is usually conducted by professional scientists of government, private companies, academic, and research institutes. Through carefully designed experimental process, the research results are generally more reliable, however, the scales of research are often constrained by available resources such as manpower, equipment, budget, and time. In comparison, citizen science, also known as civic science, is a type of scientific research that links general public with professional scientists to conduct research at multiple stages, which may include identifying research questions, collecting data, developing research methods, analyzing data, and interpreting results [4].

Large-scale researches often require vast amount of data to be collected at wide range of locations covering large geographic areas, and the research processes may span multiple time periods or even a few decades. Started in December 1900, the renowned Audubon's Christmas Bird Count (CBC) is a pioneer project that exemplifies the concept of citizen science in the field. A total of 27 observers participated in the first CBC at 25 locations in the USA and Canada. Over the years, the CBC has engaged many volunteers from different countries to participate in counting birds on Christmas. During the 116th CBC season (14th December 2015 to 5th January 2016), 76,669 observers contributed their efforts, and resulted in observation of 58,878,071 birds, and 2607 species (Fig. 2.1) [11, 12]. The CBC data collected by volunteered participants provide very valuable information not only for scientific research but also for decision-making on conservation strategies by government agencies. Based on the analysis of the CBC data, currently more than 200 peer-reviewed articles have been published [13].

Recent publications have shown that citizen science has been applied in a variety of different fields and has been accepted by the scientific community [1, 2, 5]. In this chapter, we introduce the application of VGI in spatial humanities using a case study of community resources investigation in rural Tainan, Taiwan.


**Fig. 2.1** Locations of the 116th Audubon Christmas Bird Count [12]

## 2.2 Materials and Methods

### *2.2.1 Study Site*

The study site of this research is a rural village called Chi-Shi Community, which is a part of Da-Nei District, Tainan, located in southern Taiwan (Fig. 2.2). With a population of less than 800 people, the community is situated among rolling hills, and most of the residents are engaged in agricultural production. Fruit production is the major source of income for the community, which is best known for a variety of fruits including papaya, mango, guava, orange, banana, and avocado. More than 20% of the population are of age 65 and above, therefore, health care for old people and lacking young work force are both important issues for long-term development of the community. The ancestry of residents mostly came from Fujian Province of China, and inhabited this area since the early Qing dynasty.

### *2.2.2 Volunteered Geographic Information (VGI)*

The Chi-Shi Community Association (CSCA), established by volunteers from the community, had initiated several projects aiming to protect the natural environment and preserve valuable cultural resources of the community. In addition to protecting these resources, the community also want to promote environmental education and cultural tourism. Collaborating with the CSCA, this study employed the VGI (volunteered geographic information) method to collect data about various natural

**Fig. 2.2** Location of the study site

and cultural resources of the community, including the distribution of natural resources, cultural highlights, scenic views, fauna and flora, and walking trails. The community residents were equipped with a GPS logger and a digital camera for recording GPS coordinates and images as they traveled along the trails. The GPS logger has a built-in memory capable of storing up to 250,000 waypoints recordings, and can be configured to record data at different sampling rate, which is convenient for different traveling methods including on foot, with a bike, and driving a car (Fig. 2.3).

The GPS waypoints and digital photos were then matched by using a phototagger software to automatically produce a map showing where the pictures were taken on Google Maps. The matching process is based on the image acquisition time and GPS time, therefore, the clock of digital camera must be set before taking pictures. When there is a match for photo and GPS waypoint, the GPS coordinates can be written to the image file permanently. Because the GPS coordinates are accurate within 10 m, the locations of the recorded points may appear to be incorrect sometimes. When needed, the user can modify the points and add annotation to the data (Figs. 2.4 and 2.5). The software can work with photos obtained using any type of digital cameras as long as their clocks are accurate. This is particularly useful when a group of people are working together with only one GPS logger.

The software can also convert the data into standard Google Earth KML/KMZ format for further processing or distribution of the field data (Fig. 2.6). Google Earth provides very user friendly tools for documenting the trips, including texts, symbols,

**Fig. 2.3** The GPS logger used in the study

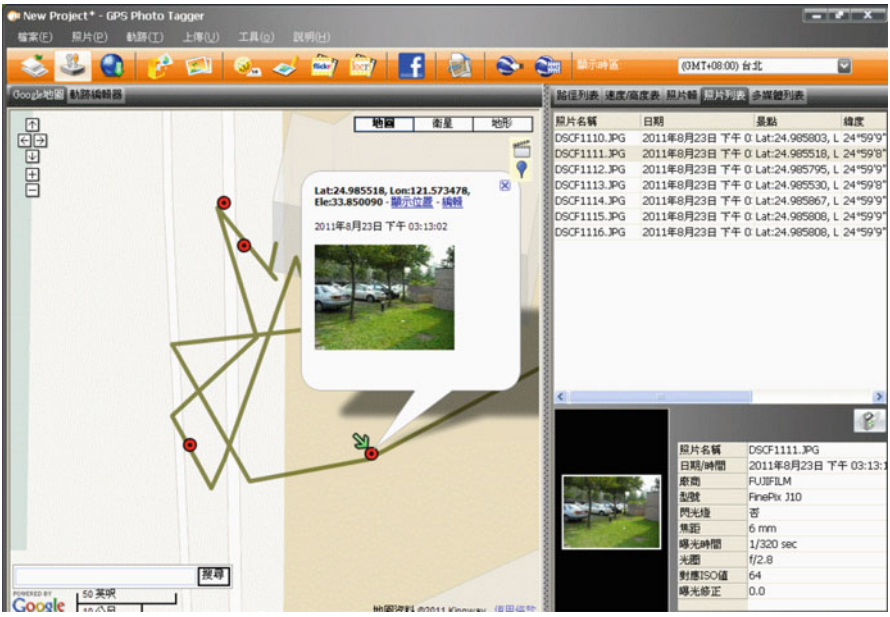



**Fig. 2.4** Automatic matching of waypoints and photos using phototagger software


**Fig. 2.5** Modifying position of photos, and adding annotation

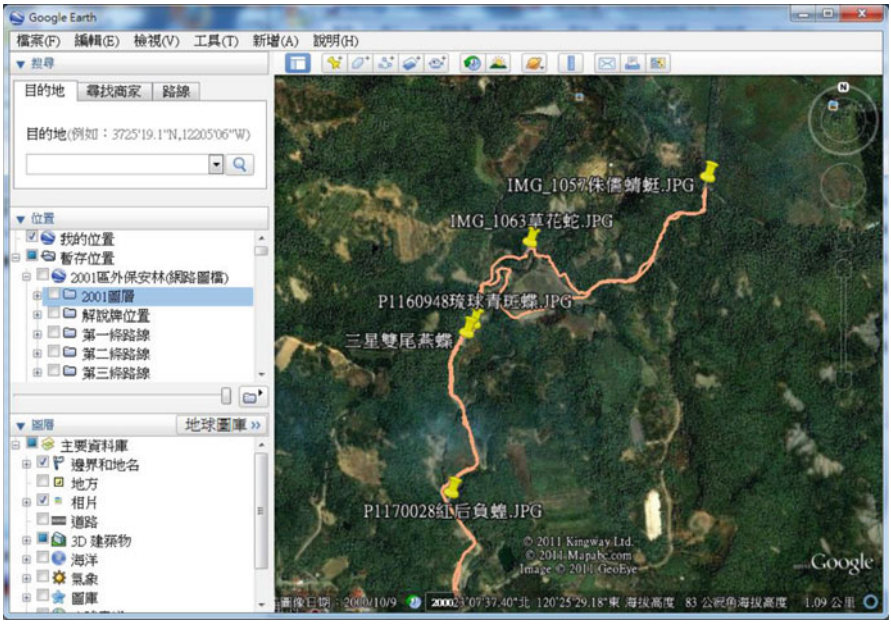



**Fig. 2.6** GPS waypoints and photos shown on Google Earth, and field data obtained at different trips are grouped into layers and folders

colors, and line styles. Moreover, with Google Earth we can group field data into folders and layers, and overlay multiple layers on the virtual Earth. This feature is very useful for managing field data, and showing historic data obtained at different time periods. Besides the community residents, visitors can also share their traveling experiences by submitting waypoints and photos to the website established for this study. By incorporating the volunteered information contributed by community residents and visitors, we can build a database with relatively low costs and high efficiency.

### *2.2.3 Building a Web-Based GIS for Community Resource Management*

GIS has been proven an effective tool for a variety of applications in wide range of disciplines. Particularly, GIS techniques are very useful for mapping and planning for applications on spatial humanities, tourism and recreation, and forest conservation and natural resource management [9, 10, 17, 20]. This study aims to develop a web-based GIS (WebGIS) platform that can be used by communities for resources management and tourism planning. Considering the system extensibility, portability, and costs for development, we used open-source software to develop an array of tools for data processing, database management, querying, display, and analyses. Specifically, the open-source software used in the study included Python, QGIS (Quantum GIS), OpenLayers, PostgreSQL, PostGIS, and Django [3, 16, 14, 15, 18, 19].

- Python: created by Guido van Rossum, it is a very popular and powerful general-purpose programming language that can be used to develop many kinds of software applications for all major operating systems.
- QGIS: written in C++, Python, and Qt, it is a cross-platform open-source desktop GIS that supports numerous vector and raster data formats, and provides many powerful processing and analysis tools.
- OpenLayers: an open-source JavaScript library designed for web mapping, it provides an API (Application Programming Interface) for building web-based GIS applications similar to the Google Maps API.
- PostgreSQL: written in C, it is an open-source object-relational database management system (ORDBMS) that runs on Microsoft Windows, MacOS, Linux, BSD, and Solaris.
- PostGIS: an open-source software that extends the PostgreSQL, it enables PostgreSQL to store geographic objects into geospatial database, and provides spatial analysis functionalities.
- Django: written in Python, it is a web framework designed to facilitate rapid development of web applications such as content management system (CMS).

## 2.3 Results

### *2.3.1 Community Resource Database*

The Chi-Shi community had participated in the Community Forestry Initiative of Taiwan Forestry Bureau for several years. Major income of the community residents are from selling agricultural produce such as papaya, guava, mango, and avocado. Besides the daily farming works, the community residents also help the Forestry Bureau by patrolling protection forests to prevent wildfire, dumping of industrial or domestic wastes and toxic chemical wastes, and illegal logging.

In recent years, the community residents formed CSCA to promote ecotourism in order to preserve environmental resources as well as increase income for the residents. The CSCA had initiated numerous programs to investigate the resources of the community, including plants, animals, cultural heritage, and scenic landscape features. To raise public awareness of invasive alien species and their impacts to the environment, the CSCA had held several activities and workshops, from which many residents and young students learned the importance of environmental protection and measures to remove invasive alien species. The CSCA also recruited volunteers to restore old houses and historic cultural relics. By using GPS logger and smart mobile devices, while the residents only had little training, the data collection and manipulation procedure was simplified, and the quality of data was found quite acceptable. The field data were processed using a phototagger software to produce maps showing the distribution of various resources. The software uses Google Maps as the base map, and it provides utilities for editing data, adding descriptions, and converting data to Google Earth KMZ format (Figs. 2.7, 2.8, 2.9, and 2.10).

By using the PostgreSQL/PostGIS software, a geospatial database of the community resources was established in order to support resources management and tourism planning for the CSCA. An array of software tools written in Python was developed to facilitate image processing, and importing field data and user-contributed data into the PostgreSQL database. QGIS (Quantum GIS) was used in this study for editing field data, adding attribute data, overlay analysis, and producing various thematic maps. Unlike commercial GIS software such as ArcGIS and Mapinfo, QGIS is an open-source software freely available for all purposes. We chose QGIS not only because of budget limitation but also it provides many features, tools, and flexibility for customization and extending its capability for the purpose of this research. In addition to the field data of community resources, the study collected a variety of spatial data such as administration boundary maps, river and basin, roads, compartments and working circles of Forestry Bureau, protection forest, topographic maps, land-use maps, and orthoimages. All of the data comprise the geodatabase for the community, and the PostgreSQL software was used to store the geospatial database (Figs. 2.11 and 2.12).

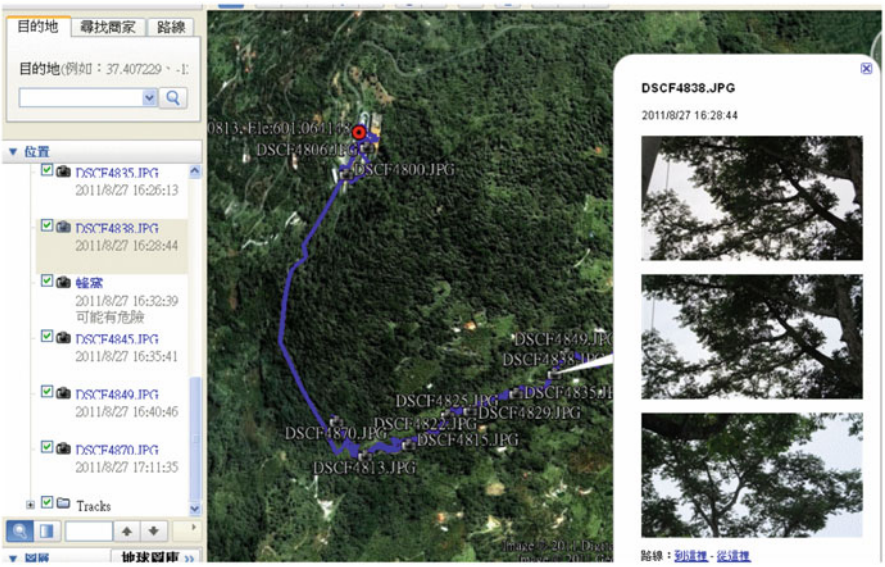



**Fig. 2.7** A field trip shown on Google Earth

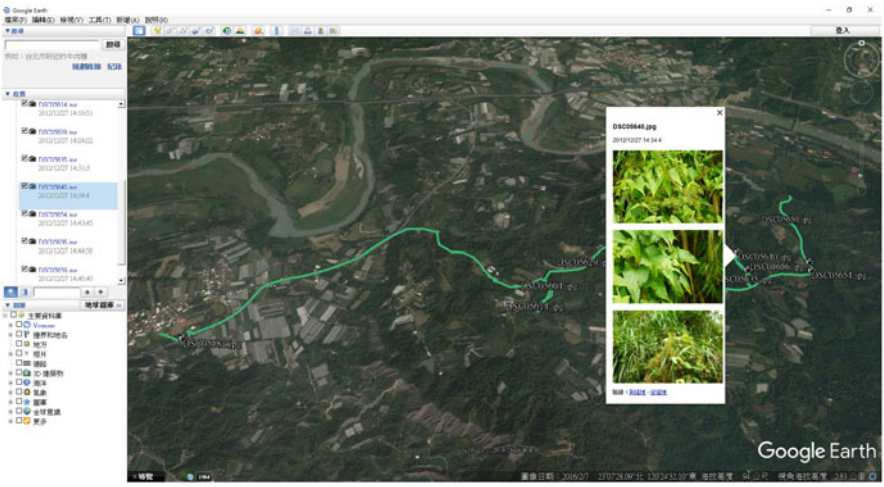



**Fig. 2.8** Invasive species (*Mikania micrantha*) investigation

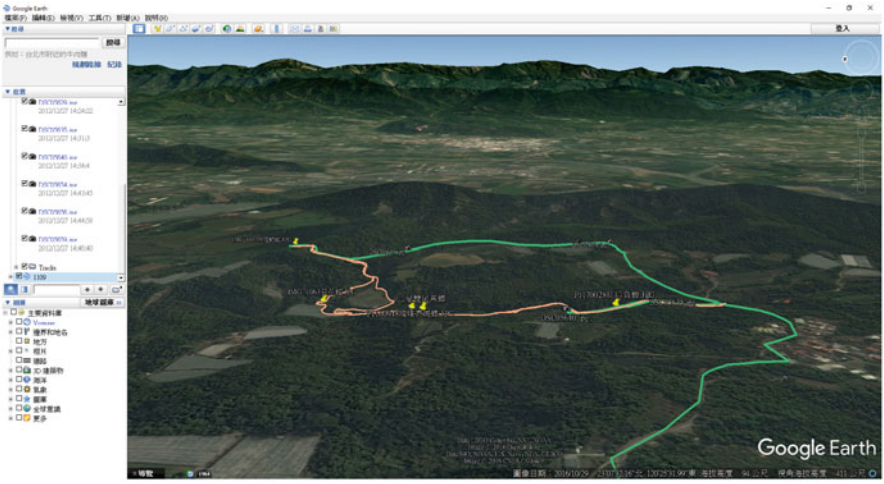



**Fig. 2.9** 3D view of field environment

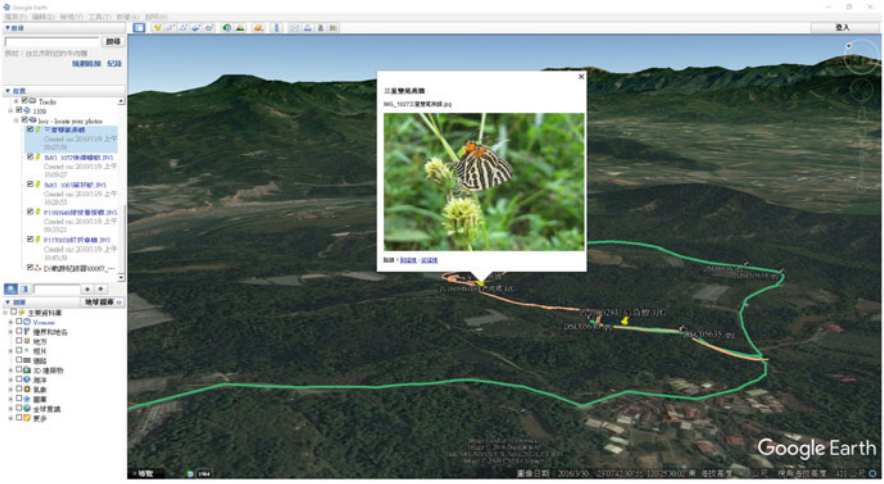



**Fig. 2.10** Resources data shown on Google Earth

### *2.3.2 Creating Virtual Tour Using Google Earth*

After adding geographic data in Google Earth, users can use Picasa or other software to embed digital photos in the map. Combined with the coordinate data collected by using GPS tracking device, each scenic spot can be marked in Google Earth and displayed with corresponding photos. In addition, after discussing with the local residents and planners of CSCA, Google Earth was used to assist in planning the most appropriate travel routes for tourists. Several travel routes for

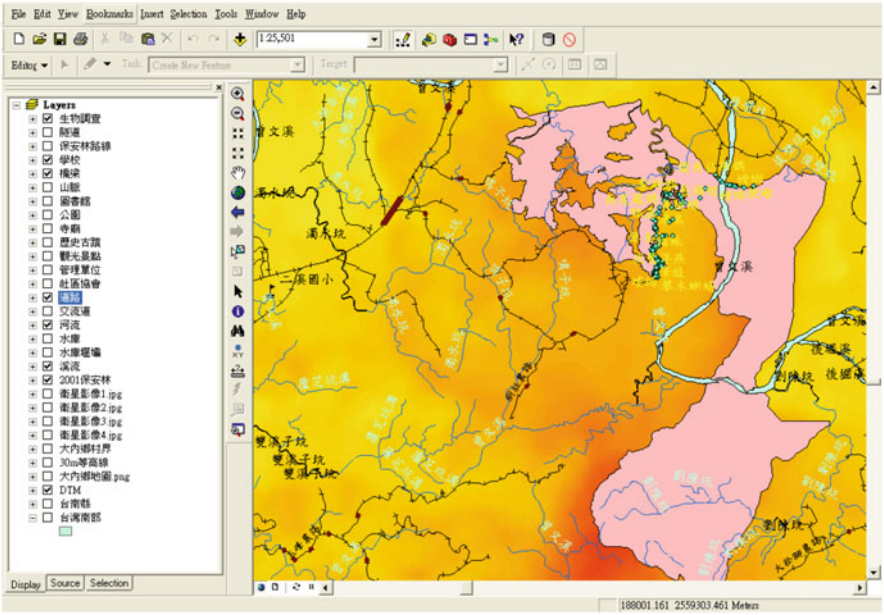



**Fig. 2.11** Topographic maps and thematic layers of the study site shown on QGIS

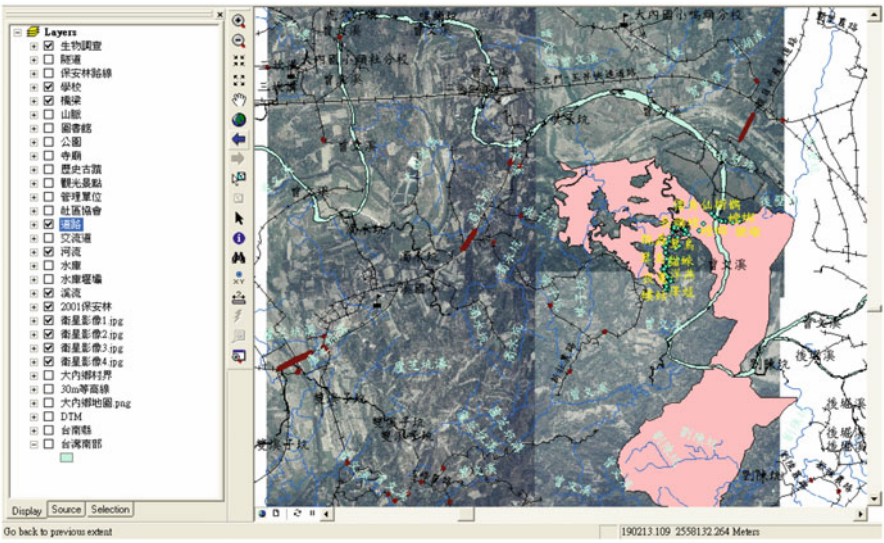



**Fig. 2.12** Thematic layers and orthoimages of the study site shown on QGIS

half-day, one-day, and two-day tours were planned. These tours are designed to let visitors experience different aspects of the community, including natural landscape, vegetation, insects, birds, historic buildings and architecture, orchard, and how the farmers plant fruits. Based on these travel routes, the tourism planner can create a simple animation by using Google Earths video recorder function. With Google Earths functionalities of viewing scenes from different directions, view angles, and altitude, the locations and photos of various feature attractions can be used to produce movies of virtual tours for these travel routes. The virtual tours are useful for the community to attract more tourists. On the other hand, the virtual tours are also helpful for the tourists to plan the trips (Fig. 2.13).

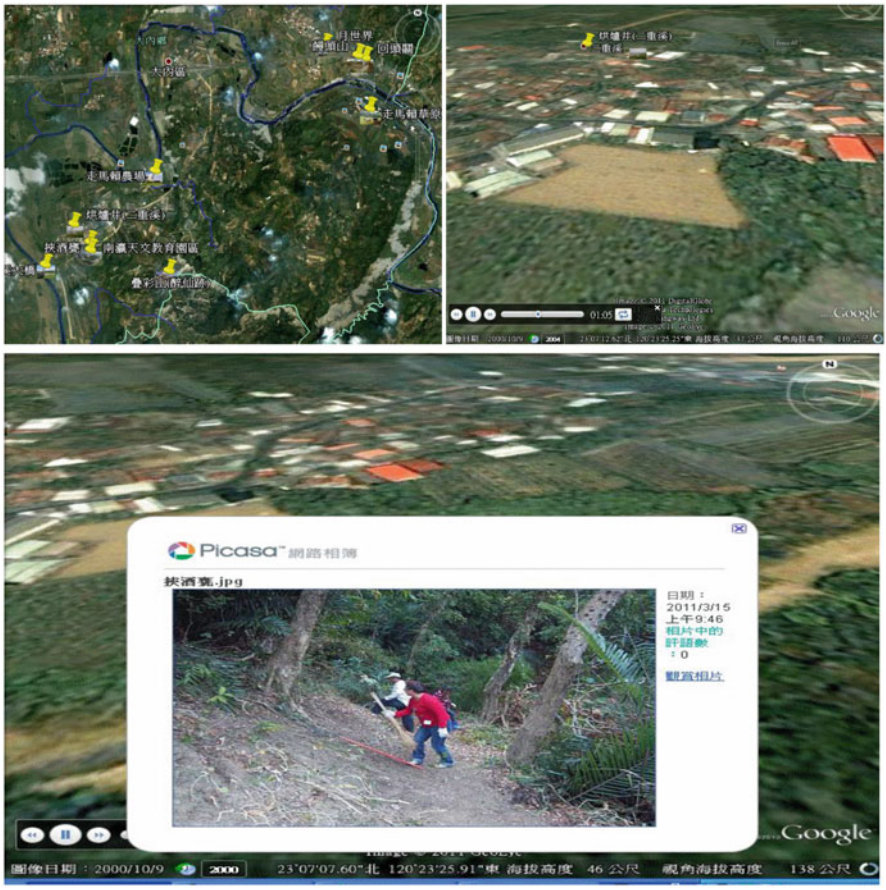



**Fig. 2.13** Video recording of virtual tour using Google Earth

### *2.3.3 Establishing the Web-Based GIS Platform*

In view of the convenience and popularity of the Internet, this study hopes to convey useful information through web pages in order to promote tourism for the Chi-Shi community. Therefore, we built a website using the Django web framework along with Apache web server, and PostgreSQL database server. The website management system allows the administrator of the website to manage and publish web pages dynamically. Moreover, the JavaScript language and API (Application Programming Interface) of Google Maps and OpenLayers were used to incorporate dynamic maps into the web pages. The Google Maps and OpenLayers API provides a variety of tools for manipulating maps using regular web browsers such as Microsoft Internet Explorer, Google Chrome, MacOS Safari, and Firefox. In addition, the website provides a route planning tool by using the Google Maps API, and a searching tool so that users can inquire data more efficiently. Thereby, the tourists can plan ahead before visiting the community. If the visitors want to share their traveling experiences, they can submit Google Earth KML (keyhole markup language) or KMZ files to the system through a web page. To avoid erroneous data and insure data quality, the system employs a double-checking mechanism to validate the data submitted by tourists. This quality assurance process is done by volunteers of the community.

Through the website, the CSCA can publish web pages to introduce all kinds of resources and make announcements of special events and activities that may draw attention from tourists. The website also serves as a platform for exchanging ideas and opinions about important issues of the community development among the residents. In general, the website can incorporate volunteered geographic information contributed from both the community residents and tourists. It also facilitates public participation in community resources inventory, as well as planning and decision-making on future development of the community. Besides, a section of the web pages introduces a range of cultural heritages, the process of growing various fruits, natural resources and unique landscape features, invasive alien species, and natural conservation concepts, which are valuable for environmental education (Fig. 2.14).

## 2.4 Discussion

The GPS logger used in this study is easy to operate and very cost effective. Some more advanced GPS loggers are waterproof, and the built-in battery allows the device to function continuously more than 2 days when fully charged. The GPS data can be used to match with photos obtained using any type of digital cameras. However, a GPS logger can only record coordinates of waypoints and lack of functionality for displaying measurements at the time of data collection. All the raw data have to be processed and manipulated after the field works are finished. Therefore, it is impossible to find if there are defects in the GPS device or poor


**Fig. 2.14** The web-based GIS platform of the Chi-Shi Community. (**a**) Website of the community, (**b**) Humanity resources of the community, (**c**) Search tool for the resources, (**d**) Result of resource inquiry, (**e**) Detailed information of resource, (**f**) Community resource management system

data quality while working on the fields. On the other hand, smart mobile devices such as tablet PC, PDA (personal digital assistant), and smart phones are capable of showing maps and much more information about the field data. Nevertheless, the costs for procuring such devices and the complexity of using these devices in rural environment may cause more limitations than benefits for this type of application.

Currently, most applications that employ embedded mapping techniques in web pages use JavaScript and Google Maps/Earth API to interact with the Google Maps/Earth server. The advantages of this method include fast response time, global coverage of the earth, free or low cost, and ease of use. However, the users of the free version of Google Maps/Earth are limited by accessible data volume, low image quality and update frequency in remote areas, availability of the other map layers, and lack of spatial analysis functionalities. This study demonstrates that open-source software can be used to develop web-based GIS application systems that incorporate the Google Maps/Earth and OpenLayers. However, it requires more training and software programming skills in order to harness the power of these tools as compared to using off-the-shelf commercial software packages.

## 2.5 Conclusions and Recommendations

This study utilized open-source software tools to develop a WebGIS (web-based GIS) platform for a rural community in Taiwan. The WebGIS in an integrating system and analysis tool for managing community resources inventory data collected by volunteered residents of the community and tourists. It appears that VGI (volunteered geographic information) is a valuable source for data collection; however, it is recommended to employ validation measures to avoid erroneous data and ensure data quality.

The WebGIS is a useful tool for the community to draw more attentions from tourists. By providing convenient tools for query and mapping service through the Internet, the system encourages more people to participate in the process of data collection, thereby the geospatial database of the community resources can be established more thoroughly and efficiently. Additionally, the WebGIS can serve as a platform for exchanging ideas and opinions about important issues of the community development among the residents. Consequently, we conclude that a well-designed WebGIS can facilitate public participation in community resources inventory, as well as planning and decision-making on future development of the community.

**Acknowledgements** This work was funded by the Forestry Bureau of Council of Agriculture at Taiwan. The author would like to thank their support and provision of various maps and images of the study site. The author greatly appreciates the members of Chi-Shi Community Association for providing assistance in website development and field investigation, and numerous individuals of the community as well as anonymous tourists for their contribution on data collection and suggestions to the Website.

## References


1. Causer, T., & Wallace, V. (2012). Building a volunteer community: Results and findings from Transcribe Bentham. *Digital Humanities Quarterly, 6*(2), 1–28. http://www.digitalhumanities.org/dhq/vol/6/2/000125/000125.html
2. Cronje, R., Rohlinger, S., Crall, A., & Newman, G. (2011). Does participation in citizen science improve scientific literacy? A study to compare assessment methods. *Applied Environmental Education and Communication, 10*(3), 135–145. https://doi.org/10.1080/1533015X.2011.603611
3. Django Software Foundation. (2016). Django website. https://www.djangoproject.com/
4. European Commission. (2013). *Green Paper on Citizen Science for Europe: Towards a society of empowered citizens and enhanced research*. Published on January 21, 2014. Retrieved November 12, 2016, from https://ec.europa.eu/digital-single-market/en/news/green-paper-citizen-science-europe-towards-society-empowered-citizens-and-enhanced-research
5. Follett, R., & Strezov, V. (2015). An analysis of citizen science based research: Usage and publication patterns. *PLoS One, 10*(11), e0143687.
6. Goodchild, M. F. (2007). Citizens as sensors: The world of volunteered geography. *GeoJournal, 69*(4), 211–221.
7. Goodchild, M. F. (2007). Citizens as voluntary sensors: Spatial data infrastructure in the world of Web 2.0. *International Journal of Spatial Data Infrastructures Research, 2*, 24–32.
8. Goodchild, M. F., & Janelle, D. G. (2010). Toward critical spatial thinking in the social sciences and humanities. *GeoJournal, 75*(1), 3–13.
9. Harris, T. (2009). *Conceptualizing the Spatial Humanities and Humanities GIS*. Keynote Presentation at the GIS in the Humanities and Social Sciences International Conference, October 7th–9th, 2009. Taipei, Taiwan: Research Center for Humanities and Social Sciences/Academia Sinica.
10. Lau, G., & McKercher, B. (2006). Understanding tourist movement patterns in a destination: A GIS approach. *Tourism and Hospitality Research, 7*(1), 39–49.
11. National Audubon Society. (2016). The 116th Christmas Bird Count Summary. Retrieved November 26, 2016, from http://www.audubon.org/news/the-116th-christmas-bird-count-summary
12. National Audubon Society. (2016). The Audubons Mapping Clearing-house: The 116th Audubon Christmas Bird Count. Retrieved November 26, 2016, from http://audubon.maps.arcgis.com/home/webmap/viewer.html?webmap=8c236c48814141d1bafd92242750079d
13. National Audubon Society. (2016). Christmas Bird Count Bibliography. Retrieved November 26, 2016, from http://www.audubon.org/christmas-bird-count-bibliography
14. Open Source Geospatial Foundation (OSGeo). (2016). PostGIS website. http://www.postgis.net/
15. Open Source Geospatial Foundation (OSGeo). (2016). QGIS (Quantum GIS) website. http://www.qgis.org/en/site/
16. OpenLayers website. (2016). https://openlayers.org/
17. Orellana, D., Bregt, A. K., Ligtenberg, A., & Wachowicz, M. (2012). Exploring visitor movement patterns in natural recreational areas. *Tourism Management, 33*(3), 672–682.
18. PostgreSQL Global Development Group. (2016). PostgreSQL website. https://www.postgresql.org/
19. Python Software Foundation. (2016). Python website. https://www.python.org/
20. Theobald, D. M., Norman, J. B., & Newman, P. (2012). Estimating visitor use of protected areas by modeling accessibility: A case study in Rocky Mountain National Park, Colorado. *Journal of Conservation Planning, 6*, 1–20.

# Chapter 3
# Telling Stories Through R: Geo-Temporal Mappings of Epigraphic Practices on Penghu


**Oliver Streiter**

## 3.1 It Takes a Story to Catch a Fish

Fishermen on Penghu, like all over the world, sit endless hours at the pier to chat and mend their nets. Stitching together piece by piece, they shape the tool that allows them to engage with their environment beyond the limits of their eyes and hands. Their stories and languages, woven centuries ago by their ancestors, are spun from the same yarn. These linguistic structures help to catch an environment in categories and causal relations based on the experience of generations and not accessible to immediate perception (Harrison 2007).

Scientists likewise have to acquire and culture a craftsmanship, which involves skills, tools, and a language. Skills are needed to operate tools and languages translate the outcome of the operation into meaningful stories that offer new perspectives on the world. Scientific endeavors thus frequently involved quests for new tools. Once developed, tools such as the telescope of Galileo (Brown 1985) or the British National Corpus (Leech 1993), have pushed scientists to perceive the old world in unpronounceable forms and shades, which then had to be cast into new languages, terms, and stories, such as *earth orbit* or *collocate* to explain, consolidate, classify, trigger, and transmit perceptions and insights.

The research presented here has been conducted in the framework of the the *ThakBong* project, a project in the Digital Humanities, which was started in April 2007 with the general intention to document and research funerary and epigraphic

O. Streiter (✉)
National University of Kaohsiung, Kaohsiung City, Taiwan
e-mail: ostreiter@nuk.edu.tw



S.-H. Chen (ed.), *Big Data in Computational Social Science and Humanities*,
Computational Social Sciences, https://doi.org/10.1007/978-3-319-95465-3_3

practices on Taiwan (Streiter et al. 2011), combining ideas and theories from different disciplines in the humanities and social sciences, such as linguistics, anthropology, sociology, and archaeology.

In the course of this long-term effort, 618 gravesites have been documented on the Taiwan archipelago, capturing and annotating 279,423 photos of 61,351 tombs.

Our craftsmanship is based on easy knitting patterns. Setting out on field trips in rhythms that follow the climate and local calendars, we visit burial grounds, clean tombstones, and harvest digital, geo-tagged photographic images. At home in our lab, our primary data are post-processed, annotated, and transcribed. Then, our catch, like processed fish, is frozen on mirrored hard drives to be preserved for the decades to come. Continuing this lossless accumulation of data, we weave a net that continues to capture unexpected phenomena whenever the size of the net increases or its meshes get smaller. New catch may include historical events like flooding, epidemics, or waves of migration.

The catch we get by casting our net over local graveyards can be read as a newspaper of regions where no reporter showed up, as a population census where no official kept records, as a *who is who* of a village where now buildings scrape the sky and as a map of a deadly battle field where people struggle over ethnicity, nationality, religion, space, symbols, and the meaning of life. Having these unwritten documents in one's hand is a truly exciting, awe-inspiring, and humbling experience. It is worth for us to get dirty hands during fieldwork and for the reader to get sweaty hands on a keyboard, exploring data and tools.

The tool we use to explore our data is the *R programming language* (R Development Core Team 2008). *R* is one of the Swiss army knifes that is popular with digital humanists. As a programming language, *R* is slightly different from some of its competitors, such as Python, and might thus provide an alternative gateway into the world of coding.[1] R is free and rich in its features. Its graphics are awesome and yet it is easy to start with. To install R on your computer, visit the Comprehensive R Archive Network web pages at https://cran.r-project.org/ and to follow instructions given there.

A second tool we introduce implicitly is *KnitR*. Through *KnitR*, one can include R commands in one's *latex* or *markdown* text to produce tables and graphics on the fly, tapping dynamically into data stored in a spreadsheet file or a database. This is what we will attempt in this paper. Graphics and tables are produced by the commands that you see within this text.

All the catch we managed to freeze in digital format has been prepared and packaged to be loaded onto your computer. A first package, small and easy to handle, contains the data: index numbers, transcriptions, and classifications. A second package, huge in size, contains hundreds of thousands media files and can be downloaded for browsing tombs and tombstones in search for inspiration, a confirmation, an example, or an explanation, in other words, something you assume

[1] Loops can frequently be avoided in R and a programming code consists of a sequence of function calls, the outcome of which is assigned to variables with meaningful names.

you cannot find in the bare data. Data and media files have been archived and can be accessed at *DANS*,[2] at http://dx.doi.org/10.17026/dans-zvy-rtju and http://dx.doi.org/10.17026/, respectively. In addition to these packages, a third package has been put at your disposal, containing programming code that has been written with the intention of facilitating the access to these funerary and epigraphic data, preparing them to be fed into standard analysis modules. The first and the third package can be downloaded from http://thakbong.dyndns.tv/R.

As a first exercise on *R*, we will download the first and third package, into an existing folder on your computer.[3] R-code like the code above reads usually from right to left. A value is created on the right side and assigned, through the assignment operator `<-`, to a variable on the left side of the operator `<-`, e.g., `source.dir`.[4]

```
source.dir <- '/home/oliver/R'
source('http://thakbong.dyndns.tv/R/ThakBongRHTTP.R')
```

After running these commands in *R*, you will notice that the folder you have specified has been populated with numerous files. The `.R` files are the functions, the `.gz` files are the data in a compressed format.

The data package is subdivided into different cans, each of them formatted as compressed comma separated value (CSV) file, uncompressed by R the moment they are read in. Each file contains one data frame, representing each one entity type, such as *tombstone, tomb, graveyard*, etc. To read them into R and connect them in the right way, you best use some functions we prepared in the third package of R code.

After downloading helper functions and data onto your computer, you can load them into your *R*-session as shown in the following code snippet.[5]

[2]The *Data Archiving and Networked Services* is an institute of the *Royal Netherlands Academy of Arts and Sciences (KNAW)* and *Netherlands Organization for Scientific Research (NWO)*.

[3]In Line 1, you specify a folder on your computer that should contain the data and functions. If not existing, you have to create that folder before. In Line 2, you start the download into this folder.

[4]R knows also a right assignment `->`, allowing to write lines from left to right. As most and earliest programming languages used the left assignment, most people are used to read and write code from right to left. With the right-assignment however, one has in R the opportunity to write highly readable code in the style of UNIX pipes using the `magrittr` package.

[5]Line 1: You specify the folder into which you downloaded the ThakBong data. Line 2: All functions and commonly used data are loaded. You can always start your analysis of ThakBong data in *R* with these two lines, unless you want to update functions and data from the online repository.

```
source.dir <- '/home/oliver/R'
source(file.path(source.dir,'ThakBongRLocal.R'))
```

Once the material loaded from the local folder into your *R*-session, variables and functions become available to you. A central variable is `df.thakbong`, a variable that contains all information on all graveyards, tombs, tombstones, and tombstone inscriptions, merged into one big `dataframe`.[6] Function you might apply to this `dataframe` are `get.stone.form`, `get.stone.height`, `get.stone.width`, `get.stone.material`, `get.tomb.altitude`, or `get.tomb.direction` which, as you might expect, extract the requested values from a `dataframe` you feed into that function as parameter.

Filtering the data frame is a central procedure that prepares every analysis. Helper functions of the type `has` or `is` operate as filters that retain relevant rows. In the example code below, we retain only graveyards located in Asia, the region we are going to focus on.

```
df.asia <- is.asia(df.thakbong)
df.penghu <- is.penghu(df.asia)
df.penghu.year <- stone.has.creation.year(df.penghu)
remove(df.thakbong)
```

Much of the success of your analysis will rely on how you segment the data. If you compare groups that have nothing in common, little follows from your comparison. Likewise, if you compare groups that are internally very diverse, you do not know which attribute of the group is responsible for the difference between groups. In the ideal case, you create a homogeneous group that distinguishes itself through well-defined features from the comparison groups. Groups may differ according to time, space, surname, or position in a social network. To create these groups, R puts at your disposal two important techniques. The first allows

[6] A `data frame` stores data tables, similar to data tables in a relational database. `df.thakbong[1,]` returns the first row, `df.thakbong[,1]` the first column, and `df.thakbong[1,1]` the first element in the first column. The values of one column are stored in a `vector`. A `vector` is a sequence of data elements of the same type, e.g., all elements must be numerical or textual. A vector can be created by combining different elements, as in `kids <- c('Mary','Bill')`. The data in this vector can be accessed, e.g., as in `kids[1]`, which would yield "Mary."

to split any dataframe into a list of smaller dataframes.[7] The second technique works on the output of the first and facilitates the run through all smaller dataframes, processing them one by one.

The splitting of a dataframe is done with the help of the split() command. The dataframe *graveyards*, for example, can be split into different sorts of *graveyards*, specified by any combination of names of columns by which you want to split. The following command splits *graveyards* according to the archipelago the graveyard resides on and the country code. The output is a list of dataframes. Adding drop=TRUE removes graveyards without reference to an archipelago or a country code.

```
df.archipelago <- is.archipelago(df.asia)
list.df.archipelago <- split(df.archipelago,list
   (df.archipelago$graveyard.archipelago,
df.archipelago$graveyard.country.code),drop=TRUE)
```

To run through such a list of dataframes, to extract or calculate some values, or to draw a graph or map for each of these dataframes, we use functions of the apply family.[8] Functions of the apply family achieve in R what other programming languages achieve through loops.[9] What would be the body of the loop can be expressed in a function which can then be repeatedly applied to a split dataframe.[10] Using this technique, we can, for example, count for each archipelago the number of tombs sampled in the ThakBong project (sum(x$number.of.tombs)), as exemplified in Fig. 3.1. For simple task, you can combine split and apply in the aggregate function which splits and applies a function with one command.

[7] A list is more flexible than a vector as it may contain any combination of simple and complex values. A list is created through the operator list as in age<-list(kids,c=(12,11)). The second (complex) element of this list can be accessed as age[[2]]. As age[[2]] is a vector, we can access the age of the first kid directly as age[[2]][1].

[8] The function lapply returns a list and sapply simplifies the output and returns a vector if possible. The mapply function applies a function repeatedly with different parameters for each application.

[9] Loops are a much debated design feature of computer languages. Hated by some and beloved by others, they are frequently believed to be hard to read and hard to be maintained. Although R offers the possibility to avoid loops in many cases, one is free to use loops of functions of the apply family.

[10] You can use an existing function in mapply or construct an unnamed function in sapply and lapply. This function is then applied to each of the split data frames. If you have to use an unnamed function but the functionality is given in available functions, you can call the existing function from within the unnamed function, as in function(x,y) YOURFUNCTION(x,y).

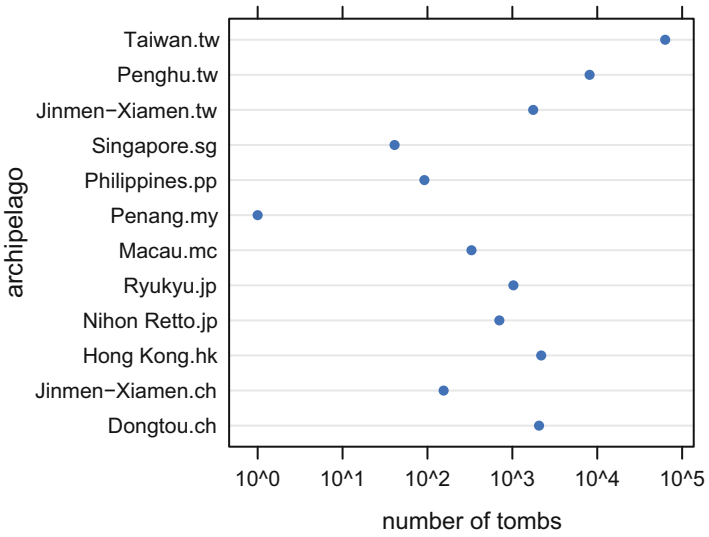



**Fig. 3.1** The number of tombs per archipelago surveyed within the ThakBong project

```
dotplot(sapply(list.df.archipelago,function(x)
   nrow(x)),ylab='archipelago',
xlab='number of tombs',scales=list(x=list(log=TRUE)))
```

## 3.2 From History to Code

Funerary and epigraphic practices vary from place to place and from time to time. When mapping these variations through space and time, we seldom see a random distribution. The variations that can be attested on Taiwan and Penghu burial grounds frequently relate to turning points in local political and social histories. These turning points, experienced by local people as states of crisis, caused funerary and epigraphic practices to be adjusted directly or indirectly, sometimes overnight, sometimes after one or two generations. In different regions, transformations may yield different forms, or we might say, find different solutions, even if driven by the same cause, depending on how local agents harmonize locally established practices with new political and social contexts. Either the local conditions before the crisis might have been different or local agents come up with new and creative ideas. Unfortunately, turning points in Taiwan's and Penghu's political history relate frequently to the rise of nationalism and colonialism in East Asia.

Back in the fifteenth and sixteenth century, sign boards like "Welcome to Taiwan" were certainly not installed along Taiwan's coastline to tell early explorers, pirates, and shipwrecked sailors where they had stepped ashore. Long time after Penghu had appeared in Chinese travel accounts[11] and on Chinese maps, Taiwan and its neighboring islands were still terra incognita to China and contemporary names used were of uncertain reference within a region that stretches from Okinawa to the Philippines. For this reason even today, after extensive research on wind directions, currents, vegetation, and landscape, one frequently cannot tell which islands travellers have actually set their foot on, basically because these travellers could not know themselves (Ptak 2015). Therefore, Ptak hypothesizes that commanders of expeditions might have systematically reported to have visited the islands they were supposed to visit to avoid any backlash.

Among the first written histories that can reliably be associated with the island that we call Taiwan today are those that come from Dutch and Spanish accounts of their patchwork colonization of the island in 1624 and 1626. The name *Taiwan* as used by the *Dutch Verenigde Oostindische Compagnie* (VOC) in the seventeenth century derives from a native Austronesian group's self-name. After 16 years of sharing the island, the Spaniards were ousted by the VOC, who in turn were besieged in 1662 by Koxinga (鄭成功 Zhèng Chénggòng), a merchant and swashbuckler loyal to the Chinese Ming Dynasty (明朝 Míngcháo). At that time, the Chinese government was under military pressure from its northern Jurchen neighbor, the Qing State (Manchurian: Daicing Gurun, Mandarin: 清朝 Qīngcháo), established in 1616 in Manchuria.

Many highly ranked Ming militaries recognized the weakness of the Chinese state and joined the new emerging colonial power. Not so Koxinga. After the Manchurian invasion of China and the suicide of the last Chinese emperor Chongzhen (崇禎 Chóngzhēn) in 1644, Koxinga and his family, surnamed Zheng (鄭 Zhèng), conquered Penghu and Taiwan in 1661 and maintained a Ming-loyal state for 20 years, until, in 1682, these islands also succumbed to the Manchu forces. The social transformation during the Zheng period were highly significant for Taiwan's social structure: The Chinese population on the island, under the Dutch at the bottom end of the social pyramid, obtained under the Zheng regime a social position that was higher than that of the Indigenous population, see Brown (2004), receiving material and legal benefits in an evolving Han-state that was threatened internally and externally by non-Han ethnicities.

Modernity started abruptly in 1895 with a Japanese surprise attack and the occupation of the Penghu archipelago, while still negotiating with Qing representatives in Shimonoseki. In the treaty of Shimonoseki, signed the same year after the occupation of Penghu, the Qing ceded both Penghu and Taiwan to the Japanese Empire, marking the beginning of a 51-year-long Japanese period during which Taiwan and Penghu were developed into a lucrative colony and a military outpost

[11] An early travel account on Penghu can be found in 汪大淵 Wáng Dàyuan (Wang Dayuan) (1339/1981).

for later attacks of the Imperial Army on Southeast Asia. As part of their critical role in the Japanese war machine, Taiwan and Penghu were bombed, similar to Okinawa, towards the end of World War II by the US military.

Unlike Okinawa however, the US military did not occupy Penghu and Taiwan after WWII and allowed instead Chiang Kai-shek (蔣介石), who had followed Sun Yat-sen as the head of the Chinese Nationalist Government, to install on Taiwan and Penghu the Republic of China, a one-party-state under the Kuomintang.

Although Taiwan has only a surface that corresponds to Switzerland, not all colonial regimes managed to rule over the entire territory of Taiwan, as Taiwan's Indigenous population fiercely defended their native lands. This limited control is especially true for the Dutch and the Spaniards, but also for the Zheng and Manchurian regimes. Would this extension of the regimes reflect in the distribution of tombs on Taiwan, which are essentially Han-tombs, in different time periods? An easy way to answer such a question is to draw with `R` a digital map.

The first step in drawing a map is to create a canvas. A canvas is a background map, which we can be repeatedly loaded as unpainted geo-referenced background for drawing geo-referenced lines and dots. Using maps provided by Google, we create and store two canvases for Penghu and Taiwan. Parameters to create these maps are the map type, the zoom level and the geo-references of the center of the map.[12] As long as we do not call a plot command, e.g., through `PlotOnStaticMap(canvas.penghu)` nothing visible happens to a canvas at this stage.

```
NEW=TRUE; if(file.exists('PenghuCanvas.png'))
   { NEW <- FALSE }
geo.penghu <- c(23.5,119.583333)
geo.taiwan <- c(23.69781,120.9605)
map.type <- 'mapmaker-roadmap' #or: roadmap,
   satellite, hybrid
canvas.penghu <- GetMap.bbox(destfile=
   'PenghuCanvas.png', zoom=10,
NEWMAP=NEW,maptype=map.type, center=geo.penghu)
canvas.taiwan <- GetMap.bbox(destfile=
   'TaiwanCanvas.png', zoom=8,
NEWMAP=NEW,maptype=map.type, center=geo.taiwan)
```

[12] As maps and geo-locations are loaded through the internet, this step requires your computer to be online.

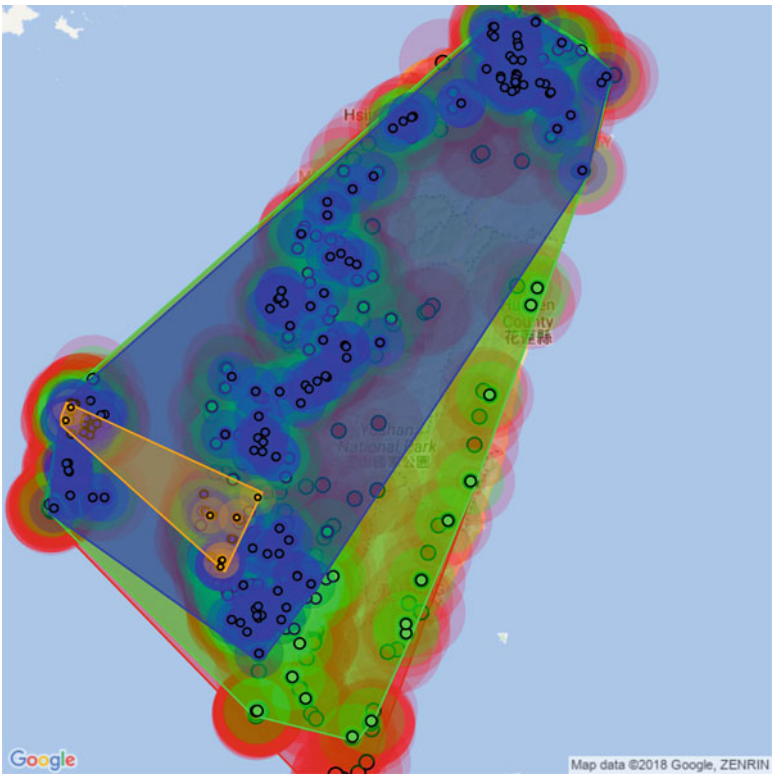



**Fig. 3.2** Tombs of the Zheng (orange), Qing (blue), Japanaese (green) and ROC (red) period on Penghu and Taiwan

Using our prepared canvases, we can draw a convex hull around all tombs of specific periods to approximate the distribution of tombs. The result is shown in Fig. 3.2.

```
map.period <- function(canvas,df,color,size) {
map.chull(canvas,df$x,df$y,col=color)
map.points(canvas,df$x,df$y,size,color)}
####
roc <- period.is.roc.roc(df.asia)
japan <- period.is.roc.japan(df.asia)
qing <- period.is.roc.qing(df.asia)
zheng <- period.is.roc.zheng(df.asia)
PlotOnStaticMap(canvas.taiwan)
```

(continued)

```
zero <- mapply(map.period,list(canvas.taiwan),
   list(roc,japan,qing,zheng),
c(color.roc,color.japan,color.qing,color.zheng),
   c(2,1.5,1.1,0.8))
```

As we can see, even though most tombstone of the Zheng area (orange) have been lost, their spatial distribution has been limited, owing to the fact that the Zheng family controlled mainly Penghu and the Tainan area. The Qing (blue) extended their control of territories, but not to the south, east, or the central mountain range. The Japanese forces (green) forcefully brought the entire island under their control, providing the ground for the first Han-settlers to establish farms and villages along the east-coast from 1910 on. Most tombs on these maps are Han tombs. Indigenous tombs appear on this map from the late Japanese period on, when the Austronesian communities, moved to lower locations and forced to bury in regular graveyards, started to use permanent grave markers. Most red dots with or without tiny green spots thus point to Indigenous communities.

Plotting data points on a map, as we did above, is a way of reporting surveyed data. Plotting a convex hull around data points (`map.chull`) is an act of generalization, an attempt to come up with a story the dispersed points cannot tell. Unfortunately, the convex hall we created glosses over a complex reality, as being based on closeness in terms of geo-references alone, not considering the nature of the terrain or the availability of transportation routes. The interpolation over the central mountain range, which separates Taiwan's east and west at an altitude of 3000 m, shows that this technique has to be handled with more delicacy. Striking the balance between reliably reporting data, on the one hand, and mapping an interesting story, on the other, remains a permanent struggle when drawing maps or timelines. Whenever values are rounded, lines smoothed or interpolated, the story should originate from that part of the graph where data and generalizations seem to converge.

## 3.3 Approaching Penghu

Doing fieldwork on Taiwan and analyzing its burial and epigraphic practices, we soon realized that our original spatial framing limited the ability to interpret the emergence, variation, and distribution of observed practices. We hypothesized that funerary and epigraphic practices outside Taiwan would hold important keys for understanding our observations on Taiwan (Streiter et al. 2010). Considering migration paths and colonial influences, the project thus extended systematically its scope. Burial grounds on Penghu, Jinmen and Mazu, Okinawa, Honshu, Zhejiang, Fujian, Hong Kong, and Macao have been documented, as well as graveyards in

places where migrants from Guangdong, Fujian, and Zhejiang have moved to in the USA, Europe, and Southeast Asia.

In this process, Penghu was studied intensively, as on our first few visits we found in most parts of the archipelago Qing and Japanese period tombs in larger quantities than on Taiwan. We hoped that this rich historical resource of epigraphic practices would allow us to identify influences on practices on Taiwan given the large-scale migration, including that of carvers, that took place in the twentieth century from Penghu to Taiwan. Yet, to our surprise, we found that Penghu was not a space of homogeneous cultural practices.

As our documentation proceeded from islands of island, we found on Penghu yet undocumented island-specific carving practices, unknown on Taiwan or regions in China we had surveyed. The formation or the preservation of these idiosyncratic practices had been facilitated by the natural fragmentation of Penghu into relatively autonomous islands, an isolation that might have ended not before the arrival of modernity in form of the Japanese colonial army and its systematic effort to construction harbors on islands that before could only be reached by rowing boats. Having thus found practices that matched those on Taiwan and practices unknown on Taiwan, we had to understand what distinguished these practices so that some were exported to Taiwan while others were not.

As a result, between 2010 and 2017 we undertook 19 field trips on Penghu documenting 77 burial grounds on 14 islands through 30,437 photos of 8154 tombs. Although these numbers seem to compare unfavorably to those of Taiwan, many small islands and many burial sites have been exhaustively documented. We thus estimate to have surveyed more than 70% of all existing tombs, which represents a larger coverage than we can ever achieve on Taiwan.

Penghu is an archipelago about 45 km to the west to Jiayi on the west coast of Taiwan and 185 km to the east of Shantou, China. China is closest when heading in the direction of Jinmen, from where most families on Penghu had migrated in the late Ming and early Qing period. The ferry from Jiayi takes 2 h to carry a swarm of tourist to the harbor of Makong. The cargo ferry from Kaohsiung travels overnight, leaving the traveller stranded on the pier among refrigerators, bags of rice, flowers, motorbikes, and huge piles of cement bags. Modern Makong lies at the periphery of Kaohsiung, at least in the world of goods and many family links. Already before entering the harbor of Makong, many of the 80 islands and islets of this archipelago can be spotted from the ferry. Military outpost and defensive structures are omnipresent, towering over harbors and cliffs. Most islands are lower than 40 m above the sea level. The highest peak can be found on Maoyu at 79 m (Tsao et al. 1999).

Most islands of Penghu came into existence millions of years ago, when basaltic lava, lava that has been formed by melting the earth's mantle, was poured into the cold ocean, where it cooled down relatively fast, contracted, and fractured, sometimes into hexagonal columns (Chen 1995, p. 452). When coral reefs settled on the uplifted basaltic structures, vegetation and humus started to develop wherever they could resist wind and waves. Depending on how much these structures have been raised above the sea level, coral stones, basalt, and, as lowest level, sedimentary

**Fig. 3.3** An eroding landscape on Dongyudongping. In front, a rubble field with remains of a former tomb, a layer of sedimentary stones, and above this a layer of basalt

sand or mud stone are visible and accessible to different degrees. Granite only occurs naturally on the most western island of Penghu, Huayu (花嶼, Huāyǔ), situated on the continental shelf. Much of Penghu's bedrock consists of softer sedimentary stones below the basalt. The basalt bedrock is weakened as this soft sedimentary stones are eroded, resulting in the collapse of basalt columns. This creates a rubble field along the coast line, a geologic landscape of strewn rocks has been used for burial grounds, e.g., on Dongji (東吉, Dōngjí), Dongyudongping (東嶼東平, Dōngyǔdōngpíng), Tongpan (桶盤嶼, Tǒngpányǔ), and Xiyu (西吉, Xīyǔ) (Fig. 3.3).

The harsher the living conditions of a region are, the more they are likely to influence the way that burial are conducted, and, consequently, how people think and talk about their practices. Deserts and permafrost, rocky ground and rain forest, each physical condition provides the framework to which cultural conceptions have to adapt physically, conceptually, and spiritually. Similar must have happened on Penghu, where a primordially continental and agrarian funerary practice and its conceptualization encountered water, wind, and rocks.

One of the most striking properties of tombs on Penghu is their dispersion. According to Chen (1953, 1995), an average of 10% of the total surface of Penghu is occupied by burial sites, a number that on smaller islands may reach up to 40%. Such a high percentage might come as a surprise, especially as Penghu is overpopulated with respect to its surface and, even more so, with respect to the fertility of the ground: Through erosion and salty rain, the agricultural production could nourish the population, statistically seen, for 4 months of a year only, see Chen (1953, 1995). A partial explanation for what seems to be an unproductive use of land might be that burial sites occupy the least profitable land, on which only some low pioneer plants, cacti, and agaves can survive. The smaller the island, the larger the

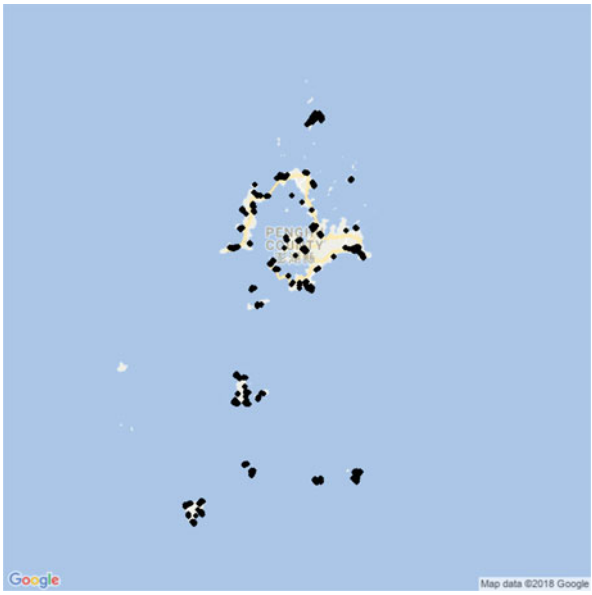



**Fig. 3.4** The spatial distribution of tombs surveyed on Penghu within the ThakBong project

percentage of coast-near, eroded, and flood-prone areas, which can be left to the ancestors without a real loss in agricultural production. Using our background map of Penghu, we can actually bring all surveyed tombs of Penghu to a map. Central in this process of mapping tombs is the translation of geo-references into pixels of the map. This is done by the function `LatLon2XY.centered()` which extracts all relevant parameters for this translation from the canvas that this function requires as parameter. Indeed, we see in Fig. 3.4 that smaller islands disappear under a cloud of tombs.

```
PlotOnStaticMap(canvas.penghu);
list.pix <- LatLon2XY.centered(canvas.penghu,
df.penghu$tomb.y,df.penghu$tomb.x)
points(list.pix$newX,list.pix$newY,pch=23,bg='black',
    cex=0.2)
```

The consequences of this dispersion of tombs are an objectively noticeable state of abandonment of tombs and an almost complete absence of social control over whether and how families maintain their tombs. We interpret this as an implicit agreement on these islands, which allows families, especially poor families, to let the tombs of a certain number of ancestors to be taken over by nature to avoid the financial burden that would require an exponentially growing number of ancestral

tombs. The dispersion, matched by the repeatedly told story of having forgotten the location of an ancestral tomb, allows for a wilful shaping of an ancestral line in accordance with one's financial possibilities, without having to articulate a violation of cultural codes. Dispersion is the local answer to the unrealistic requirement of an ancestral worship that did not originate on these islands but was imported in the cultural luggage of migrants. Considering the climatic conditions on these islands, we easily understand the gap between the cultural claim and reality of practices.

A feature that shaped Penghu as much as the waters around the islands are the winds of the winter monsoon, which hold the islands in their bitter grip. In the 6 months from October to March, the average wind speed in Makong is 17 m/s (Chen 1995, p. 453), see also (澎湖縣政府民政局 Pēnghúxiàn Zhèngfǔ Mínzhèngjú (Penghu County Civil Affairs Bureau) 2005, pp. 51–54). These winds cause a rough sea, irregular transportation between the islands, an interruption of the fishing period, massive erosion, salty rain through flying salt crystals, and an evaporation that exceeds the amount of rainfall (Chen 1995, p. 456). These winds and their physical consequences shape almost everything on the islands: Landscapes, vegetation, land utilization, agriculture, fishing, transportation, architecture, and the design of villages.

Typical for the architecture of Penghu, including tombs, is the extensive use of the resources that are locally accessible. Importing construction material not found on Penghu, such as wood, clay, or granite, was difficult and expensive as distances to both sides were long, currents dangerous and transportation facilities limited. A systematic net of harbors was cast over the islands as late as during the Japanese period, requiring during Qing period or earlier that each beam and each slate be reloaded from the cargo vessels to rowing boats that could safely cross the shallow waters around islands.

Coral rocks and basalt have thus been the fundamental building material on the islands, used for houses, temples, tombs, terraced fields, and windbreaks. Coral rocks, located above the layer of basalt, are accessible to everyone in need of building material. Their low density of $1,5$ g/cc makes them also more easy to move and to carve than basalt with a density of 3 g/cc. Where basalt is more easily accessible than coral rocks, e.g., on the island of Tongpan, it became the primary building material. Although basalt is much harder than coral rocks, both erode within the storm of sand and salt particles that blows over the islands half of a year.

## 3.4 Penghu Epigraphies

In this harsh environment, tombs and tombstones that spread out along cliffs, beaches, and hills are so brittle and fade away so easily that one might wonder, how epigraphic practices can be perpetuated from generation to generation. Many of these practices have no other source or documentation than previously inscribed tombstones. There is no administrative regulation, no religious norm, and no

discourse in relation to tombstone inscriptions, as death is a taboo in Han societies. Only professional carvers might maintain such a discourse. All other inhabitants of these islands construct their proper conception of what they perceive to be a local tradition through the abstraction, generalization, and interpretation of observed practices. A fundamental question we thus have to address is whether epigraphic practices would be readable for a time span that is long enough to guarantee their perpetuation or whether the perpetuation of a carving practice would depend entirely on the discourse maintained by professional carvers.

The time period for which epigraphic inscriptions are readable to the local people can only be estimated as unreadable tombstones usually can no longer be dated. For our estimate, we follow the assumption that local people, unlike researchers and professional, see mainly their family tombs and the easily accessible tombs on their way to their family tombs. We estimate that they do not see the oldest 25% of those tombstones that we have documented in our exhaustive research, removing vegetation, cleaning tombs, digging into the ground, and using specific photographic techniques to get the last readable stroke out of each tombstone. We thus use the second quartile of the tombstones of our digital documentation as the time that tombstones are no longer visible or readable.[13]

Calculating the readability of tombstones for different islands and different materials as shown in Table 3.1, we notice that on some islands erosion may affect the readability already after two generations (60 years). After three generations (90 years), tombstones usually become unreadable. This means that upon the death of the parents, some of the tombstones of the generation of the grandparents and

**Table 3.1** The effect of location and material on the readability of tombstones on Penghu, estimated in the number of years

| | Magong | Baisha | Xiyu | Wangan | Qimei | Mean |
|---|---|---|---|---|---|---|
| Basalt | 113 | 84 | 94 | 137 | | 107 |
| Concrete | 72 | 78 | 59 | 74 | 71 | 71 |
| Coral | 94 | 43 | 116 | 56 | 99 | 82 |
| Granite | | 77 | 149 | 270 | 155 | 163 |
| Mean | 93 | 71 | 104 | 134 | 108 | 102 |

[13]The assumption of a cutoff point at 75% is arbitrary in absolute terms. Yet, if applied consistently, the estimate would allow to compare the readability of tombstones depending on the region or the material. And still, this assumption seems to match the extent to which local people know their local graveyards. Talking after a survey to local people about the graveyard we have visited, our interlocutors are usually surprised to hear us talking about readable Qing tombstones. Likewise, people refer to tombs of the 1960s as old tombs, although we have found tombs of the Japanese period in the same graveyard. In one particular instance, locals on the Tongpan island told us that people started to inscribe their tombstones about 100 years ago, although we found among the ten remaining inscribed tombstones two Qing tombstones, dating back at least 120 years. The knowledge that local people have of their graveyard is thus astonishingly precise, but not as precise as the data that we get through our digital documentation.

most tombstones of the generation of the great grandparents have been eroded. The erosion affects particularly tombstones carved into concrete and coral rocks and islands with exposed graveyards, Baisha, with no natural protection against winds from the north, and Qimei where most graveyards are located along the northern cliff. The surprising long readability of coral rocks relates to the fact that this soft stone is usually carved up to more than a centimeter deep, while hard materials, such as concrete, basalt, and granite, are usually carved only a few millimeters deep.

```
df.penghu.mat <- stone.has.material(df.penghu.year)
df.penghu.read <- inscription.has.family(df.penghu.mat)
df.penghu.read <- stone.creation.is.before
   (df.penghu.read,1970)

list.islands <- split(df.penghu.read,
list(df.penghu.read$graveyard.island,df.penghu.read
   $stone.material), drop=TRUE)
readable <- sapply(list.islands,function(x) 2010 -
   quantile(get.year(x))[2])
```

This substantial erosion is a key feature for development of epigraphic practices through time. We argue that when older tombstone inscriptions become unreadable, a newer carving will turn into a visible model from which tradition, identities, and narratives are constructed. For Penghu, this means that traditions are based on still visible practices that date back not much longer than two generations. A person burying, for example, one of his or her parents after WWII might find as a template only family tombstone carved during the Japanese period, unable to distinguish aspects of these inscriptions conditioned by the Japanese occupation from those that reflects older practices. Thus, although the visibility of tombstones on Penghu may be long enough to perpetuate a practice from generation to generation, it is difficult for local people to understand how inscriptions have historically been formed. A mental conception of an epigraphic tradition is thus either formed from a limited point of view or put into the hands of a few professionals, who are interested, more than in historical truth, in the economic success of their craft enterprise. Interestingly, in tombstone inscriptions we find both paths along which traditions are constructed, craggy carvings of dodgy content, as well as perfectly shaped and structured gravestones. While the former reflect the individual effort to make sense of epigraphies, the latter embody the constructed tradition the carver might have presented to his clients, as well as the economic model that underlies the carving of this myth.

The material of the tombstones, granite, basalt, or coral rocks, not only differentiates tombstones as to their durability but also as to their geographic origin, the

route they have taken from the quarry and, accordingly, the type of carver or even a specific carver family.

The relation between the material of the tombstone and the type of carver is relatively straightforward. Professional carvers went through an apprenticeship with a master, frequently an elder family member. They have learned how to handle and inscribe hard stones, granite and basalt, and their inscriptions are based on a syntax, a system of abstract rules. Trying to distinguish themselves from nonprofessional carvers, they tend to use hard materials as an expression of their concern for quality, to rely on a specific syntax to produce auspicious tombstones and, if required, to explain to a customer the advantage of their carvings in comparison to that of a competitor.

Semiprofessional carvers are construction workers who set up the tomb and take over the carver's job of procuring a tombstone and carving the inscription. This is technically possible, as the carver tends to be subcontracted by the construction worker. Thus, unless required by the family of the deceased, there might be a tendency of construction workers to infringe into the realm of the carver to increase their profit margin. Their material chosen for the tombstone is similar to the material the tomb has been made of. Examples are tombstones made from bricks and then covered with a layer of cement, or concrete slates, into which characters are scratched while the cement is drying. As bags of concrete and bricks belong to the general purpose material that can be found even on smaller islands, tombs are easily and quickly set up, avoiding the logistic problems involved in procuring a tombstone from a carver located on a larger island. Semiprofessional carvers may or may not have acquired the syntax of a tombstone inscription from their trained colleagues (Figs. 3.5, 3.6, and 3.7).

**Fig. 3.5** A tomb and tombstone on Dongyudongping, very likely to have been set up by the one and same craftsman, potentially a semiprofessional carver according to our classification. The tombstone has been made from concrete and characters have been scratched into an outer layer of fresh fine cement

**Fig. 3.6** A tombstone found on Dongyudongping that is likely to have been carved by a nonprofessional carver. The inscription is carved into a soft sedimentary stone. The inclusions of sea shells are well visible. The stone has not been shaped into a particular form and characters are irregular in size and arrangement. The content of the inscription has been reduced to two semantic roles and a total of seven characters, while professional carvings require four semantic roles and at least 21 characters

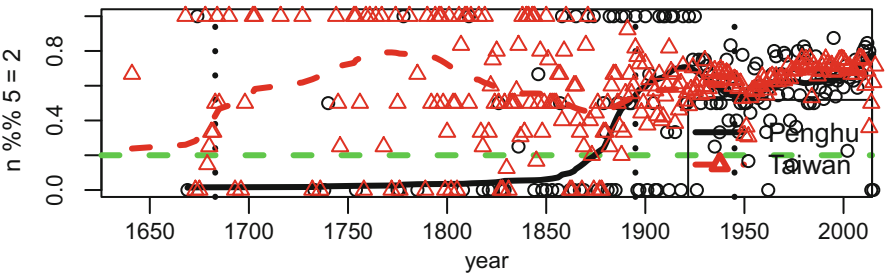



**Fig. 3.7** The relative usage of auspicious character numbers on Penghu and Taiwan through time

Nonprofessional carvers, the group we know very little about, may be family members of the deceased. They usually carve the tombstone if the tombstone is the only visible marker of the tomb, meaning that even the second group of professionals is not involved in setting up the tomb. On Penghu, this third group prefers to carve into coral rock, an omnipresent material which is easy to cut, to transport, and to inscribe. Nonprofessional carvers are not in professional exchange with carvers. They might copy and adapt the surface form of professional inscriptions found on a burial ground, without necessarily understanding the syntax of the expressions. This may result in the violation of the less obvious syntax rules that professional carvers tend to follow.

We distinguish three tombstone properties that professional carvers might take into their consideration when designing a tombstone. Classifying tombstones with respect to these properties, we are in state not only to distinguish professional from nonprofessional carvers but also to distinguish different carving practices, potentially of different carving families.

Most stone carvers of Penghu and Taiwan, when interviewed as to how they design a tombstone inscription, will sooner or later come to the point to explain a complex numerical grammar that is supposed to render a tombstone auspicious. As part of this grammar, the middle line of the tombstone inscription, the line that refers to the deceased, should be composed of seven, twelve, seventeen, etc., characters. Mathematically, this requirement can be expressed as $(n\ modulo\ 5) = 2$, i.e., the number of character modulo 5 equals to 2. Why this should be the case is a story in itself, telling much about how practices are formed:

Having its origin in Buddhism, the phrase '生老病死' (shēng lǎo bìng sì, birth, aging, illness, death) represent the circle of rebirth and suffering. These four elements are thus '苦' (kǔ bitter). In Sanskrit *jāti-jarā-vyādhi-maraṇaṃ*, the equivalent of '生老病死' are equal to *dukkha* (suffering). In Korea, for example, the phrase '生老病死' is known, while the phrase '生老病死苦' is not. In the Chinese folk religion practised on Penghu and Taiwan, this wisdom transformed into the aphorism '生老病死苦' (*idem* kǔ, *idem* suffering), which is pronounced in a circle '生老病死苦生老...' to assign auspicious or non-auspicious meaning to numerals. Most tombstones on Taiwan and Penghu thus have turned into a documentation, carved into stone, how Chinese folk religions transformed Buddhist teachings. The first (sixth, eleventh, etc.) character is associated with '生' (shēng, birth), the second (seventh, etc.) with '老' (lǎo, old), etc. '生' (shēng, birth) and '老' (lǎo, old) are considered auspicious characters, while the others are not. The numbers 7, 12, 17, etc. would be considered to be auspicious numbers for the central line in combination with the right and left line ending in '生' (sheng). These two '生', when read in a row, yield a '老' (lǎo), cf. 6 + 6 =12. Adding a focus in top of two characters, a place name or a loyalty expression, adds another '老' (lǎo). Three '老' (lǎo) add to the tombstone another '生' (sheng), cf. 12 + 7 + 2 = 21. In total, there are three '生' (sheng) and three '老' (lǎo), the '老' (lǎo) generated from '生' (sheng) and vice versa. Analyzing the number of characters in the central line is thus

a simple way to check the syntax of the tombstone inscription, especially if other parts of the inscriptions, usually with smaller characters, have become unreadable.

Using a timeline, we can pinpoint the transformation from an unorganized random state into a regulated state. Plotting the relative percentage of various values ("yes") of one attribute ("lao"), in relation to a random distribution of that attribute, we can identify when the number of values that represents a regulated state exceeds the baseline of a random distribution.

```
funct.plot.line <- function(x,y,mycol,mylty,myf) {
points(x,y,pch=mylty,col=mycol)
lines(x,lowess(y,f=myf)$y,col=mycol,lwd=3,lty=mylty) }
####
lao.per.year <- function(df,mycol,mylty) {
vec.middle <- get.inscription.length.middleline(df)
vec.lao <- (as.integer(vec.middle) %% 5) == 2
df.lao.year <- aggregate(vec.lao,FUN=mean,
by=list(get.stone.creation.year(df)))
####
funct.plot.line(df.lao.year[,1],df.lao.year[,2],
   mycol,mylty,0.2)}
df.middle <- inscription.has.length.middleline(df.asia)
df.middle <- stone.has.creation.year(df.middle)
plot(0,0,xlim=c(1640,2000),ylim=c(0,1),xlab='year',
   ylab='n %% 5 = 2',col='red')
abline(v=c(1683,1895,1945),lty=3,lwd=3)
abline(h=1/5,col='green',lty=2,lwd=3)
legend('bottomright',c('Penghu','Taiwan'),lty=c(1,2),
   pch=c(1,2),lwd=3,col=c(1,2))
zero <- mapply(lao.per.year,list(is.penghu(df.middle),
   is.taiwan(df.middle)),c(1,2),c(1,2))
```

Our claim here is that the success of this feature was the result of the commercial advantage that those carvers who offered this feature had over carvers who did not. On Penghu, for example, we can observe that the probability for this feature to be applied was before 1850 below random chance (20%), c.f. Fig. 3.8. After 1850, this feature propagated quickly and developed within a short time into a local standard. The reason for this general spreading is easy to understand. Once a feature is used by a carver with an important market share in combination with a convincing story, the remaining carvers had to apply the same feature likewise, if they did not want to appear as an incompetent carver who carves inauspicious tombstones. Why carvers on Penghu have not used this feature before 1850, as it has been in Taiwan, and how it arrived on Penghu is currently still unclear.

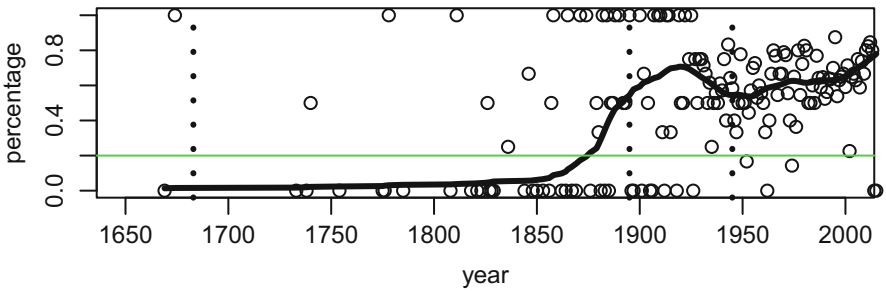



**Fig. 3.8** The percentage of tombsstones on Penghu with an auspicious number of characters

```
df.penghu.lao <- inscription.has.length.middleline
    (df.penghu.year)
vec.length.middle <- get.inscription.length.middleline
    (df.penghu.lao)
df.penghu.lao$lao <- as.integer(((vec.length.middle)
    %% 5) == 2)
df.penghu.lao.mean <- aggregate(df.penghu.lao$lao,
FUN=mean,
by=list(get.year(df.penghu.lao)),SIMPLIFY=F)
plot(0,0,xlab='year',ylab='percentage',xlim=c(1650,
    2000),ylim=c(0,1))
funct.plot.line(as.character(df.penghu.lao.mean$
    Group.1),df.penghu.lao.mean$x,1,1,0.2)
abline(h=0.2,col='green',lty=1)
abline(v=c(1683,1895,1945),lty=3,lwd=3)
```

Plotting in Fig. 3.9 the spatial distribution of Qing tombstones with (green) and without (red) an auspicious number of characters on four maps for different materials, we can identify two separated areas. In the north-west of Penghu, we see a few potential centers of professional carvers, located in close neighborhood to each other, i.e., the north of Xiyu, Xiaomen, and Baisha. In these places, an auspicious number of characters is used on granite and basalt, presumably by professional carvers. In Wang'an, the percentage of stones with an auspicious number is close to the random baseline of 20%. Further to the south and east of Penghu, this relatively new feature is not found. Additional analyses will be required to clarify whether either carvers in the south were generally not professionals or the professionals located on the southern islands did not have been in dialogue with the north-western carvers, and had simply missed a new trend.

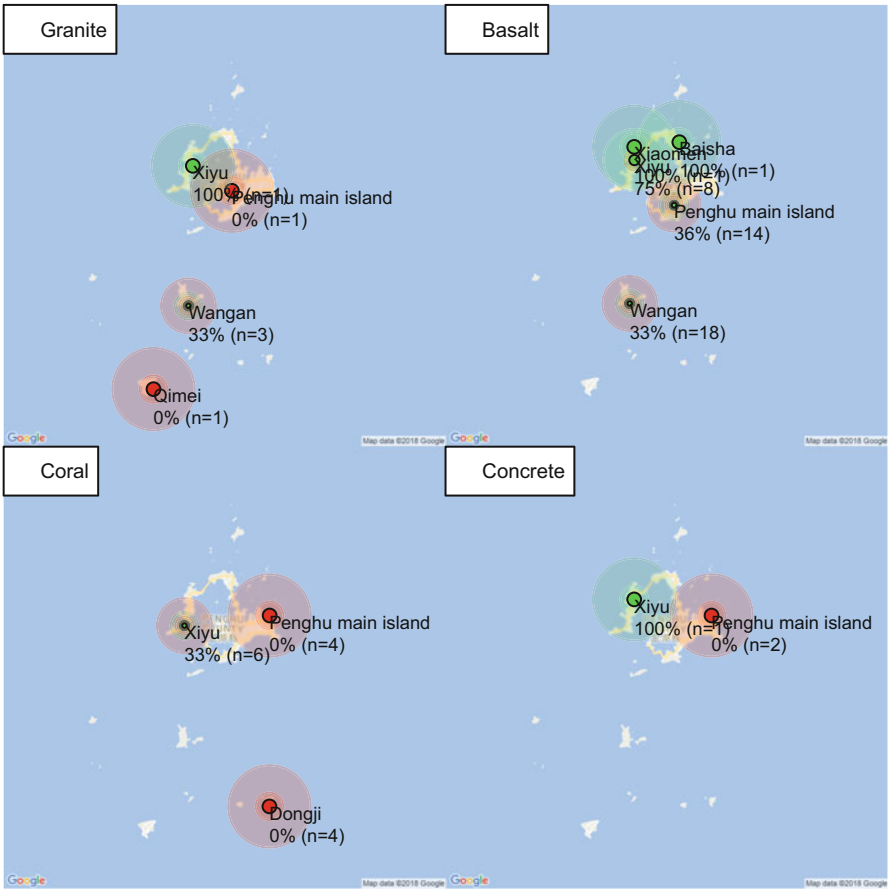



**Fig. 3.9** Tombstones with (green) and without (red) an auspicious number of characters

```
percentage <- function(x) { return(sum(x)/length(x))}
####
plot.dot.on.island <- function(df,col.index,text.cex,
   message){
vec.criteria <- df[,col.index]
list.stats <-mapply(aggregate,
list(vec.criteria,df$x,df$y,vec.criteria),
list(list(df$graveyard.island)),
list(percentage,mean,mean,length),SIMPLIFY=FALSE)
PlotOnStaticMap(canvas.penghu)
```

(continued)

```
map.points(canvas.penghu,list.stats[[2]]$x,list.stats
    [[3]]$x,
3* (1-list.stats[[1]]$x),'red')
map.points(canvas.penghu,list.stats[[2]]$x,list.stats
    [[3]]$x,
3* list.stats[[1]]$x,'green')
legend('topleft',message,bg='white')
map.text(canvas.penghu,list.stats[[2]]$x,list.stats
    [[3]]$x,cex=text.cex,
paste(list.stats[[1]]$Group.1,'\n',round(list.stats
    [[1]]$x*100),'% (n=',list.stats[[4]]$x,')',sep=''))}
####
df.penghu.lao.qing <- period.is.roc.qing(df.penghu.lao)
col.index <- grep('lao',names(df.penghu.lao.qing))
par(mfrow=c(2,2))
plot.dot.on.island(is.granite(df.penghu.lao.qing),col.
    index,0.9,'Granite')
plot.dot.on.island(is.basalt(df.penghu.lao.qing),col.
    index,0.9,'Basalt')
plot.dot.on.island(is.coral(df.penghu.lao.qing),col.
    index,0.9,'Coral')
plot.dot.on.island(is.concrete(df.penghu.lao.qing),col.
    index,0.9,'Concrete')
```

A tombstone with a rounded top is in Taiwan and Penghu assumed to be more auspicious than a rectangular tombstone. The reason for this is that cosmical energy is believed to be transmitted along mountain peaks from high mountain ranges in China. A mountain-shaped tombstone is thus assumed to capture this cosmical energy and tunnel the energy through the bones of the deceased onto the family members. As rounded tombstones are difficult to carve, they might also be an indicator for professional carving. Drawing a timeline of the categorical distribution of tombstone forms, we can indeed observe a systematic trend towards rounded tombstones. This trend started in the beginning of the eighteenth century and continued into the middle of the twentieth century. During the Japanese colonial period, the Japanese tombstone in the form of a column became popular for reasons that relate to the position of local people within the society under the Japanese (Figs. 3.10 and 3.11).

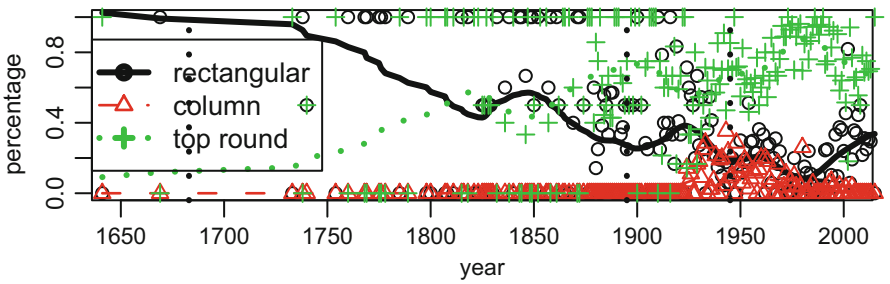
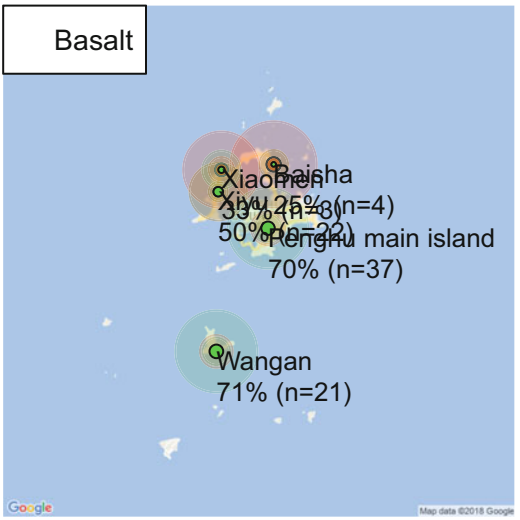



**Fig. 3.10** The categorial distribution of tombstone shapes on Penghu through time

**Fig. 3.11** The percentage of top-rounded tombstones on Penghu through time




```
df.penghu.form <-subset(df.penghu.year,
   grepl('(4c$|^top round|column)',stone.form))
vec.form <- get.stone.form(df.penghu.form)
vec.year <- get.stone.creation.year(df.penghu.form)
tab.form.year <- table(vec.form,vec.year)
tab.prop <- prop.table(tab.form.year,2)
plot(0,0,xlab='year',ylab='percentage',xlim=c(1650,
   2000),ylim=c(0,1))
zero<-mapply(funct.plot.line,list(colnames(tab.prop)),
   get.tab.row(tab.prop),1:3,1:3,0.2)
abline(v=c(1683,1895,1945),lty=3,lwd=3)
legend('left',c('rectangular','column','top round'),
   col=1:3,lty=1:3,pch=1:3,lwd=3)
```


**Fig. 3.12** A wooden Luban measure for yang fengshui. Used for measuring tables, chairs, windows, doors, etc.


**Fig. 3.13** A wooden Luban measure for yin fengshui. Used for measuring tombs and tombstones

Differently from the auspicious number of characters, stones are rounded in the south of Penghu but less so in the north. Roughly speaking, the auspicious stone shape and the auspicious number of characters seem to exist in complementary distribution, pointing to two professional carving practices, one with its center on Xiyu, the other with its center on Wang'an.

```
df.penghu.form.qing <- period.is.roc.qing(df.penghu.
   form)
df.penghu.form.qing <- stone.material.is.basalt(tomb.
   has.xy(df.penghu.form.qing))
df.penghu.form.qing$form2 <- 0
df.penghu.form.qing$form2[df.penghu.form.qing$stone.
   form=='top round'] <- 1
col.index <- grep('form2',names(df.penghu.form.qing))
plot.dot.on.island(is.basalt(df.penghu.form.qing),col.
   index,0.9,'Basalt')
```

A third indicator of professional carving is the width and the height of a tombstone. From 1850 on, professional carvers tended to determine the width of the tombstone by the auspicious measures for the living (yin) and the dead (yang) as printed on a Fengshui tape measure (魯班尺 lǔbānchí, 文公尺 wéngōnchí, 風水尺 fēngshuǐchí), cf. Figs. 3.12, 3.13, 3.14 and 3.15. Before 1850, only the yin fengshui was used to determine the size of a tombstone.

Notice, the number of characters on a tombstone and the width of a tombstone showed significant changes around 1850. Yet, as both curves show a different slope, we assume that these transformations were not introduced by the same agent or process (Fig. 3.16).


**Fig. 3.14** Modern metal Luban measure, combining the measure for good yin and yang fengshui. Yang is arranged at the top, yin at the bottom of this measure

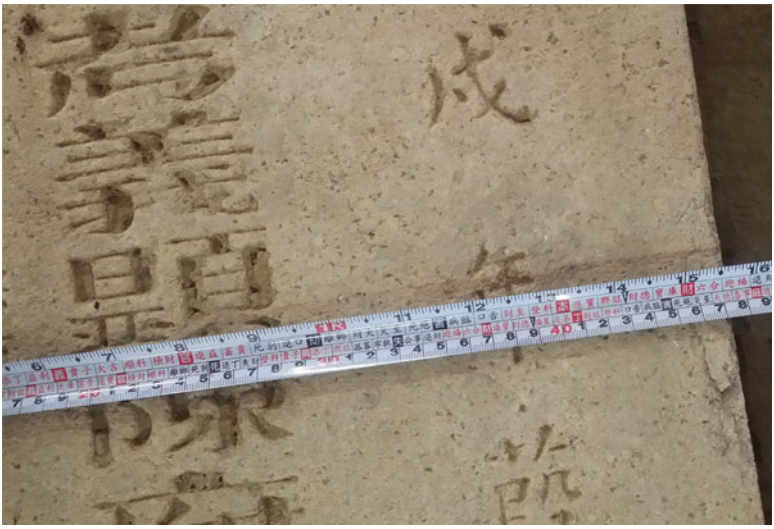


**Fig. 3.15** A tombstone of the year 1898 having a width of 47 cm. This width is auspicious, as 47 cm are red (auspicious) for yang, marked above, and yin, marked below

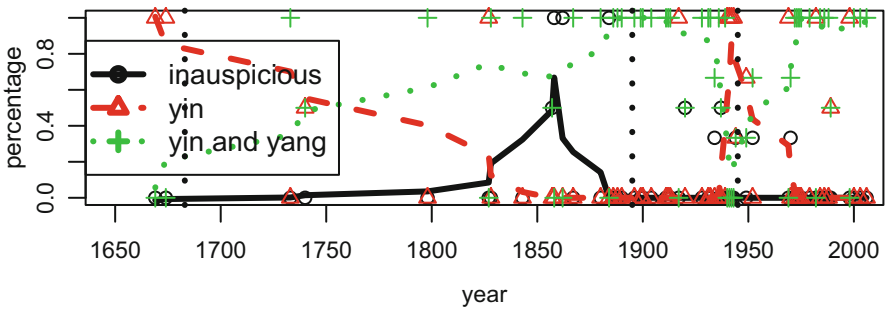



**Fig. 3.16** The percentage of tombsstones with an auspicious tombstone size on Penghu through time

As our analysis in Fig. 3.17 and Table 3.2 shows, basalt tombstones show more features of a professional carving than coral rocks.


**Fig. 3.17** Tombstones with an auspicious yin and yang (green) versus tombstones with only an auspicious yin (red)

**Table 3.2** The material and properties of tombstones

| | Basalt | Coral | Granite |
|---|---|---|---|
| *N* | 665 | 310 | 26 |
| Usage | 1754–1979 | 1669–1979 | 1733–1973 |
| Char. in LAO | 61% | 48% | 46% |
| Yin and yang width | 64% | 50% | 75% |
| With Tanghao | 1897–1979 | 1911–1978 | 1943–1959 |
| *N* 1895–1945 | 228 | 142 | 9 |
| Tanghao 1895–1945 | 36% | 18.3% | 11.1% |

```
df.penghu.width <- stone.has.width(df.penghu)
vec.penghu.width <- get.stone.width(df.penghu.width)
df.penghu.width$black.yang <- measure.is.black.top
   (vec.penghu.width)
df.penghu.width$black.yin <- measure.is.black.bottom
   (vec.penghu.width)
df.penghu.width$red.yin <- measure.is.red.bottom
   (vec.penghu.width)
df.penghu.width$red.yang <- measure.is.red.top
   (vec.penghu.width)
df.penghu.width$red.black <- df.penghu.width$red.yin *
df.penghu.width$black.yang
df.penghu.width$red.red <- df.penghu.width$red.yin *
df.penghu.width$red.yang
df.penghu.width$luban <- 'inauspicious'
df.penghu.width$luban[df.penghu.width$red.black==T]
   <- 'yin'
df.penghu.width$luban[df.penghu.width$red.red==T]
   <- 'yinyang'
df.penghu.width.year <- stone.has.creation.year
   (df.penghu.width)
vec.year <- get.stone.creation.year
   (df.penghu.width.year)
vec.luban <- df.penghu.width.year$luban
tab.form.year <- table(vec.luban,vec.year)
tab.prop <- prop.table(tab.form.year,2)
plot(0,0,xlab='year',ylab='percentage',xlim=c(1650,
   2000),ylim=c(0,1))
zero<-mapply(funct.plot.line,list(colnames(tab.prop)),
   get.tab.row(tab.prop),1:3,1:3,0.2)
abline(v=c(1683,1895,1945),lty=3,lwd=3)
legend('left',c('inauspicious','yin','yin and yang'),
   col=1:3,lty=1:3,pch=1:3,lwd=3)
```

```
df.penghu.width.qing <- period.is.roc.qing(df.penghu.
   width)
df.penghu.width.qing <- subset(df.penghu.width.qing,
   luban!='inauspicious')
```

(continued)

```
col.index <- grep('red.red',names(df.penghu.width.
    qing))
par(mfrow=c(2,2))
plot.dot.on.island(is.granite(df.penghu.width.qing),
    col.index,0.9,'Granite')
plot.dot.on.island(is.basalt(df.penghu.width.qing),
    col.index,0.9,'Basalt')
plot.dot.on.island(is.coral(df.penghu.width.qing),
    col.index,0.9,'Coral')
```

Some additional features of tombstones inscription we have to discuss are the *semantic roles* and the *focus* of tombstone inscriptions, as well as the various forms the *focus* might take, either as *loyalty expression*, as jiguan (籍貫 jíguàn), or as tanghao (堂號 tánghào).

In linguistic theories, a semantic role describes the kind of semantic contribution that a phrase adds to its larger linguistic unit. For example, in the sentence "*Kim comes today*," *today* functions as a temporal complement. In the same line, we define the *semantic role* of a tombstone inscription as the semantic contribution that a phrase adds to an epitaphic inscription. Semantic roles vary from community to community. Examples are: *the deceased*, *date of birth*, *date of death*, *date of burial*, *date of tombstone erection*, *placename*, *religious affiliation*, *generation number*, *profession*, *political affiliation*, *buried with*, *buriers*, etc.

As not all semantic roles can enter a tombstone inscription, and communities and carvers like to maintain a certain order in the arrangement of the inscription as a proof of their professionality, semantic roles are classified into paradigms of grammatical functions that contain mutually exclusive semantic roles. Syntactic functions in many natural languages are subject, predicate, and object, but also topic, focus, and contrast. Not only have different languages different grammatical functions, they also have different limitation as to which semantic role can enter which grammatical function. German, for example, has the semantic role of *experiencer* which can enter an indirect object (*Mir ist kalt*, literally: *To me is cold*, translated as: *I have cold*, a combination that is not possible in modern English. Likewise, the tombstones of different carvers or different regions have their sets of semantic roles and grammatical functions they can enter.

Of particular interest, because showing a remarkable variation between regions, time periods, and religious or political regimes is the grammatical function that we call *focus*, a formally identified function that identifies a social community as the main source of a social identity. A *focus* can be a placename, a religious symbol, or a reference to a government. As there is usually only one main source of a social identity, there tends to be a competition as to which semantic role will enter this function.

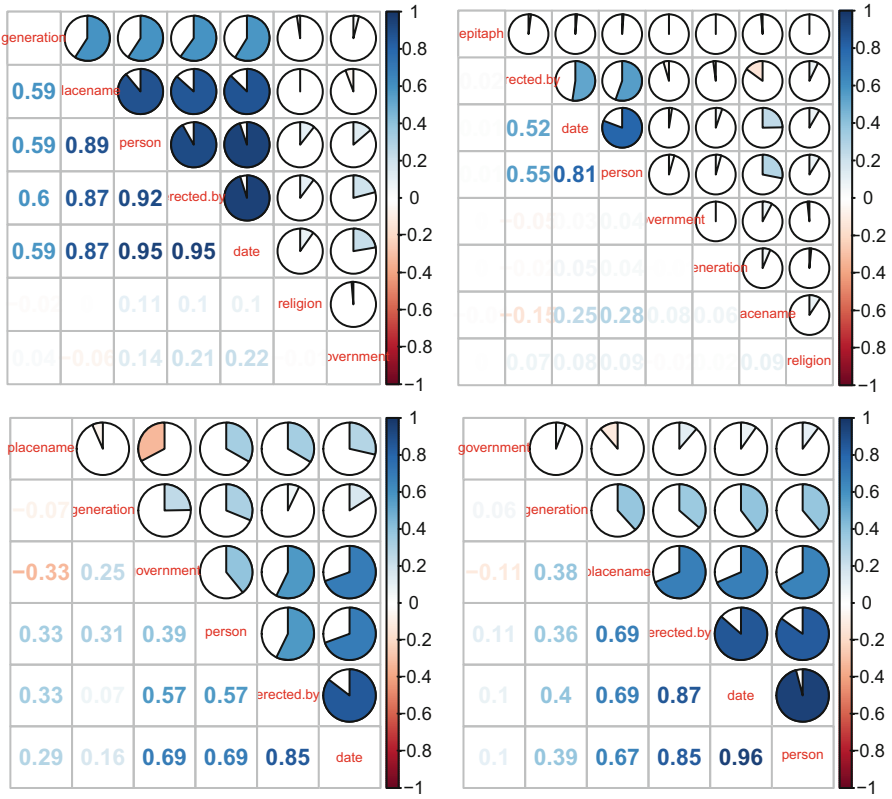



**Fig. 3.18** The correlation of semantic roles as indicators of syntactic functions in tombstone inscription, top-left: Penghu, top-right: Wenzhou, bottom-left: Penghu in Qing period, and bottom-right: Penghu in the Japanese period

Comparing in Fig. 3.18 the correlation matrixes of semantic roles in Penghu and Wenzhou, we notice that in both cases, the *the person*, co-occurs freely with *date*, *placename*, and *erected.by*. The function of the *placename* however is different in both cases. While in Penghu *generation*, *placename*, *religion*, and *government* compete for the same *focus* position, with the *placename* as default, in Wenzhou, where there is no formally marked focus position, the *erected.by* competes with *government*, *generation*, and *placename* for the function *social.identity*. Looking at Penghu in Qing and Japanese period, we see that the opposition of *government* and *placename* is a historical one. In Qing dynasty, the default *focus* was *government* and the *placename* was marginalized. In the Japanese period, the opposition still existed, but the quantitative relations of these two roles were inverted. It is this inversion which is the heart of our investigation.

```
funct.empty.col <- function(x){return(sum(x)==0)}
####
funct.corr.sem.roles <- function(mydf) {
library('corrplot')
mydf <- transcription.is.completed(mydf)
vec.role.idx <- grep('inscription.xml',names(mydf))
tab.sem.roles <- !mapply(is.na,mydf[,vec.role.idx])
vec.emtpy.col <- -1 * which(mapply(funct.empty.col,
split(tab.sem.roles,col(tab.sem.roles))))
tab.sem.roles <- tab.sem.roles[,vec.emtpy.col]
colnames(tab.sem.roles) <- mapply(gsub,'inscription.
   xml.','',colnames(tab.sem.roles))
corrplot.mixed(cor(tab.sem.roles),upper='pie',
   order = 'hclust',tl.cex=0.6)}
####
par(mfrow=c(2,2))
zero <- mapply(funct.corr.sem.roles,list(df.penghu,
   is.wenzhou(df.asia),
period.is.roc.qing(df.penghu),period.is.roc.japan
   (df.penghu)))
```

Most tombstones in Taiwan are written with a Chinese script, usually from top to bottom, from left to right. The upper part of these tombstones, written mainly from right to left, is what we call the *focus* position. The focus is thus marked by its flipped writing direction in that area of the tombstone, from where the reading of most lines starts. The content of a focus can be described by its semantic role, e.g., *generation number, place name, loyalty expression, religious affiliation*, etc. All other rows in the tombstone can also be ascribed as semantic role, such as *deceased, mourners, date*, etc., the content of most semantic roles which are not focused reports distinctive information, i.e. information that distinguishes the deceased from all other deceased on this burial ground. The focus, however, through its form or through its content, through its type or through its token represents real or imagined communities, in this or beyond this burial ground. A cross on a Christian tombstone, or the placename 台南 (Taínán) on the largest graveyards of Tainan, Taiwan are good examples of how these foci function.

A common way to fill in the focus position is by using a placename. We distinguish three reference types placenames: The *local placename*, which references the birth or living place of the deceased on Taiwan or Penghu, the *jiguan* (籍貫), where a family was registered in China before migrating to Penghu or Taiwan (Fig. 3.19), and the *tanghao* (堂號), a maybe historical, maybe mythological place from where the surname is thought to have originated 2000 years ago.


**Fig. 3.19** A placename in China in focus position

Before the appearance of the *tanghao* on tombstones of Penghu and Taiwan, the *tanghao* was a component of statal strategies to define its space. For centuries, *tanghao* have been traded in the canonical book *Baijiaxing* (百家姓), which is usually translated as *The Hundred Family Surnames*. Through this book, people acquired literacy over the last eight centuries, along with the Neo-confucianist *Three Characters Classic* (Sanzijing 三字經) and the earlier *Thousand character Classic* (Qianzi Wen 千字文).[14] This book listed originally 400 Chinese surnames in quadrisyllabic rhyming couplets. It is assumed to have been composed during the period of the Northern Song (960–1127), as the first surname in the book, Zhao (趙), is the surname of the Song dynasty, to whom the author of the *Baijiaxing* might have wanted to pay a homage.[15] In later dynasties, new editions of the *Baijiaxing* rearranged the order, so that the surname of the respective ruling dynasty was in first position.[16]

[14]Peng, *Five Hundred Years Ago, It Was One Family*: 16.

[15]Ibid.16.

[16]Examples of these derivations are the Liu Zhongzhi (劉仲質), the *Yuzhi* Baijiaxing (御制百家姓), and the *Baijiaxing Er2bian* (百家姓二編) Theobald (2011).

The most common *tanghao* referred to in Taiwan, *Yingchaun* (潁川) is the *tanghao* of families surnamed Chen (陳), Zhong (鍾), Lai (賴), Wu (烏), and Gan (干).[17]

We refer as *loyalty express* to those expressions which mention the ruling dynasty, the government, a form of a government, or the loyalty or devotion to the government. The highest percentage of tombstones with a loyalty expression can be found, according to our data, on tombstones of the Zheng era. After all, the Zheng family claimed to be loyal to the Ming dynasty, even though the Manchurian Qing had conquered China. During that time, we find frequently the expression 明 (míng) or 皇明 (huángmíng) on the top of the tombstone. Much of the loyalty expression is represented in the Chinese character '皇' (huáng, emperor) which represents the emperor (王 wáng) under the white light of the sun. Putting this on top of the tombstone, above the name of one's ancestors means to accept this hierarchy of commoners under the emperor, who himself is gifted with cosmical power.

The interpretation of these expressions as loyalty expression derives also from the fact that after a government had been replaced, a loyalty expression has usually been avoided for a couple of generations. Not using the loyalty expression was a form of protest. When the loyalty expression reappeared again on the tombstones of Taiwan and Penghu in Qing dynasty, it was written 皇清 (huángqīng) and only occasionally 清 (qīng).

Towards the end of the Japanese colonial period, the loyalty expression reemerged in various forms, of which 皇日 (huángrì), 皇民 (huángmín) and 皇恩 (huángēn) are the most common.

## 3.5 Penghu Epigraphies Under the Ming and Qing

For a long time, Penghu could neither nourish its population nor provide a product the local people could use for trading. The settlements on Penghu thus have been and are to our days not the outcome of economic projects of the local people, but of political and economic projects of the state and its elites. Local people were needed as work forces, food providers, and placeholders who prevented pirates, smugglers, or outlaws to settle on these islands. For a long time, Penghu thus functioned only as a state subsidized trading hub and as military outposts, creating a small market on which local people could access imported products. The economic dependence on merchants, soldiers, and officials influenced tombstone inscriptions depending on how close or far people were to the center of this military and economic power.

[17] *Yingchaun* can neither be found on a modern map of China nor in *The China Historical GIS*. Yet, numerous publications locate *Yingchaun* in Henan (河南). Most of these publications, unfortunately, work without historical sources and copy instead widely from each other. Without proper sources, these publications do not lend themselves to corroborate or refute the historical existence of Yingchuan (潁川). Many other *tanghao* are without doubt historical placenames and some have even equivalents in the modern world.

After the last evacuation of Penghu under the military ban in the fifteenth century, fishing communities were resettled under the Ming towards the end of the sixteenth century and military forces were stationed on Penghu from 1603 on. In 1622, the Dutch Vereenigde Oost-Indische Compagnie (VOC) (United Dutch East India Company) occupied Penghu but eventually had to leave in 1924 and moved to Taiwan. The VOC was ousted in 1661 by the Ming-loyalist Koxinga (鄭成功 Zheng Chenggong), which caused the Qing to issue another maritime ban with the unintended side effect of triggering a migration wave not landward, but to Penghu between 1662 and 1664. After the battle of Penghu in 1683, the Qing conquered Penghu and shortly later also Taiwan.

Having experienced maritime ban, relocation, foreign occupation, reoccupation, maritime ban again, migration, the loss of the emperor, and his replacement by a swashbuckler, the local people were aware that their presence on the islands would depend on the government and its ability to stay in power. Especially, during the period of the Zheng regime, the time most known Ming tombstones fall into, a considerable part of the population were Ming loyal soldiers, prepared to face the Qing who were gathering their strength at the Chinese coast. Loyalty and support of the current government, whatever it was, might have been the tactics of most inhabitants to avoid future uncertainties. It thus seems almost natural, that on Penghu as in Taiwan, the most common focus position on tombstones was the expression of loyalty to the Ming, expressed mostly as 皇明 (huang2ming2).

Unfortunately, we cannot reconstruct the transition of tombstones on Penghu from Ming to Qing in detail, as the earliest preserved Qing tombstone dates from 1733, 50 years after the establishment of the Qing on Penghu. This tombstone can be found on the burial ground of Dong'an on Wang'an, a few hundred meters from the harbor, with a loyalty expression to the Qing carved into granite. Granite tombstones can be usually found where boats were unloaded and thus indicate a direct shipping link to China. The majority of loyalty expressions however can be found not on Wang'an, but on Xiyu, where the Qing took over the West Fort from Koxinga and added the East Fort in 1883.

Wang'an finally adopted the *jiguan* as its principal focus on tombstones about 1750, while the loyalty expression continued to be used on Xiyu. For this to happen, it was not enough that Wang'an was unreachable from Makong during the months of the winter monsoon. In order not to become the periphery of another, more central culture, Wang'an must have been a cultural center on its own, a center that could promote its own practices and thus avoid the import of practices. Traces of this cultural center can be attested in the form of unique local carvings. The character '穀' (gǔ, corn or lucky), employed as synonym of '吉' (jí, lucky), is found almost exclusively on tombstones on Wang'an (Fig. 3.20). The most likely interpretation for this distribution is that for more than 150 years a tombstone carver family must have been active in Dong'an, using this character in its carvings, where other carving families would use the character '吉'.

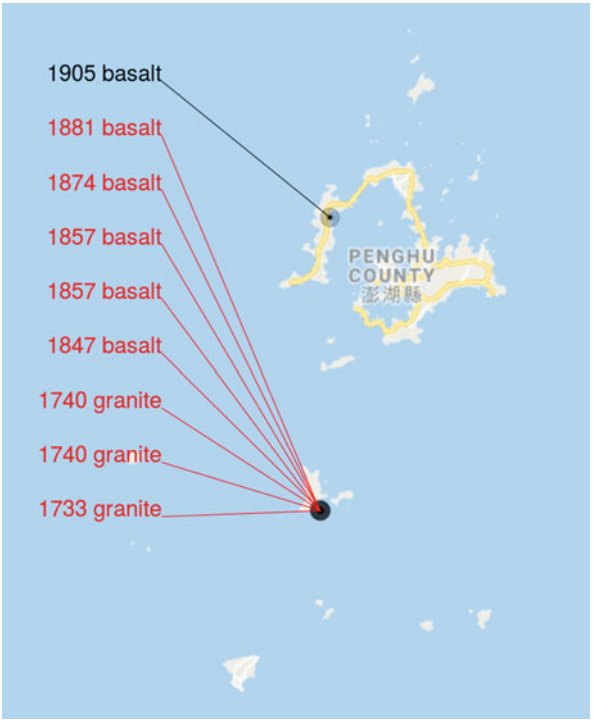



**Fig. 3.20** The spatial and temporal distribution of tombstones with the character '穀' (gǔ), in the sense of 'auspicious' on Penghu

```
df.penghu.mat <- stone.has.material(df.penghu.year)
has.GU <- grepl('穀',df.penghu.mat$inscription.xml.date)
df.penghu.GU <- subset(df.penghu.mat,has.GU)
mysorted <- sort.by(df.penghu.GU,df.penghu.GU$stone.
   creation,TRUE)
l <- nrow(df.penghu.GU)
x <- mysorted$tomb.x
y <- mysorted$tomb.y
x2 <- rep(119.35,l)
y2 <- 23.8 - rep(0.05,l)*1:l
lab <- paste(mysorted$stone.creation,mysorted$stone.
   material)
col <- rep('orange',l)
col[mysorted$stone.creation.year > 1683] <- 'blue'
col[mysorted$stone.creation.year > 1895] <- 'green'
PlotOnStaticMap(canvas.penghu)
map.points(canvas.penghu,x,y,1,'green')
```

(continued)

```
map.text(canvas.penghu,x2,y2,lab,cex=0.9,adj=c(1,0),
   col=col)
map.lines(canvas.penghu,x,x2,y,y2,col=col)
```

Wang'an was not the only center during the Qing period. A second center can be located on Xiyu, identified by the character '㫃', a variant to '時' (shí, time), in expressions like (在)辛未孟夏吉旦 (shí (zài) xīnweì mèngxià jídàn, time (is) metal-sheep year first lunar month auspicious morning), which themselves are quite particular, not only by the character '', but also by its phrasing.

```
df.penghu.mat <- stone.has.material(df.penghu.year)
df.penghu.mat <- stone.creation.is.before(df.penghu.
   mat,1940)
has.SHI <- grepl('㫃',df.penghu.mat$inscription.xml.date)
df.penghu.SHI <- subset(df.penghu.mat,has.SHI)
mysorted <- sort.by(df.penghu.SHI,df.penghu.SHI$stone.
   creation,TRUE)
l <- nrow(df.penghu.SHI)
x <- mysorted$tomb.x
y <- mysorted$tomb.y
x2 <- rep(119.4,l)
y2 <- 23.88 - rep(0.026,l)*1:l
col <- rep('orange',l)
col[mysorted$stone.creation.year > 1683] <- 'blue'
col[mysorted$stone.creation.year > 1895] <- 'green'
lab <- paste(mysorted$stone.creation,mysorted$stone.
   material)
PlotOnStaticMap(canvas.penghu)
map.points(canvas.penghu,x,y,1,'green')
map.text(canvas.penghu,x2,y2,lab,cex=0.8,adj=c(1,0),
   col=col)
map.lines(canvas.penghu,x,x2,y,y2,col=col)
```

What appears to be a relatively unsystematic spreading of a specific character variant represents however a very precise sequencing in terms of market power and market share. We observe the loss of a proper tombstone tradition on Wang'an after 1881 and the import of tombstones from Xiyu to Wang'an and Qimei after 1900, see Fig. 3.21. We interpret this as a relative loss of economic power and cultural independence towards the end of the Qing period.

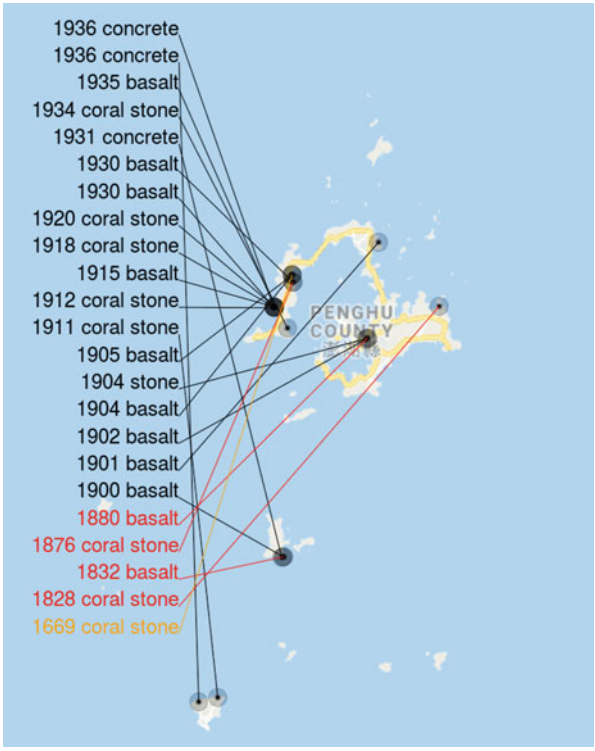



**Fig. 3.21** The spatial and temporal distribution of tombstones with the character '𥫱'(shí), in the sense of 'auspicious' on Penghu

After a carver from Xiyu took over the work on Wang'an, we find inscriptions typical for Wang'an on Xiyu and vice versa. First, there is a single instance of the character '穀' (gǔ), originally unique on Wang'an, carved into a tombstone with a '𥫱'(shí) on Xiyu. We interpret this occurrence as an inspiration the Xiyu carver got on Wang'an. Second, as Fig. 3.22 shows, a tombstone with a loyalty expression has been imported to Wang'an in 1894, most probably from Xiyu, where this loyalty expression has been very common. The shift of carvers from Wang'an to Xiyu thus took place between 1881 (last '穀' on Wang'an) and 1894 (first loyalty expression on Wang'an). In addition, the first '𥫱'is found on Wang'an in 1900 and the first '穀' on Xiyu in 1905.[18]

Also, Makong and Baisha were probably served from Xiyu in the very late Qing period, before Makong developed into the unrivalled center of Penghu. After 1902, we see no export to Makong or Baisha, both places are by then probably taken over by a carver from Makong. Which used neither '穀' (gǔ) nor '𥫱'(shí).

[18]The only tombstone on Wang'an with a '𥫱'(shí) that predates this hypothesized shift after 1881 has unfortunately been used as target for shooting games by bored ROC soldiers. One of their bullets hit what might have been the character '𥫱', right in its center, leaving only a few millimeters of the upper character component unscattered.

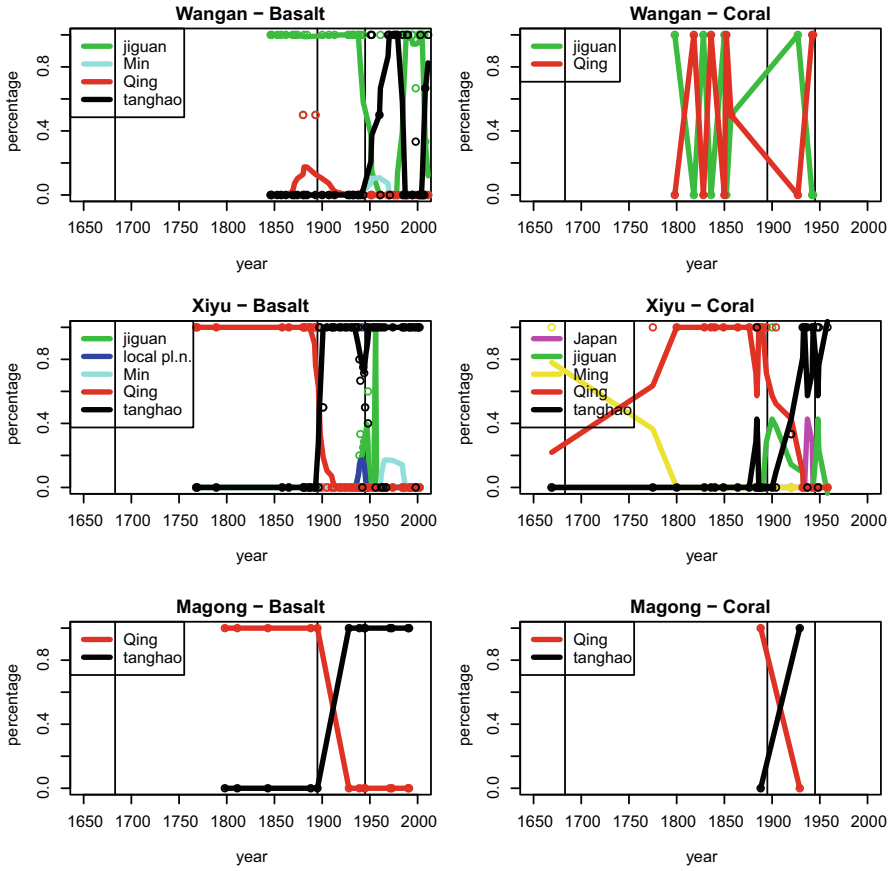



**Fig. 3.22** The transformation of the focus on Penghu through time

The distribution of the character '𥑮' provides us with an additional piece of information: Exported tombstones are more likely to be of basalt. The reason for this might be that, when ordering a stone from a professional stone carver and not from a local worker, one wants to obtain a professional product. Likewise, the professional carver wants to show his professionality with a hard stone, applying to the stone knowledge and skills that the untrained worker does not have. Coral rocks travelled an average distance of 5 km, from Xiaochijiao in Xiyu and basalt travelled 11 km. Surprisingly, tombstones made of concrete seem to have travelled furthest, an average of 13 km. According to our understanding however, tombstones in concrete have not travelled physically, as they are the product of local workers who create the tombstone on the graveyard, usually where and when professional carvers do not serve tombstones.

We thus observe in the Qing period the existence of different cultural centers, not all of which are also military centers. Xiyu was a military center, while Wang'an was

not. Xiyu used the expression of loyalty, Wang'an maintained the *jiguan*. Makong acquired the central position it has today towards the end of the Qing period and then became the unrivalled center with the arrival of the Japanese navy. This rise of power of Makong under the Qing is reflected by a shift from the *jiguan* to the loyalty expression from 1820 on, a period in which the Qing started to invest in Makong, as documented by a row of construction projects during that time.

## 3.6 皇日, The Rising Sun, Penghu Under the Japanese

The loyalty expression on the tombstones devoted to the Qing, however, plunged the inhabitants of Penghu into despair when the Japanese forces landed, as the practice to carve this focus was no longer opportune and dangerous or even life-threatening when continued. Having witnessed the violent occupation of Penghu, the local people understood the risk of continuing a practice that obviously sympathized with the former regime. The solution to the question of how to carve a focus for their tombstones could not follow any historical example and had to be flexible enough to adapt to a completely uncertain future. Even the option of an opportunistic expression of loyalty to the Japanese emperor (皇日, huángrì) had to be discarded, as Penghu might have returned at any time in the future to the Qing. The Qing also would after an eventual reoccupation of Penghu not have been amused by such a welcoming attitude towards the Japanese. The desperation was probably biggest in Xiyu, where the loyalty expression had been used most extensively for 200 years and it was in Xiyu where, within a very short time, a very important epigraphic invention took place. This invention was actually only possible on Penghu, where wind and weather had eroded the visibility into the past. For centuries, the only way to continue a practice had been to copy the practice of the previous generation, or to receive an imported practice that came by boat from a hired carver. But, Xiyu and Makong were not importing tombstones, they were exporting. They were the trendsetters and there was no way back. Copying the practice of the previous generation had become impossible and the practices of many generations before had been eroded and equally become inaccessible. People were forced to invent a tradition.

Since the tombstones on Xiyu had no alternative semantic role to fill the focus position, i.e., no *jiguan*, no generation number that counts up from a selected ancestor to the deceased, such as found in Meinong, the vacant position was filled with a new element, the *tanghao* (堂號 tánghào). This element, one might claim was not completely new, as all over Taiwan and Penghu there have been seven tombstones with a *tanghao* in the period from 1798 to 1884. Most of them are individual occurrences without systematic relations. Yet, two of them are located in Penghu. In 1830 and in 1884, we find two tombstones of a Wu (吳, Wú) family that uses the *tanghao* '延陵' (Yánlíng), the first in Makong, Caozhang and the second

in Xiyu, Zhuwan. Whether the second tomb, carved 11 years before the Japanese invasion, has served as inspiration for the systematic application of the *tanghao* remains currently an open question.

In contrast, in Makong, where tombstones retain their readability for more than 80 years, carvers could reanimate the *jiguan*, which had fallen out of usage in Makong only 70 years before. The tombstones in Wang'an, which were most probably served in the Japanese period by professional carvers from Xiyu, retained the local tradition of a *jiguan*, although the same carvers invented and promoted in Xiyu the *tanghao*. Comparing in Table 3.2 the material of tombstones in the Japanese period, we see that the *tanghao* was primarily introduced on basalt tombstones and thus most probably through professional carvers.

```
plot.focus <- function(df,island,mat,colors){
plot(0,0,ylim=c(0,1),xlab='year',ylab='percentage',
xlim=c(1650,2000),main=paste(island,'-',mat))
abline(v=c(1683,1895,1945),lty=1,lwd=1)
vec.y <- get.inscription.focus.sub(df)
vec.x <- get.stone.creation.year(df)
tab.form.year <- table(vec.y,vec.x)
tab.prop <- prop.table(tab.form.year,2)
n <- length(get.tab.row(tab.prop))
col <- colors[[rownames(tab.prop)]]
zero <- mapply(funct.plot.line,list(colnames
   (tab.prop)),get.tab.row(tab.prop),col,1,0.18)
legend('topleft',rownames(tab.prop),col=col,lty=1,lwd=3)}
####
plot.mat <- function(df,label,colors){
list.mat <- split(df,list(df$mat),drop=TRUE)
zero <- mapply(plot.focus,list.mat,label,c('Basalt',
   'Coral'),list(colors))}
####
df.focus <- inscription.has.focus.sub(df.penghu)
df.focus.year <- stone.has.creation.year(df.focus)
df.focus.year <- stone.has.period(df.focus.year)
df.focus.year$inscription.semantic.roles.focus.sub
   <- rename.vec.val(
df.focus.year$inscription.semantic.roles.focus.sub,
   'tw','local pl.n.')
df.focus.year$inscription.semantic.roles.focus.sub
   <- rename.vec.val(
df.focus.year$inscription.semantic.roles.focus.sub,
   'ch','jiguan')
df.focus.year$inscription.semantic.roles.focus.sub
   <- rename.vec.val(
```

(continued)

```
df.focus.year$inscription.semantic.roles.focus.sub,
   'th-other','tanghao')
df.focus.year$inscription.semantic.roles.focus.sub
   <- rename.vec.val(
df.focus.year$inscription.semantic.roles.focus.sub,
   'th-bjx','tanghao')
vec.focus <- unique(get.inscription.focus.sub
   (df.focus.year) )
library(hashmap)
hash.focus <- hashmap(vec.focus,1:length(vec.focus))
df.focus.year$wangan <- grepl(', Wangan,',df.focus.year
   $graveyard.name)
df.focus.year$xiyu <- grepl(', Xiyu,', df.focus.year
   $graveyard.name)
df.focus.year$magong <- grepl(', Magong,',df.focus.year
   $graveyard.name)
df.focus.year <- df.focus.year[which(
(df.focus.year$wangan + df.focus.year$xiyu + df.focus.
   year$magong) >0),]
df.focus.year$mat[df.focus.year$stone.material=='basalt']
   <- 1
df.focus.year$mat[df.focus.year$stone.material=='coral
   stone'] <- 2
df.focus.year$ming <- period.is.roc.ming.logic
   (df.focus.year)
df.focus.year$qing <- period.is.roc.qing.logic
   (df.focus.year)
df.focus.year$japan <- period.is.roc.japan.logic
   (df.focus.year)
df.focus.year$roc <- period.is.roc.roc.logic
   (df.focus.year)
par(mfrow=c(4,2))
df<-df.focus.year
label <- c('Wangan','Xiyu','Magong')
list.island <- split(df,list(df$wangan,df$xiyu,df$magong),
   drop=TRUE)
zero <- mapply(plot.mat,list.island,label,list(hash.focus))
```

As the data in Fig. 3.22 show, professional carvers and untrained carvers behaved very differently during this period. Professional carvers adopted quickly a new solution and applied it systematically. Untrained carvers show more random-like patterns, as each tombstone reflects a different approach to handle a particular situation. Some untrained carvers continued to use the expression of loyalty to the Qing, either as an act of bravery that is supposed to challenge the Japanese empire or simply because a carver ignored that world history had arrived on Penghu. While

nonprofessional carvings reflect more personal attitudes, knowledge, life styles, and identities, the professional carvings are designed as market products.

We hypothesize that a systematic nature of tombstone inscriptions is at the heart of any professional tombstone carver, simply because this systematicity leads to a belief in his authority, which is at the foundation of his commercial model. Through a fast and systematic reaction, carvers could show that they reliably master the situation.

We further hypothesize that the reason why carvers on Xiyu did not return to the *jiguan*, which they practised simultaneously on Wang'an, was that the *jiguan*, a family history, had been simply forgotten on Xiyu. Carving a *jiguan* requires the memory of a specific historic placename for each family, which, contrary to the *tanghao*, listed for many surnames in the *Baijiaxing*, could not be provided by the

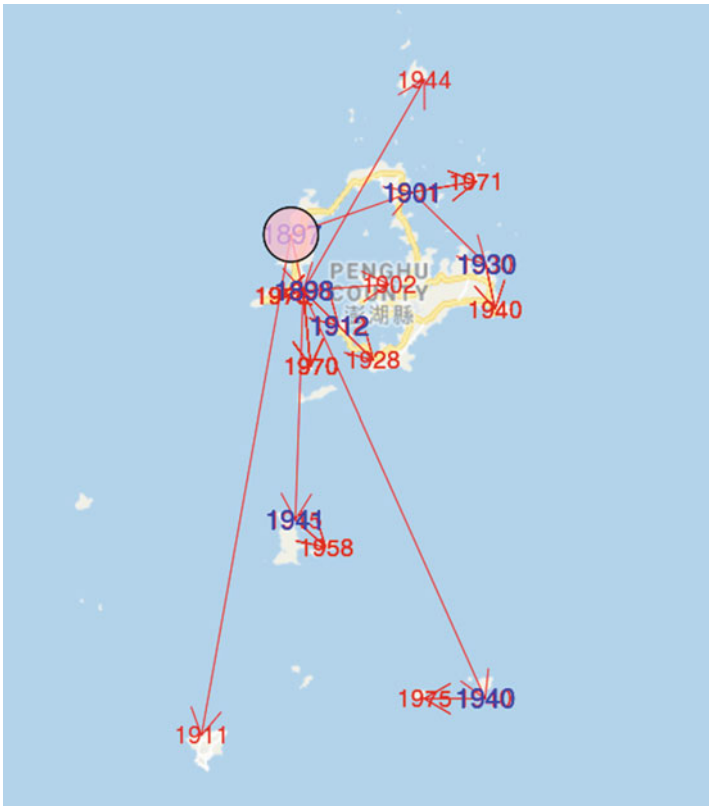



**Fig. 3.23** The propagation of the *Baijiaxing tanghao* from Xiyu over the Penghu archipelago following a model that limits the *p*-value of the space–time correlation to 0.005. Minimizing in addition to this the average spatial distance, the model identifies potential multipliers, i.e., potentially professional carvers. The spreading from carver to carver is colored blue, the spreading from a center to a spatial periphery is colored red

carver. Once forgotten by families on Xiyu, the *jiguan* would have been an imperfect solution to a general problem. At Makong, where we observe, differently from Xiyu, the return to the *jiguan*, the *jiguan* might have been readable on family tombstones and thus have still been part of the active memory. After all, the tombstone is not only the expression of a social identity, it is for many families also the memory of this social identity.

After the WWII, the practice of the *tanghao* expanded to all corners of the archipelago, even to the smaller islands, such as Jiangjun'ao, where before the arrival of the *tanghao*, tombstones had no inscription. This massive application of the *tanghao* was largely the result of the import of tombstones from two carvers, one in Baisha and one in Makong, who towards the end of the 20th assumed an absolute monopoly on the archipelago. What unites these carvers is that the carver in Baisha and the father of the carver in Makong did their apprenticeship in Yanshui, Tainan, Taiwan, with a Master from Penghu. After their apprenticeship, they returned to Penghu to open their workshops.

Modeling the propagation of the *tanghao* over the islands of Penghu, we have to make first assumptions on how such a feature might spread, forming a conceptual model, and find a solution among thousands of randomly created propagation paths that best fits the general assumptions. One of its many similar outcomes is shown in Fig. 3.23. Marked in pink is the onset of the practice, in blue the potential location of multipliers, i.e., carvers, and in red and orange the arrival of the practice on an island.

```
eval.model<-function(argtab) {
gvs <- unique(argtab$gv1)
pval <- cor.test(argtab$time.diff.norm,argtab$loc.
   diff.norm)[3]
if(is.na(pval)) { return(1000000000000) }
if(pval > 0.01) { return(1000000000000) }
for (i in 1:length(gvs)) {
smalltab <- argtab[which(argtab$gv1==gvs[i]),]
if(length(smalltab[,1]) < 2) { return(1000000000000) } }
return(mean(argtab$loc.diff))}
####
st<-graveyard.has.xy(df.penghu.year)
st<-stone.creation.is.after(st,1894)
st<-inscription.has.placename(st)
st<-inscription.has.date(st)
st<-inscription.loc.is.tanghao(st)
st<-stone.creation.is.before(st,1980)
# merge graveyards that are very close
st <- merge.graveyards.dist(st,0.04)
```

(continued)

```
# create a tab that contains all possible graveyard
   combinations
st.early <- earliest.token(st)
mytab <- cbind(
expand.grid(x=st$graveyard.id,y=st.early$graveyard.id),
expand.grid(x=st$stone.creation,y=st.early$stone.
   creation),
expand.grid(x=st$x,y=st.early$x),
expand.grid(x=st$y,y=st.early$y))
# giving names to the columns
colnames(mytab) <- c('gv1','gv2', 'date1','date2',
   'x1','x2','y1' ,'y2')
# remove gv1==gv2
mytab<-mytab[which(mytab$gv1!=mytab$gv2),]
# adding time difference and distance
mytab$time.diff <- mytab$date2 - mytab$date1
# remove negative time difference
mytab <-mytab[mytab$time.diff > 0,]
# calculating the distance between the points
mytab$loc.diff <- distVincentyEllipsoid(cbind(mytab$x1,
   mytab$y1),
cbind(mytab$x2,mytab$y2))
# adding pixel points
mypix1 <- LatLon2XY.centered(canvas.penghu,mytab$y1,
   mytab$x1)
mypix2 <- LatLon2XY.centered(canvas.penghu,mytab$y2,
   mytab$x2)
mytab$pixx1 <- mypix1[[1]]
mytab$pixy1 <- mypix1[[2]]
mytab$pixx2 <- mypix2[[1]]
mytab$pixy2 <- mypix2[[2]]
# normalizing distance and time diff
mytab$time.diff.norm <- normalize.vector(mytab$time.
   diff)
mytab$loc.diff.norm <- normalize.vector(mytab$loc.diff)
mytab$id <- 1:length(mytab[,1])
min <- 1000000000000; runs <-100000; trial <- 0; mytab
   $connected <-0
while(trial < runs | min == 1000000) {
trial <- trial+1; tab<-mytab
while(length(tab[,1]) > sum(tab$connected)) {
# reuse a random number of connections of the most
   successful model
if(min < 1000000 & (sum(tab$connected)==0) & (1
```

(continued)

```
   < sample(c(1:10),1))) {
reuse <- sample(model$id,sample(1:length(model[,1]),1))
connect <- which(is.element(tab$id,reuse))
}
# connect randomly one graveyard to another unconnected
   graveyard
else { # select one graveyard from which to connect
selected.gv <- tab[sample(which(tab$connected==0),1),]
   $gv1
connect <- which(tab$gv1==selected.gv & tab$connected==0)
l.sel <- length(connect)
if (l.sel > 5) { connect <- sample(connect,sample
   (6:l.sel,1)) }
}
if (length(connect > 0)) {
tab[connect,]$connected<-1 # mark as connected
exclude <- tab$gv2[connect] # exclude: connected
   graveyards
# remove connections to the same
   graveyard
tab <- tab[which(tab$connected==1 | !is.element(tab$gv2,
   exclude)),] }
}
comp <- eval.model(tab)
if (comp < min) {
min <-comp; trial <-0; model <-tab;
PlotOnStaticMap(canvas.penghu)
arrows(model$pixx1,model$pixy1,model$pixx2,model$pixy2,
   cex=2,col='red')
text(model$pixx2,model$pixy2,col='orange',labels=model
   $daté)
mult <-sort.by(model,model$datē,dec=F)
mult <-unique.by(mult,mult$gv1)
text(mult$pixx1,mult$pixy1,col='blue',labels=mult$datē)
starting.points<-which(!is.element(model$gv1,model$gv2))
points(model$pixx1[starting.points],model$pixy1
   [starting. points],
col='black',cex=3,pch=21,bg=adjustcolor('pink',
   alpha.f=0.2)) } }
```

Yet, we still can only speculate, how the *tanghao* come towards the end of WWII to Makong and Wang'an, where, after all, the *jiguan* had been predominantly carved during the Qing and the Japanese period. Our interpretation of the data is the following: As the local tombstone carving tradition expired on Wang'an in the

late Qing period, tombstones were served from Xiyu by a carver who respected the local tradition of the *jiguan* on Wang'an. Towards the end of the Japanese period, probably one or several sons of the Xiyu carver moved to Makong to increase their revenue, introducing the *tanghao* in Makong and Baisha. The hypothesis of this movement is supported by fact that around this time the practice of carving a *tanghao* started to decrease in Xiyu.

## 3.7 Conclusion

Tombstone inscriptions, like languages, can thus be said to have life cycles and to play different roles in different periods of this cycle. Tombstone inscriptions may at one point in time be the reflection of an identity based on a geographic origin and turn in later generations into a means of how identities and geographic origins are transmitted. We argued that the difference between Xiyu, where the *tanghao* was invented as systematic focus and Wang'an and Makong, where a *jiguan* was used in the Japanese period, is the result of a literally eroded memory of the *jiguan* on the tombstones of Xiyu.

Through time, the inscription may become obsolete and require an adaptation. These adaptations do not reflect necessarily a changing identity, but the best possible inscription under unstable and potentially unpredictable conditions. Especially, the carvings of professionals, who try to establish systemic solutions that highlight their professional status and guarantee their economic continuity, do not reflect family traditions or social identities. The constructed narrative, that has to introduce the transformation of their product, might leave the first customers in speechless disbelieve, and yet, it might ascend to a national rhetoric or national ideology, from where it provides the next generation with an interpretation of their ancestor's tombs. The newly invented inscription is perceived as a tradition from the moment that previous generations of inscriptions have become unreadable. The interpretation given to these traditions is that elaborated in a state-mediate discourse, potentially inspired by the carvers' original narratives. It is this interpretation of a practice perceived as tradition which is then projected onto the past.

Finally, our work stipulates, in addition to a life-cycle model for the development of cultures and practices, a three-party model of power structures. Theories in the Marxist tradition underline the influence of the superstructure, e.g., the cultural hegemony, or as de Certeau calls it the *strategy* (de Certeau et al. 1980/1990, p. 59). But, de Certeau reminds us equally that there is the creativity of the oppressed, the ruse of the powerless, or as he calls it, *tactics* (de Certeau et al. 1980/1990, p. 60), an ingeniosity however, that usually remains without a lasting impact on the flow of history. In our analysis, we could notice the absence of a significant influence of the untrained carvers on the transformation of tombstone inscriptions. Instead, a third group sneaked into the limelight, the mediator: a poet, painter, priest, or tombstone carver, who has the power to pick up elements of the arts of the powerless, similar to what we have seen on Xiyu, where a carver potentially picked up a singleton

occurrence of a *tanghao* in focus position, and to merge these elements in form and function with the requirements of the ruling class. As our analysis demonstrates, the influence of this third party cannot be underestimated and merits further studies. Accordingly, most of our hypothesis we had to formulate was related to the question how these mediators reasoned, where they moved, and how they tried to survive politically and economically.

## 3.8 A Case for Digital Humanities

Having summarized in a few sketches the transformation of epigraphies on Penghu, we have not yet evoked the scientific paradigm this research is embedded in, the Digital Humanities (DH). Our personal conception of Digital Humanities is that of an empirical approach to the Humanities and Social Sciences that relies on the digitization of the study object and, subsequently, computational methods to analyze or visualize the digitally represented objects. In this sense, DH can be brought to the center of a wide range of academic disciplines. Some of these disciplines involve by definition digital representations, such as corpus linguistics or computational linguistics. Other disciplines integrate the digital approach quite naturally, such as archaeology, geography, history, and musicology. Yet, no matter how much the individual disciplines embrace the digital approach, this trend will transform the scientific landscape and our way of teaching Social Sciences and the Humanities. This transformation is driven by the digitization of the research object, in our study tombs and epigraphic inscriptions. This common denominator in various research activities levels the barriers between traditionally defined disciplines, as the creation and exploration of digital data requires similar if not identical procedures, techniques, and tools across different fields. And, even if tools are different, they are frequently mere variances of more general approaches to extract meaning from data. One of these techniques is the analysis of co-occurrences, called in linguistics *collocations*, in geography *spatial correlations*, and in many other disciplines simply *correlations*. These correlations can be established on the basis of inherent categorial values, such as colors, or by mapping objects onto a common referential space, such as embodied in maps and timelines. High correlations indicate similar meanings. Negative correlations indicate paradigmatic oppositions, i.e., different functions within a common larger referential system. Network analyses as used in many fields ranging from linguistics to art history are equally based on correlations and allow for more systemic views on collections of objects.

As techniques for the creation and exploration of data are similar or transferred from one discipline to another, thematic distinctions become difficult to maintain. This claim might come as a surprise. Yet, using their scientific denomination as pretext, many academic disciplines only pretend to be defined by their research object, as *anthropology* the study of humans, *sociology* the study of societies, etc. Yet, historically these disciplines have been divided by their research method, qualitative in anthropology and quantitative in sociology, or using observation in

anthropology and the experiment in psychology. The advance of polythematic data in one digital format promises to bridge the academic islands that in the last two centuries have become consolidated through institutions, journals, and an academic habitus.

As we show in this paper, the bridging between the academic islands is not only something we can observe to happen, but it is something we need and long for. The study of products of the human hand and brain, the Humanities, and the Social Sciences, the study of human behavior, are fragmentary when considered in isolation. If working with rich data, pursuing a fragmentary approach might produce only scattered insight. If, however, the data themselves are fragmented and incomplete, as in the study of previous societies and languages, every piece of information, no matter which scientific field it might be formally associated with, is worth of being pulled into a pool of data to overcome knowledge gaps through various reasoning techniques which can integrate these data, be they interpolations, analogies, or models.

The research presented here addresses the transformation of practices and to do so, looks narrowly at intertwined linguistic, anthropological, sociological, economic, and archaeological aspects of epigraphic practices. None of these academic disciplines would be in a state to analyze their transformation on Penghu alone. As we have shown, the geological conditions of the islands influence funerary and epigraphic practices and it is the choice of the material, reflecting the economic conditions of the family of the deceased, that triggers the involvement of different kinds of carvers. Different carvers slate stones of different shape and size and produce distinctive linguistic features with a distribution in time and space that reflects the economic state of the home village of the carver, his market share at different times, and thus the potential to disseminate specific epigraphic practices over a certain area. These practices, which evolved out of political and social transformations, ascend to perceived traditions once older inscriptions have become erased by the storm of salt and sand particles that batter the islands. The *tanghao* as a systematic focus of tombstone inscriptions proved to be a successful invention, as it allowed the carver to present a coherent narrative and the community to have an feature that could be woven into the fabric of a social identity with interpretations that ranged from resistance to assimilation. Each of these aspects involved in the transformation of practices and the emergence of a tradition would be irrelevant when considered in isolation. After all, the transformation of practices and the emergence of a tradition involves a material culture, a physical space, a geopolitical constellation, professional groups, their economic models, a psychological reception of their products as well as a discursive elaboration of the products' interpretations.

How to bring these entities together, and how to interpret their correlations is what I tried to show in this contribution, using the ThakBong database and R as principal tools. Similar to fishermen, who maintain their nets and develop the art of casting them out in the sea, digital humanists have to study, elaborate, and sharpen their tools. Only when attaining the agility to manipulate one's tools freely in conventional and unconventional ways, one can trace the hidden story in one's

data that eventually would keep the audience breathless, when the world around them, like basalt, seemingly petrified, transpires as scalding liquid lava of meaning.

**Acknowledgements** The research presented in this paper would have not been possible without the generous support of NSC and MOST for the research projects 99-2631-H-390-002-, 104-2420-H-390-003-MY2, and 106-2420-H-390 -002 -MY3. The author is also grateful to Ann Meifang Lin, Sandy Ke-jui Yen, Naiyu Chen, Sandy Lilun Lin, Ares Jin Tang, Hanna Yaqing Zhan, and James X. Morris for their longstanding support and committed fieldwork on Penghu.

## References

Brown, H. I. (1985). Galileo on the telescope and the eye. *Journal of the History of Ideas, 46*(4), 487–501.

Brown, M. J. (2004). *Is Taiwan Chinese? The impact of culture, power, and migration on changing identities*. Berkeley: University of California Press.

Chen, C.-s. (1953). The Pescadores. *Geographical Review of Japan, 26*(1), 77–88.

Chen, C.-s. (1995). The Pescadores. In *Geo-essays on Taiwan* (pp. 450–466). Taipei: SMC Publishing Inc.

de Certeau, M., Giard, L., & Mayo, M. (1980/1990). *L'invention du quotidien. Arts de faire*. Paris: Gallimard.

Harrison, K. D. (2007). *When languages die: The extinction of the world's languages and the erosion of human knowledge*. New York: Oxford University Press.

Leech, G. (1993). 100 million words of English. *English Today, 9*(1), 9–15.

Ptak, R. (2015). *Fujian – Penghu – Taiwan. Frühe Kontakte, nach Texten zusammengefaßt (ca. 200–1450 n.Chr.)*. Abhandlungen für die Kunde des Morgenlandes. Wiesbaden: Harrassowitz Verlag.

R Development Core Team. (2008). *R: A language and environment for statistical computing*. Vienna: R Foundation for Statistical Computing. ISBN 3-900051-07-0.

Streiter, O., Goudin, Y., & Huang, J. C. (2011). ThakBong, digitizing Taiwan's tombstones for teaching, research and documentation. In *TELDAP 2010 - The International Conference on Taiwan e-Learning and Digital Archives Program* (pp. 146–157). Taipei: TELDAP agency.

Streiter, O., Goudin, Y., Huang, J. C., Lin, A. M.-f., & Yen, S. K.-j. (2010). Places of shared histories: Spatial patterns of placename types on Taiwan's tombstones. In *GeoInformatics for Spatial-Infrastructure Development in Earth and Allied Sciences (GIS-IDEAS)* (pp. 29–34). Hanoi: Japan-Vietnam Geoinformatics Consortium (JVGC).

Theobald, U. (2011). *Chinese literature: Baijiaxing* 百家姓 *"the hundred family names"*. Last modified 2011, accessed August 16, 2013.

Tsao, S., Song, S.-r., Lee, C.-y., & Shea, K.-s. (1999). *Explanatory Text of the Geological Map of Taiwan, Scale 1:50,000 Sheet 73, Sheet 74, Sheet 75, Sheet 76, Penghu Islands*. Taipei: Central Geological Survey.

汪大淵 Wáng Dàyuan (Wang Dayuan). (1339/1981). 島夷誌略校釋 *Dǎo yí zhì è xiào shì (A Brief Account of Island Barbarians)*. Beijing: Zhonghua.

澎湖縣政府民政局 Pēnghúxiàn Zhèngfǔ Mínzhèngjú (Penghu County Civil Affairs Bureau). (2005). 續修澎湖縣志・卷二・地理志 *Xù xū pēnghúxiàn zhì. Juǎn èr. Dìlǐ Zhì (The Annals of Penghu County. Volume XIII. Geography)*. Magong: Government of Penghu County.

# Chapter 4
# Expressing Dynamic Maps Through Seventeenth-Century Taiwan Dutch Manuscripts


**Ann Heylen**

## 4.1 Objectives and Rationale of Seventeenth-Century Dutch Formosa in the Project

Although the application of digital humanities (DH) methods and technology in Taiwan academia is not a recent phenomenon, it has only been in recent years that researchers from a wide spectrum of disciplines in the humanities have been invited to apply for project funding and creating platforms for academic exchange. The purpose of this chapter is to guide the reader through the processes involved in applying digital software to seventeenth-century Dutch handwritten manuscripts that document the presence of the Dutch community in Taiwan. This chapter uses as a case study the digitalized version of the Consistory Records (*Kercboek*) of the manuscript *Kercboek, Brievenboek van Formosa, 23 januari 1642-4 maart 1660* in Dutch transcription and English translation. This manuscript has been preserved as *Resoluties van de kerkenraad van Taiwan (Formosa), 1643 oktober 5–1649 juni* and counts as one of the significant parts of the Resolutions of the Tayouan consistory. This text will be used as an illustration of how to create a geocultural visualization of a semi-structured text with an educational purpose. This intellectual exercise tests the feasibility of DH in bringing the forth the distant past to a general public, and more specifically to an academic audience. As such, the chapter outlines steps in geospatially mapping the mobility of a number of Dutch United East India Company (Verenigde Oostindische Compagnie, hereafter VOC) personnel in Dutch Formosa between 1643 and 1649. My ambition is to demonstrate how the application of exploratory techniques in the spatial humanities offers new insights about the geographies embedded in a diverse range of texts (such as letters, works

A. Heylen (✉)
National Taiwan Normal University, Taipei, Taiwan
e-mail: annheylen@ntnu.edu.tw



S.-H. Chen (ed.), *Big Data in Computational Social Science and Humanities*,
Computational Social Sciences, https://doi.org/10.1007/978-3-319-95465-3_4

of literature, and official reports). This is a starting point to include other materials to broaden and deepen our historical understanding of the period. Inspiration for this idea has come from different directions, but all are equally inspired by the world of virtual and interactive technology that are ideal in representing mobility in "movement." However, prior to outlining the steps taken, attention is to be paid to positioning this project within a larger scope of the NCCU research project and "handbook objectives." This will be followed by a brief setting of the text used as case study for this DH geocultural exercise.

In view of the integrating projects with maps provided by David Blundell, as the principal investigator, the appropriate title and approach to incorporate has been designated as "Mapping Mobility through seventeenth-century Dutch Formosa Historical Writings: GIS on seventeenth-century Dutch Textual Information in Taiwan: A Spatiotemporal Mapping of the Region." Other partners on the team are David Blundell, working on "Mariners, Merchants, Monks: Early Historical Religious Maritime Networks in Monsoon Asia based on the Spread of *Dharma* and Finding Austronesian Connective Evidence utilizing Collective Digital Humanities GIS Mapping Resources"; Jihn-Fa Jan "Integrated Geospatial Techniques for Local Community-Based Volunteer Research: A Method for Digital Humanities and Natural Resources Investigation"; Oliver Streiter with "Digitizing, archiving and analyzing the gravesites in Taiwan: Combining GIS approaches with theories of Social Sciences," and Ching-Chih Lin's "GIS Spatio-Temporal Analysis and Linked Open Data of the Chinese Woodblock Prints (*Nianhua*)."

The integrated project will set up a complete infrastructure for the storage, processing, and archiving of GIS-related data in the digital humanities. This will include a file server, a central data server running PostgreSQL and PostGIS, and a Web-server. The data of the different projects are merged into one database in the same relational data model that this handbook attempts to illustrate and should be seen as a reader companion. As a consequence, data are automatically linked and research can cross academic borders. These data modules will include a set of relevant symbols, languages, scripts, motives, and colors, a gazetteer with place names in various languages and scripts, plus their geo-references for various times, a model on artifacts, e.g., texts, tombs, boats, prints, etc., a model on people of the past, their interrelations, their relations to the artifacts, and of all entities, their relations to place and time.

With this in view, the research objectives of this project are lined up as including data extraction from seventeenth century transcribed and translated VOC manuscripts using GIS and PN recognition; visualizing the mapping of the mobility patterns through an integrated platform, and preparing a scholarly database ready to support not only large-scale text input from the Taiwan-stored digitized VOC archival materials at National Taiwan University Library but also links to existing databases that store historical materials and research articles on seventeenth century Formosa. The collaboration with the other individual projects can be described on the following bases. With Blundell's project, itshares generating PN candidates

that connect to maritime and merchant activities in the Asia-Pacific region. Jan's work will help generate challenges and opportunities for the overall theoretical and technical development of PN identifiers in the geospatial database, because of the non-standardized Dutch language spelling format in seventeenth century historical texts. With Streiter it has in common the application of geospatial mapping seventeenth century place names in Taiwan and the mobility of Dutch VOC personnel in Asia-Pacific region. Finally, with Lin, it seeks the solutions for the application software tools that facilitate genealogy for indexing Dutch people in Taiwan (and Asia-Pacific region) as well as testing tools for the input of Chinese translation data.

## 4.2 Manuscript Kercboek van Formosa: Mission History Goes Digital

The manuscript *Kercboek van Formosa* is a transcription of the Consistory Records or Church Minutes in the Tayouan Consistory from October 1643 to June 1649, and constitutes a particular category of historical writing known as mission history materials. It is written by the clergy commenting on the moral state of the Dutch community and the progress made by the church in evangelizing the natives. Within the totality of Formosa-related documents, church texts are the least studied. Extrapolating this to the study of the VOC, Formosa-related documents are at best complementary, especially in view of the longstanding Dutch presence in other VOC and Dutch West India Company (WIC) settlements. Reasons for the focus on the years 1643–1649 are not arbitrary. The Hollanders had been in Formosa for 20 years; in 1642 the Spanish left, and the Dutch mission post came to include the central and northern parts of the island.

At the same time, Batavia implemented new policies, and the VOC expanded, gaining control over the Strait of Malacca. A limited number of reference works on the settlement of the Dutch community in the seventeenth-century Formosa have been published, but to date the main emphasis has been on tracing the cultural and geographical history of the interaction of the indigenous peoples with the Dutch and Chinese, or positioning Taiwan as a trading depot in the larger framework of Dutch East India relations with neighboring states and ethnic communities (Blussé 2003; Heyns and Cheng 2005; Chiu 2008; Heylen 2016). The Dutch arrived in Formosa not long after the VOC had been established in Batavia and, interestingly, came at a time when the Reformed Church had not been long established in the Dutch Republic. The advantage this text offers is that it lends itself to documenting moves in the spatial mobility of VOC personnel. How did either local or VOC demands that resulted in relocation, displacement and sometimes imposed career changes impact daily life activities?

The analysis of the text has been inspired by research questions from a "literary turn in historical studies" approach, a contextualist perspective that seeks out points of thematic reference, one of which is the observation of social and economic change.[1] Social change is mainly recorded in terms of educational expansion and the development of the Calvinist community, but the Consistory Records also document the scope of mobility by preachers, attendants-to-the-sick, schoolmasters, and other VOC personnel returning to Batavia or the Dutch Republic. This multi-method reading does not merely enable us to note the "conventional practices of time" and how they may have changed; we can also decipher the unusual and various methods that individuals resorted to in response to policies from above and how the Calvinist spirit played out in the lives of its members. What were voyagers' prospects for mobility in going overseas in the service of the VOC, and how did they work out once the travelers had arrived in the distant settlement?

Mobility as a transdisciplinary term caught on in literary and cultural studies, after having been the main emphasis of the sociology and history disciplines (Greenblatt 2010). One of the dynamics through which we can illustrate mobility is education and literacy. Seventeenth-century society witnessed the advancement of popular literacy that had begun in the previous century with the advent of the Reformation (Houston 1988; Graff 1991). Literacy defines the commonality among these different professions in education, and as such helps us to reflect on processes of both upward and downward social mobility. Given the shortage of ordained ministers, a great variety of ecclesiastical ranks, such as licentiates, catechists, and soldiers, were appointed as schoolmasters. Literacy was not especially widespread among the Dutch settlers, so those who could read and write were promoted to the post of schoolmaster. However, this demand for literate personnel was conditioned by the role the church played.

In a recent article (Heylen 2017) I illustrate how Dutch VOC personal were conditioned by a horizontal (lateral or spatial) mobility in their careers in Formosa. My findings suggest that although the voyagers to the East Indies were originally driven by the incentive for upward mobility, for many who landed in Formosa this translated into a sequel of spatial mobility. I showed this through the dynamics of relocation among several villages coupled to the changes in profession, that either revealed an upgrade or downgrade of schoolmaster to soldier, attendant-to-the-sick or interpreter to political commissioner. For instance, promotion to the post of schoolmaster did not always work out well; promoted attendants-to-the-sick tended to ask to be released from the post after sometimes only a few months, whereas the reports on the promoted soldiers were favorable. But of immediate attention, we ask how this text can be treated as a digitalized narrative that extends far beyond its immediate content and provides a digital platform for Dutch Formosa and the VOC?

[1]The approach is exemplified in C. W. Watson, Of Self and Nation: Autobiography and the Representation of Modern Indonesia (Honolulu: University of Hawai'i Press, 2000).

With a focus on mobility and its prospects for visualization, the following three areas of investigation are anticipated to render some new findings. First, detect particular routes where the same villages surface. This can be done through a visible connection between two or more villages with rotating personnel. In addition, the recording of the quarterly visit by the clergy to the villages contributes to the spatial demarcation and its scope of geographical occupation. Second, elucidate the temporal distance in correspondence between Formosa, Batavia, and the patria. Post was by definition surface mail, and usually the names of ships embarking in Formosa, either coming from Japan, or sailing for Batavia are mentioned. It enables outlining the frequency of letter writing and correspondence to and from Batavia, and contributes to the letter post history and its relation to literacy practices. Third, highlight information about the Dutch community and the daily activities of its actors seen from a grassroots level. In particular, their mobility within the island, often coupled to the interchangeability of positions (schoolmaster, attendant-to-the-sick, soldier), suggests a number of theoretical hypotheses regarding the status and role of the Tayouan factory in the VOC. The purpose of this chapter is thus not to present a state-of-the-art DH reading, but to illustrate how researchers from the field of humanities come to learn about and apply one of their topics of their expertise to the digital field. Researchers must identify which steps are to be taken, the opportunities and challenges, and what do we expect to find that is new.

## 4.3 Identifying Terms in Semi- and Unstructured Text

Spatial humanities, a sub-discipline of digital humanities based on GIS and timelines provides an effective integrating and contextualizing function for geo-cultural attributes. As crosswalks for information from multiple sources and in multiple formats they create visual indexes for diverse cultural data. Spatiotemporal interfaces provide new methods of integrating primary source materials into web-based interactive and 3D visualizations. This enables researchers to chart the extent of specific traits of cultural information via maps using GIS gazetteer style spreadsheets for collecting and curating datasets. The system is based on GIS point locations, routes, and regions linked to enriched attribute information. These are charted and visualized in maps and can be analyzed with network analysis, creating an innovative digital infrastructure for scholarly collaboration and creation of customizable visualizations. This method gives the researchers an expanse of data in layers of time across space providing new tools to advance humanistic inquiry. This in turn becomes a web-based bulletin board for local community and scholarly knowledge exchange.

Within the broader context of spatial humanities, the manuscript *Kercboek* furnishes a rich example of geocultural space. Geocultural space means mapping cultural practices and attitudes onto the geographic regions with which they are associated, or using the words of Jane Stadler (2015), p. 134 “referring to a ‘sense of place’ that encompasses the historical and cultural events and narratives associated

with a location." The location selected for this essay is Dutch Formosa, or the seventeenth century Dutch presence on the present day island of Taiwan, and the narrative invocates the mobility of Dutch VOC personnel (lower servants) between 1643 and 1649 based on the Consistory Records (*Kercboek*). Geocultural space thus exists in relation to both the actual location, i.e., the island Taiwan, and the mobility recorded on the map representing events based on spaces of narrative and historical memory. The setting is a cultural narrative depicting the interaction between Dutch, indigenous, and Chinese against the background of the nascent seventeenth century Dutch Calvinist community building. The mobility is visualized through the narrative with georeferents and annotates the accompanying map in a three-fold dimension: local, regional, and global in seventeenth century perspective. The cultural mapping relates directly to the physical landscape, but grapples with the ambiguity and uncertainties such as disappeared place names. In its interdisciplinary mode, it bridges historical source (manuscript), literature (narrative), and geography (placename mapping) with technology (interactive cartography), which is a combination of the "literary turn" in history with the "spatial turn" in cultural theory. The following sections illustrate the steps taken towards the visualization of the manuscript text onto the map.

Preparatory work resulted in the Dutch transcription and English translation of the text. This was achieved, thanks to a previous research grant awarded by the National Science Council (GKH). What was needed now was the formation of a team, consisting of a few dedicated students and an IT instructor to assist in the specific tasks that span from preparing the text, to digital uploading and taking classes in proceeding with the actual data extraction necessary for the compilation of the blueprint map. This team of people I found in Nung-yao Lin, doctoral candidate in geography at NTU with expertise in historical GIS, Dutch national He-On Tsao, master student in Chinese as a foreign language at NTNU, Meng-Lun Teng, doctoral student at the department of Taiwan Culture, Languages and Literature at NTNU, and Benjamin Hlavaty, who sadly recently passed away in August 2016. The first step consisted in preparing the text for digital upload: setting the text into a word or PDF document, so that it can be transferred and copy-pasted into the model software that enables the data extraction process. The usage of https://regex101.com proved ideal and became the working website with which the team learned to give commands for text search and data extraction. The advantage of this website tool is that one can upload any text. Unlike a Microsoft word search, what characterizes this tool is the acquisition of a computer logic language that teaches one to formulate commands based on key words that then search the corpus, and extract the data. Evidently, the specificities of the text and the nature of the word search define the instruction of how to use the tool (Figs. 4.1 and 4.2).

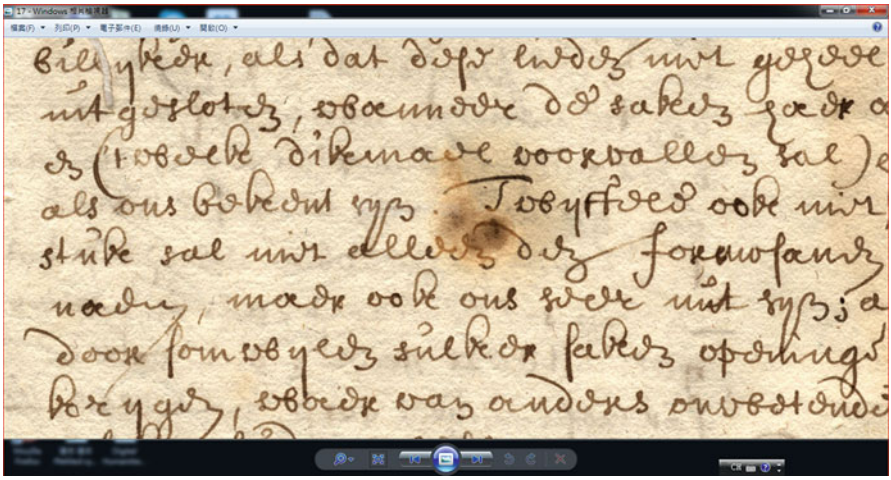


**Fig. 4.1** Manuscript text

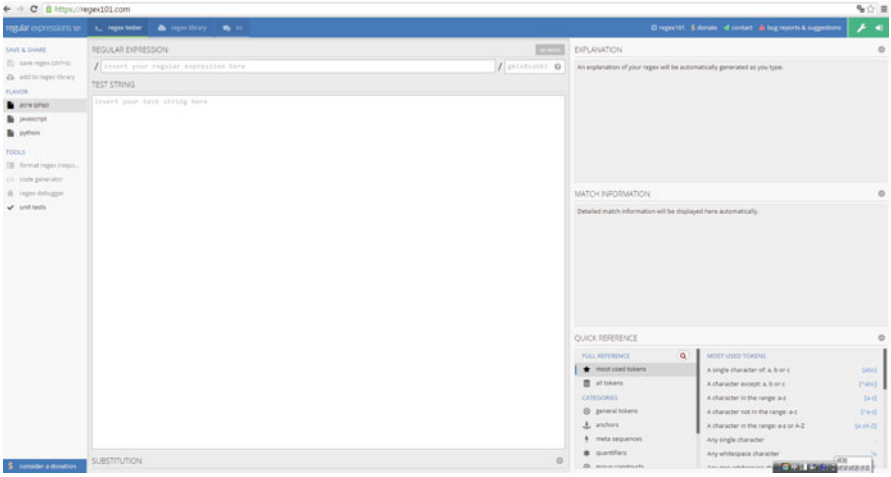



**Fig. 4.2** Regex example

The first series of lessons taught the team how to think in computing language. The fascinating part is the logic needed to adapt to for creating and inserting commands in order to capture terms in the corpus. The text that is to be inserted into the system is called the corpus. The most basic form is commanding Regular Expressions. Computing logic means that every space and character counts.

Capture terms in Corpus- Regular Expression

(可以用 javascript)

G= global

Whole word: /Dominee/ /schip/

空白字: 如 \t \n \r \f \v \d\w

字元族(character class) : 如 o[fn] [abc]e not the whole word, but will only find letters!!

或[a-z] 或 [abcd], [A-Z], [0-9] , [Ii]n \s[Ii]n\s = 有空白 \s

補字元族(negated character classes): 如 [^abc] = abc 不要 ^=not

還原字:找得到符號如：\? 或\. 或 \, \\

Also regular expression find: (dot),*

| | |
|---|---|
| \d= digit | |
| \n= new line | |
| \w= ANY word CHARACTER | |
| | |

\b= boundary \bthethetheater is other

\B= not boundary \bt

^From= 出現在前面一行的句子。 From London and from Taipei

bye$= 出現在後面一行的句子。 Good Bye

r.t rat, rot r..t root

* cu*t Winscho*ten

+至少會出現一次 Winscho+ten cu+t cuter, cuuter, cuuuuter

{} Winscho{1,2}ten

John|MaryCesae?r

---

Tijden:

2 S, 9 October= \d\s[A-Z][a-z]+

3 S, 28 October= \d\d?\s[A-Z][a-z]+

4 S, 23 November= \d\d?\s(October|November|December)

6 S, 3 December= \d\d?\s(October|November|December)

1st December= \d(\d|st|nd)?\s(October|November|December)

2nd November= \d(\d|st|nd)?\s(October|November|December)

Or \d\d?(st|nd)?\s(October|November|December)

2 Session= \d\sS(ession)?

---

Plaatsen:

In Taoyuan

From Soulang

((from|in)\s[A-Z]\w+)

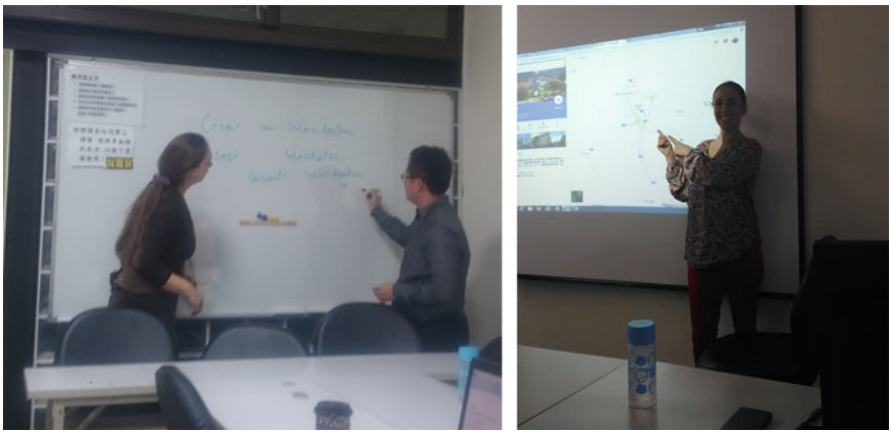


**Image 4.1** Pictured: Dutch VOC personalities "Cæsar van Winschooten" and "Cornelis Cæsar" (left), Mihaela Ionescu illustrates spatial reality of location in a Romanian village (right)

One of the challenges this research presents that is inherent in the historical nature of the texts is the lack of standardized writing in seventeenth-century manuscripts. The lack of a standardized orthography especially pertains to place names and personal names. One of the Dutch VOC personalities who appeared frequently throughout the narrative is Cæsar van Winschooten. The spelling of his name is at times van Winschooten, van Winschoten, or simply as Cesar, which can also refer to another VOC employee called Cornelis Cæsar, as shown in (Image 4.1). Therefore, from the very start the instructor paid particular attention to how to search for this diversity in spelling. Time was thus spent on reviewing the character classes and each character's representative commands. An example would be [abc]e which would search for ae, be, and ce. We also reviewed negated character class and regular expression in the software's command language. We learned the three kinds of systems where one can make a command and the different ways one can type the command (e.g., to search for a digit: \d, [0–9], or [:digit:]). These review lessons were helpful in moving on to the next step of how to search and replace terms and word sections within our body of text, as well as the concept of "boundary." For instance, to find "the" by itself, type \bthe\b, with "b" being boundary, or in this case a space. However, typing the string \bthe will find the word "the" but with a space before it and not after it, e.g., "theatre." This along with several other commands in the regular expression has allowed us to search for a wide variety of place names with a single string of commands. For instance, the place name will most likely be a single word starting with a capital letter and a space before it. We also learned how to check for alternate spellings, as many proper names in the text are spelt differently throughout, e.g., Backluan, Bacloan, Backluang, etc. We have also learned how to search for dates within the corpus, which is actually much simpler to do than searching for locations. An illustration was also provided in class to show us the spatial reality of location in everyday life in action (Image 4.1).

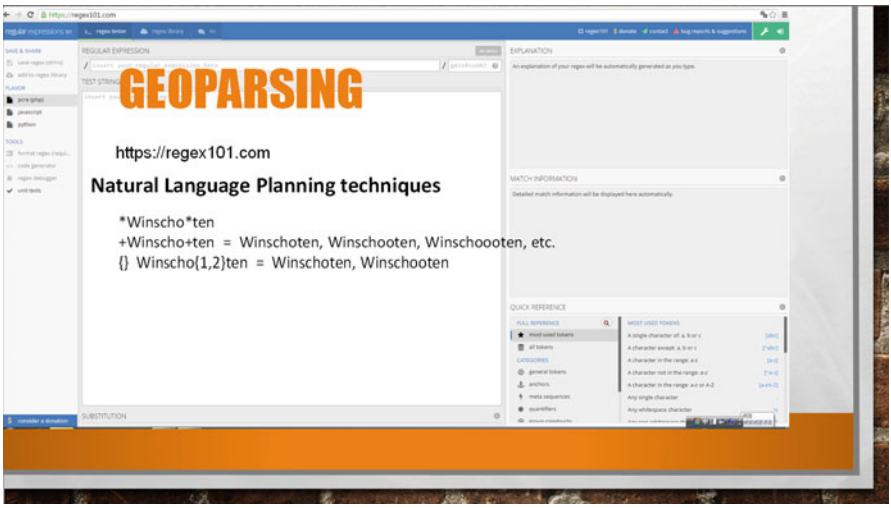



**Fig. 4.3** Geoparsing

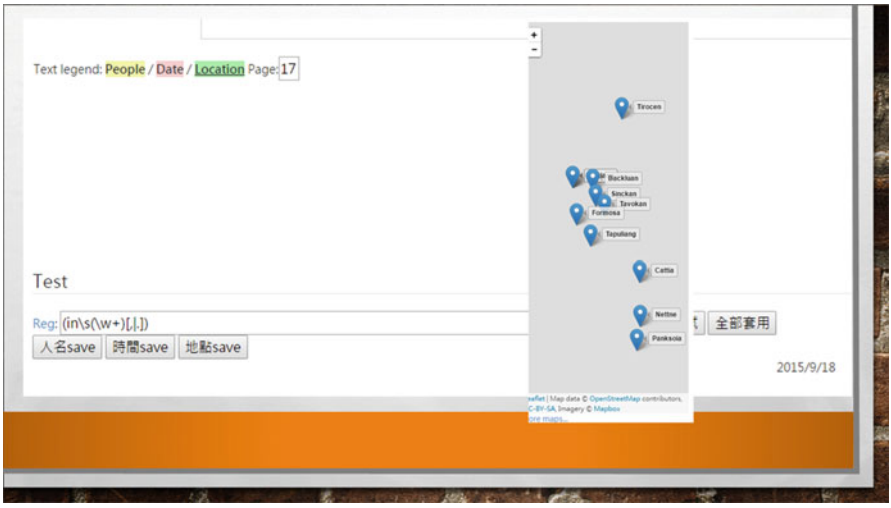



**Fig. 4.4** Map

Armed with these commands, as shown in Fig. 4.3 above, the team was able to generate PN codes in the corpus, which resulted in the website http://church17tw.appspot.com/. The corpus text is inserted and the keywords related to place, date, and names have been highlighted in a different color (Fig. 4.4). Instructor Nung-yao Lin's expertise as a geographer has a specific interest in the compilation of a historical Taiwanese place names map. An excel sheet with georeferents illustrates the coordinates for pinpointing the location. This resulted in a first blueprint

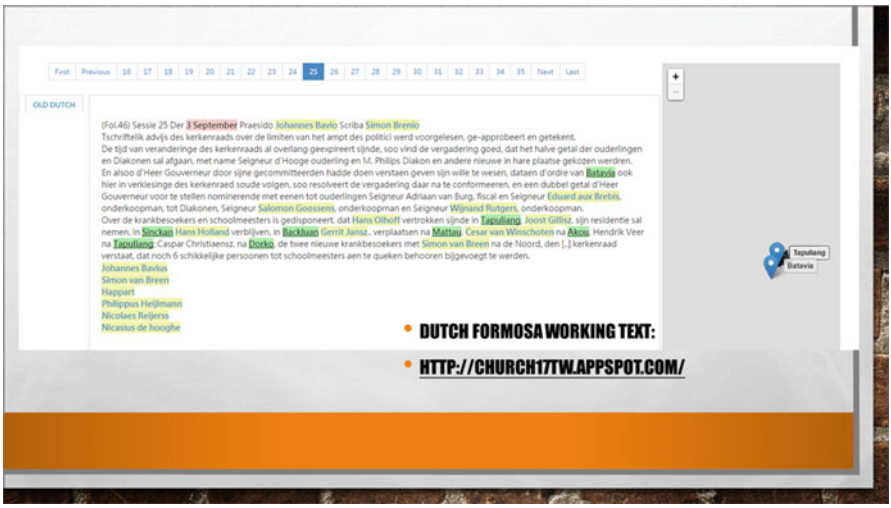



**Fig. 4.5** Map text

integrated text and map platform website. In order to produce the website, the background work involves compiling data onto the excel sheet, which we now term a fusion table, and will come to our attention in the next section (Fig. 4.5).

## 4.4 Visualizing the Narrative with the Help of Faulkner

*The Kercboek* corpus pinpoints the place names and personal names aimed to visualize the narrative. Said otherwise, it brings the Dutch community to life and creates a network cluster that visualizes the movement of schoolmasters, attendants-to-the-sick, and soldiers around the island. This section features the process of bringing this visualization to realization. Our "creative interpretation" centered on how to visualize the narrative, and to which as to paraphrase (Gregory et al. 2015, p. 2) we employed "geographical technologies to develop new knowledge about the geographies of human cultures past and present".

The kind of layout we were looking to implement our "creative interpretation" was found in what we have termed the Faulkner model. This model can be found at the Virginia University website.[2] In the same way we set out to map the people and places in the 17th Dutch-Formosa Consistory Report. In accordance with the

[2]See *Digital Yoknapatawpha*, http://hero.village.virginia.edu/~rwb3y/Faulkner/. Many thanks go to Dr. John Corrigan whose lecture at National Chengchi University, Taipei, in December 2015 inspired the idea.

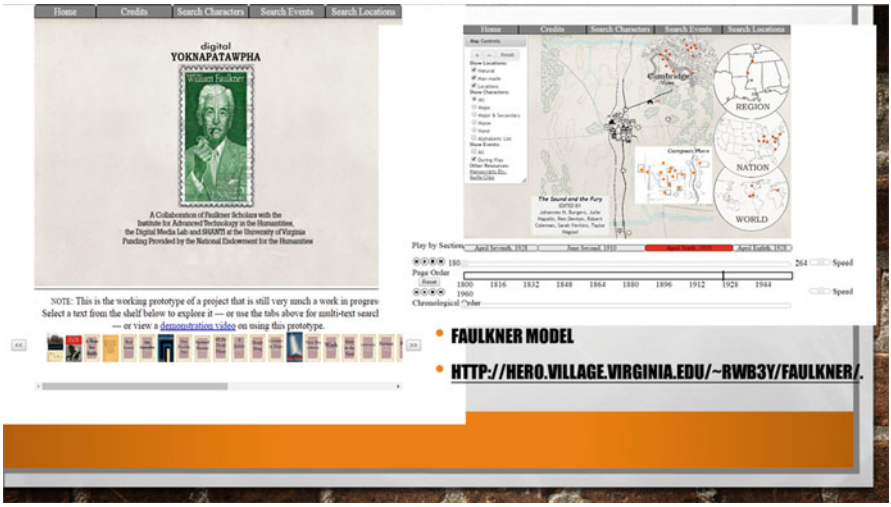



**Fig. 4.6** Faulkner map

teaching method above, we put together a fusion-table spreadsheet mapping out the above sessions from the corpus of text that we are using (Fig. 4.6).

Inspired by this Faulkner model, our geospatial database will provide a list with the personal names that feature in *Kercboek*. We will see at a later stage if it is feasible to include all names, or merely the more significant ones. Our main concern is the location of the place names to show the spatial mobility of the VOC servants around the island. A connection link will be made to region, nation, and world. The Faulkner format also ties in with the further developments of the project. This means compiling a map that has the place names in its different spellings and languages, such as Dutch, Chinese, aboriginal, and also in Spanish where possible. The several variations will be included that will be used as references in the scholarly database, and place names will be complemented with secondary materials accessible in Taiwan. To date a comprehensive database complete with an historical map are lacking.

Figure 4.7 shows that we have utilized what we have learned to highlight the following in the corpus: Personal Names, Place Names, and Dates. To these data are added columns that include the geocoding of the locations, icons which show the professional status of the person (social status of each person regarding his or her title), and additional markers to show the means of mobility (such as various appearances of ship names). These data input are necessary to procure visualization on the blueprint map (Fig. 4.8). The blueprint map itself is divided into three sections on the screen.

In Fig. 4.9 the left side map displays Formosa and serves as the main platform. Figure 4.10 provides an enlarged vision of the Anping region where Castle Zeelandia was built. In the seventeenth century this region was still a landmass

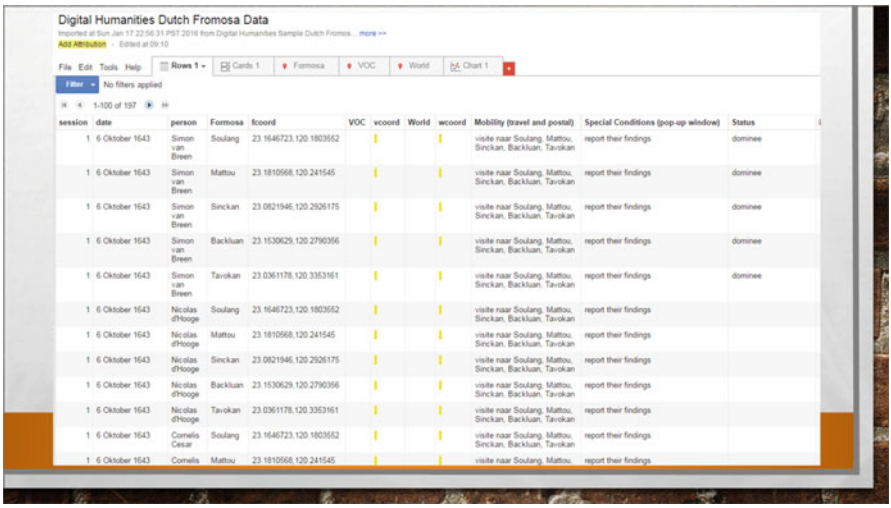



**Fig. 4.7** Fusion table


**Fig. 4.8** Sections blueprint map

whereas now it is sunken in the sea. Anping as we know it today is located further inland compared to its location in the seventeenth century. The Consistory where the Church Minutes (or Consistory Records) were taken was located in the Castle. The Consistory icon thus forms the central point on the blueprint map. At a later stage, the map can be expanded with icons of other buildings mentioned in the manuscript text, such as the church, the poorhouse, schools in the villages, the prison in the castle, or the warehouses. The two maps to the right side of


**Fig. 4.9** Sections blueprint map

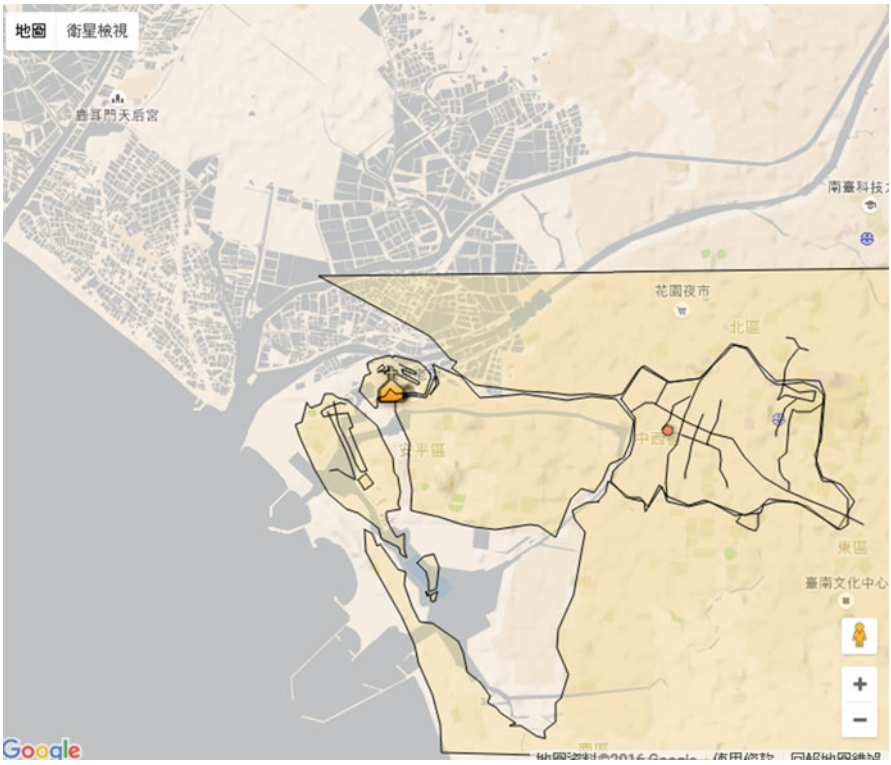



**Fig. 4.10** Enlarged angle

Fig. 4.9 display the East Indies region (top) and the world map (bottom). These are instrumental for demonstrating the spatial mobility of the VOC personnel, whether in terms of arriving from or returning to the VOC headquarters Batavia, cities in the Netherlands, and/or other VOC Factories in the region.

The advantage of the Faulkner model was that it allowed the creation of a similar structure with the manuscript text. The *Kercboek* of the Tayouan Consistory is an ecclesiastical writing consistent with the church regulations in Calvinist countries. Consistories or church councils (*kerkeraad*) were assemblies of ministers and elders who administered church discipline, translating as a form of ecclesiastical justice. Although it was exercised in a purely spiritual way it reveals the all-powerful and pervasive force of religion in the lives of the believers. In this fashion, the Tayouan Consistory kept detailed records of the moral state of mind of its members. The Consistory sessions were chaired by the minister, who acted as the *praeses* (or president) and who was assisted by the scribe, and each session bore signatures of approval from the elders, the deacon, and the political commissioner. The entire text is thus chronologically divided into sessions. The minutes of the Consistory reveals a wealth of data. Each session has a date, the names of the *Praeses*, the scribe, and the board members signing the Minutes. This structure enables us to represent the content of each session on the blueprint map, and this is done through the Play by Section function. Both the Dutch transcription and the English translation are included, and the user can interchange between both languages.

Due to the quantity of data each session contains we feel that it is necessary to make some choices. We decided to associate personal names with the place names. This is in accordance with the setup: to demonstrate the mobility of VOC personnel. The character icons that appear on the map have extended pop-up window functions when selected. When we click on the icon the pop-up window is as comprehensive about the personality of the icon as it is in the entire narrative (Figs. 4.11 and 4.12).

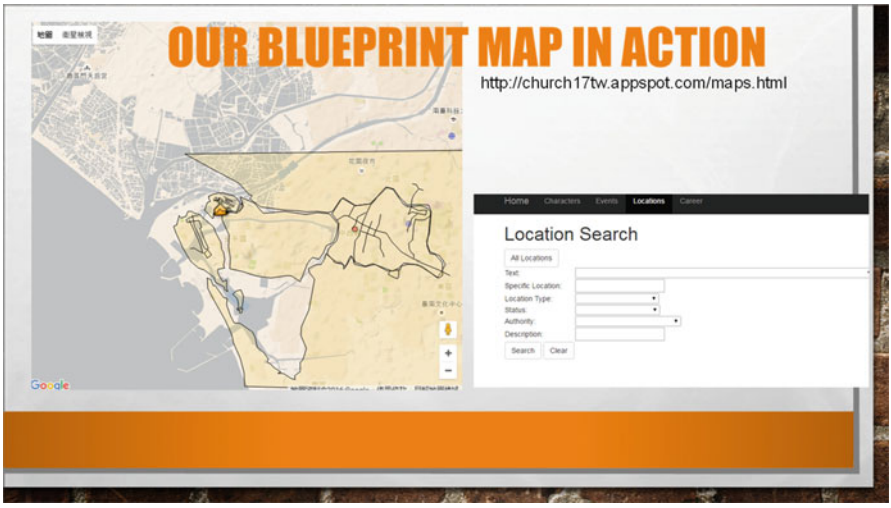



**Fig. 4.11** Sessions text

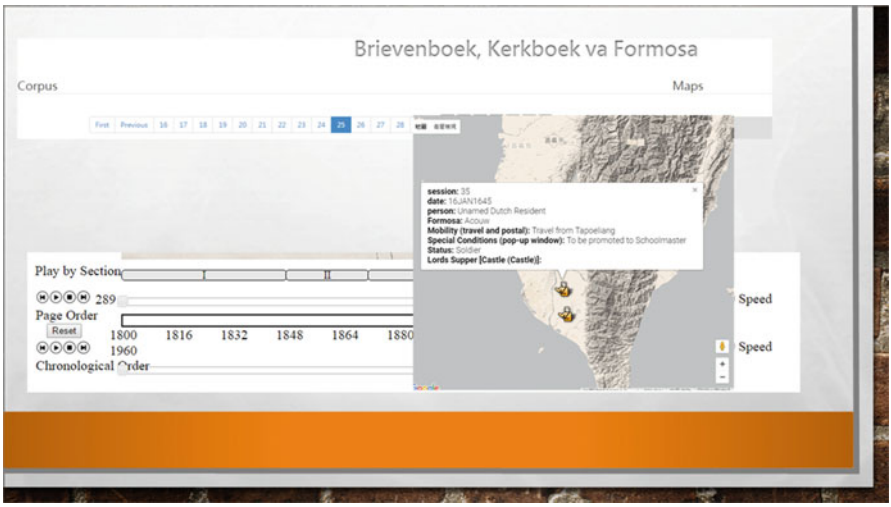



**Fig. 4.12** Pop-up window

The names of the board members who signed each session do not appear as icons in the map, but instead appear in the Consistory. When the Consistory is clicked on the window that appears is not as comprehensive as other icon windows, this instead shows the board members involved in the meeting and the minutes according to the session highlighted. However, the people in the minutes narrative do appear as separate icons, and clicking on them generates a window with comprehensive information about that individual as it is shown throughout the entire narrative. The reason why the Consistory is different here is because it is the actual location of every session that takes place and where the Church Minutes or Consistory Records were written.

## 4.5 Digital Exercise in Mobility

Not all designations for keywords and icon placement are equally straightforward. This section will highlight some examples from the text and draw attention how to visualize its mobility. On the toolbar in the map each section or text entry of the corpus is arranged in a numbered bar that corresponds to the number of the session, and as can be seen in Fig. 4.11. The first example features the sentence in bar/session 9, which is the entry on January 2, 1644:

> Paulus Stroes Vendrig requests support for the widow of some soldier who died during the latest excursion to the south.

This calls for a classification of "nameless people"—rather than widow—and "excursions" as part of the mobility. Equally pertinent is the question how to

visualize the notion of future movement on the blueprint map. Bar/session 14 with the entry on 27 January 1644:

> The visitors, having returned from the south, report to the meeting: Merkinius professed bad conduct in [carrying out] his duty: Hans Valland was worse, and many substantial complaints were heard against him from the elders in Nettne and Cattia.

On the map we see a Personality Icon for Merkinius and Hans Valland appearing on the villages Nettne and Cattia pinpointed as icons in the Formosa section. When you click on the icon of these individuals, the pop-up window provides the information about these two people in the text. When you click on Nettne and Cattia, there should be a pop-up of an element such as a school or a list indicating baptism records depending on the relevance to the session's content. The same entry on January 27, 1644 also reported on intended movement:

> The Church Council considered the last point of the 7th session of 5 December 1643 and took into account what the visitors had reported about Merkinius and Hans Valland: [thus they] gave their approval to Hans Olof taking up his residency in Sinckan: Merkinius being kept under closer supervision in Soulang: one attendant-to-the-sick, Gerrit Jansz, being ordered to reside in Tapuliang: Jan Boudewijnsz. in Tirocen: Hans Valland in Backluan: Jacob Sandbergen in Tavokan: Daniel Hendriksz. in Soulang.

The visualization thus goes as follows: Apart from Merkinus who stayed in Soulang, the other personalities in this session are all about to be on the move. The Consistory decided over their location and for some this translated into relocation. Thus, when we click on the session, the mobility of the person is shown with an arrow giving direction to. To the left of the Hans Olof Icon appears an arrow showing direction to Sinckan marked on the Formosa map (but he has not yet moved there). The same representation goes for the other personalities in that session. Although, when we click on the icon mobility, and search for the name Hans Olof, we see all the movements he made throughout Formosa. In an extended version, this is visible for all the key persons in the narrative, with the purpose to show their spatial mobility.

Bar 31, with an extract taken from (Fol 60) 1645, 35th Session of 16 January. Chairman D. Joannius Happartus, clerk Jannius Bavius

> The Church Council agrees that Acouw, Netne and Katia, which are lacking in schoolmasters, will each be provided with a Dutch resident from the best ones who are currently living in Tapoeliang; also that Joannes Olhoff will take good care that the children of the several villages who are still running free in the countryside will be attracted to the schools.

This relocation of Dutch schoolmasters from Tapoeliang to Acouw, Netne, and Katia is thus visualized with an icon for the unnamed schoolmaster with an arrow from Tapoeliang pointing towards the three southern villages. However, in the pop-up window information section, it could be mentioned what the follow-up was, but that is not necessary, only if the unnamed schoolmaster becomes known later on. It could be that there was no follow-up of this relocation, in which case it should be marked as equally meaningful.

The following entry brings to the fore some of the more tricky elements in the visualization process. In addition to the textual representation below, the image in the fusion table shows how it will be made visible on the map.

## 4.6 (Fol. 38) 21 Session Chair Johannes Bavio, Clerk Simon Brenio, 26 July

The Rev. D. Johannes Happart [ICON], former preacher in Walcheren [WORLDMAP], appears before the meeting. He had been sent here from Batavia [VOCMAP], to be the preacher in this district. After showing his Hon. his attestation, he was welcomed by the meeting and accepted into service. Attendant-to-the-sick in Tapuliang [FORMOSAMAP], Gerrit Jansz. [ICON], who had complained several times about Cæsar van Winschoten in his letters [BOOK ICON], which were quoted at the meeting, presents his accusations in writing against Cæsar. The Church Council agrees that, after both parties have been heard separately and together, D. Johannes Bavius and Simon van Breen and also the elder Sr. Boon would be appointed to hear the case once more, keeping the resolution for the complete meeting [to be held]. Alexander [ICON], who had come here as attendant to the sick with the ship Haerlem [MOBILITY], asks to be put in the service of the church in Formosa. The Church Council is of the opinion that the matter should be postponed until the arrival of François Caron [ICON], which will be any day now.

Johannes Bavius [CONSISTORY]
Simon van Breen [CONSISTORY]
Nicasius de Hooge [CONSISTORY]
Pieter Boon [CONSISTORY]
Philippus Heijlman [CONSISTORY]
Nicolaes Reijerss [CONSISTORY]

For the visualization on the map, we chose icons available on the web. Interesting in this entry is the representation of the notion of literacy, exemplified through the sentence in which Gerrit Jansz complains about Cæsar van Winschooten. In order to illustrate these notions of literacy and its mobility frequency at the time in the narrative, a keyword selection such as “letter,” “read,” “send,” “sent,” “write,” and “writing” will be used for our digital exercise. The purpose of this exercise is to deduct how many letters were written, the length of time they were en route, and their frequency. This interest in the relation between literacy and mobility equally pertains to identifying information relating to the movement of information such as the distance, routes, and time it took information to spread. Ideally, we like to extract information about how many times and in which months a ship arrived, how to picture the “postal service” on the island. By all means, this information is complementary to the existing work and database on the VOC shipping history. Closer to home, in Formosa, the documentation of seventeenth century Taiwan geography features the findings for the keyword “visits” and related geographical terms “north” and “south.” This will enable us to measure the distance between the villages. Bringing all these findings together will result in a cluster network representation, and ideally reveal some findings that we do not see when reading the text manually. Such is the contribution of the intellectual digital exercise. What we have achieved so far is presenting the shift in focus from information about

the text (person A in place B) to information about the world that the text is presenting (his relationship with other people). We thus are able to map the spatial mobility, for instance, Van Winschooten moves from Tapoeliang to Akauw, we also map the social mobility, Van Winschooten as the schoolmaster becomes a political interpreter, and we map the interactions with others which is revealed through the administered church discipline that has consequences for his career moves in the island inextricably related and impacting his spatial mobility.

## 4.7 Generating a Scholarly Database

The final part of this project consists in linking a scholarly database to the mapping platform. This is our aspiration with enhancing the academic application and making it accessible to as wide and varied a public as possible. At this point, we need to steer away from the purely digital treatment of the narrative and say a few words about the role this seventeenth century Dutch historical episode occupies in the larger framework of Taiwan Studies and Taiwan historical materials. For the informed researcher and audience the study of Taiwan society, culture, and history is new in a sense that it is being formatted into an academic discipline of recent origin. To date there is still a lack of a comprehensive historical place name map and a centralized database that encompasses the scattered smaller and individual efforts on bibliographical references documenting Taiwan in the seventeenth century. Nonetheless, major efforts are being undertaken in the Chinese translation of seventeenth century manuscripts. Moving towards open access and public domain research accessibility is becoming an option within the Taiwanese academic world. Hence, we see it our task to create a digital scholarly platform adjacent to the digital map that provides relevant and updated information about where to find either full downloadable access to scholarly articles, databases, and other academic and popular products that guide and feed our knowledge about seventeenth century Dutch Formosa in a "glocal" setting. In order to make the scholarly database attractive and fitting with the overall purpose of mapping, this project gives digital access to a number of historical maps complementary to the creation of its own digital map, as Figs. 4.13a–c show.

## 4.8 Conclusion

This chapter has treated a semi-structured text as a geocultural space, and applied the notion of geocultural space in cartographic initiatives, showing implications for the ways in which geographic and environmental spaces are culturally understood. Working with a digitalized content will enable a view of the key figures in the Dutch community, their social networks and mobility in Formosa and other VOC factories, which calls for an expansion beyond Formosa, and interlinking with

existing digitalized sources of the VOC. The setting is a cultural narrative (the interaction between Dutch, Chinese, indigenes, plus seventeenth century Dutch Calvinist community practices). The mobility visualized through the narrative has geographic referents (georeferents) and annotates the map (in three layers, dimensions: local, region, world) with seventeenth century perspectives. Its cultural mapping relates directly to the physical landscape, but grapples with the ambiguity and uncertainties such as disappeared place names. The interdisciplinary field of

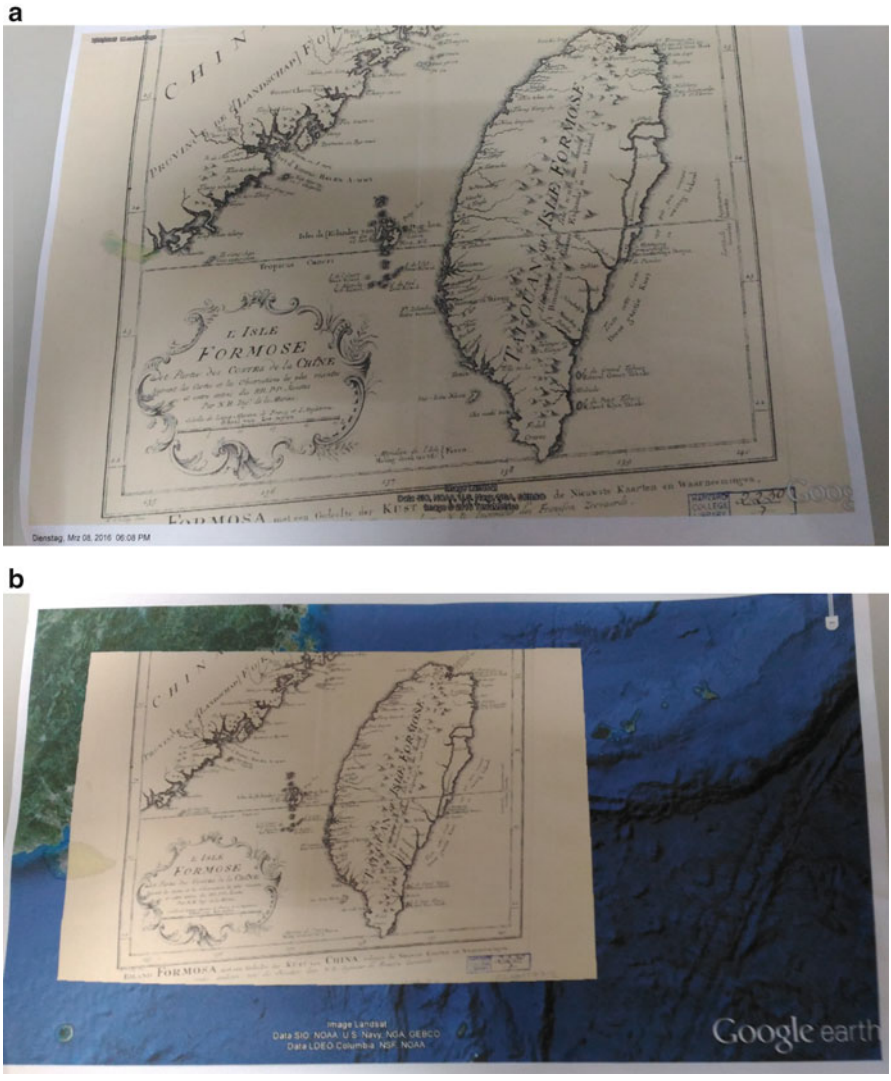



**Fig. 4.13** (**a**) Physical map. (**b**) Digitized map inserted into GIS platform. (**c**) Digital map

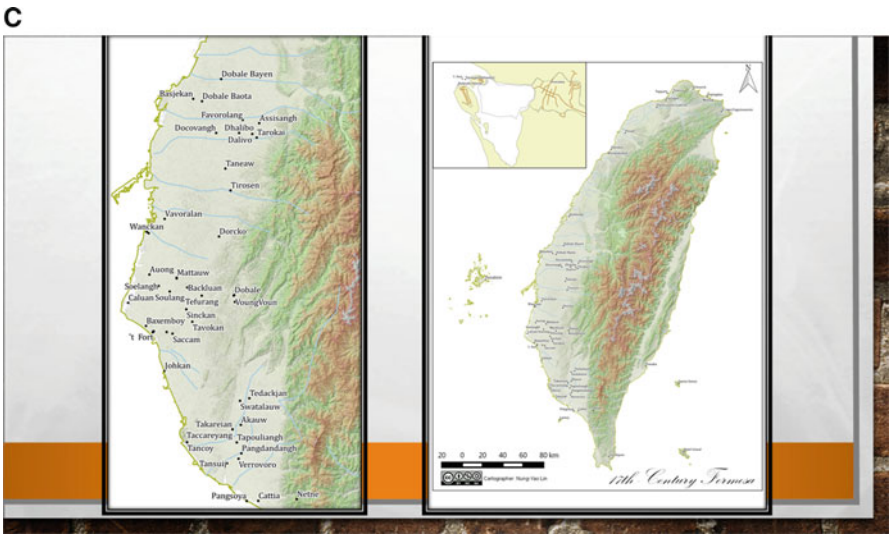



**Fig. 4.13** (continued)

research features historical source (manuscript), literature (narrative), and geography (placename mapping), with technology (interactive cartography) that allow to record the mobility on the map, with deep mapping of events based on spaces of narrative and historical memory. Dutch Formosa in a kindle format brings a visualization that shows something new about the cultural understanding we have today about Dutch Formosa's geographic and environmental spaces.

## References

Blussé, L. (Ed.). (2003). *Around and About Formosa. Essays in Honor of Professor Ts'ao Yung-ho*. Taipei: Ts'ao Yung-ho Foundation for Culture and Education.

Chiu, H. (2008). *The colonial 'civilizing process' in Dutch Formosa, 1624–1662*. Boston: Brill.

Graff, H. J. (1991). *The legacies of literacy: Continuities and contradictions in western culture and society*. Bloomington, IN: Indiana University Press.

Greenblatt, S. (Ed.). (2010). *Cultural mobility: A manifesto*. Cambridge: Cambridge University Press.

Gregory, I., Donaldson, C., Murietta-Flores, P., & Rayson, P. (2015). Geoparsing, GIS, and textual analysis: Current developments in spatial humanities research. *International Journal of Humanities and Arts Computing, 9*(1), 1–14.

Heylen, A. (2016). Taiwan in late Ming and Qing China. In G. Schubert (Ed.), *Resoluties van de kerkenraad van Taiwan (Formosa), 1643 oktober 5–1649 juni*, Hoge Regering 4451, Arsip Nasional Republik Indonesia, Jakarta.

Heylen, A. (2017). "Mobility and consistorial discipline in Dutch Formosa: An examination of kercboek, 1643–1649, Tayouan Consistory. *Taiwan Historical Research, 24*(1), 1–36.

Heyns, P., & Cheng W. (Eds.), (鄭維中, 韓家寶譯著). (2005). 荷蘭時代臺灣告令集婚姻與洗禮登錄簿 [Dutch Formosan placard-book, marriage, and baptism records]. Taipei: Ts'ao Yung-ho Foundation for Culture and Education 《荷蘭時代台灣告令集婚姻與洗禮登錄簿》, 台北:曹永和文教基金會.

Houston, R. (1988). *Literacy in early modern Europe: Culture and education* (pp. 1500–1800). London: Longman.

*Resoluties van de kerkenraad van Taiwan (Formosa), 1643 oktober 5–1649 juni*, Hoge Regering 4451, Arsip Nasional Republik Indonesia, Jakarta.

Stadler, J. (2015). Conceptualizing and mapping geocultural space. *International Journal of Humanities and Arts Computing, 9*(2), 133–141.

Watson, C. W. (2000). *Of self and nation: Autobiography and the representation of modern Indonesia*. Honolulu: University of Hawai'i Press.

# Chapter 5
# Has *Homo economicus* Evolved into *Homo sapiens* from 1992 to 2014: What Does Corpus Linguistics Say?


**Yawen Zou and Shu-Heng Chen**

## 5.1 Introduction

*Homo economicus* has its early roots in the age of classical economics, when economics was dominated by political economy. Mill (1874) described humans as solely caring about wealth possession and making sensible decisions on their own behalf. These characteristics of *Homo economicus* then evolved over time, especially after Walrasian general equilibrium analysis eventually superseded political economy, when *Homo economicus* became super-rational and infinitely smart, and human decisions were neither dictated nor influenced by emotions. The use of a mathematical optimization framework which was heavily borrowed from operations research, and was developed from the 1940s to the 1960s, laid the analytical foundations of such human behavior and hence helped this paradigm to evolve.

In addition to economics, *Homo economicus* has also been introduced to other sister disciplines in the social sciences, a spread generally known as *economic imperialism*. However, not all social scientists can embrace *Homo economicus* as economists did; in fact, a negative side of economic imperialism is that economics became more alienated from her sister disciplines, and economists sometimes considered economics to be a "brother" of the physical sciences, including mathematics. The communication between economics and other social sciences turns out to be much weaker than an outsider can possibly imagine. Humans, in the eye of

Y. Zou (✉)
Center for Biology and Society, Arizona State University, Tempe, AZ, USA

The Chinese University of Hong Kong, Shenzhen, China
e-mail: yzou20@asu.edu

S.-H. Chen
AI-ECON Research Center, Department of Economics, National Chengchi University, Taipei, Taiwan



S.-H. Chen (ed.), *Big Data in Computational Social Science and Humanities*, Computational Social Sciences, https://doi.org/10.1007/978-3-319-95465-3_5

most economists, behave in a way that is rather peculiar in so far as other social scientists can perceive. This difference, well characterized by the increasing gap between what people *ought to do* and what they *actually do*, created a degree of uneasiness between economists and other social scientists. This gap also gave rise to the distinction between "good economics" and "bad economics"; a world filled with *Homo economicus* cannot possibly experience a financial crisis (McDonald 2009).

To narrow the gap and lessen this uneasiness, economists began to work with an alternative *Homo*, namely *Homo sapiens*. Over the past few decades, behavioral economics has illuminated the role of the social, cognitive, and emotional aspects of humans in economic decision-making (Dohmen 2014; McDonald 2009), and hesitates to characterize humans as self-interested, utility- or profit-maximizing machines. In his popular book *Not Just for the Money*, Frey (1997) distinguished intrinsic motivation from extrinsic motivation (the pecuniary motivation) and highlighted the neglected importance of the former in understanding human choices and behavior. Through his life-time devotion to the study of cooperation and altruistic behavior, Nowak and Highfield (2011) argued that both altruism and cooperation were both significant in terms of the survivability of an individual or even the whole of mankind.

Unlike *Homo economicus*, *Homo sapiens* is what Herbert Simon coined as being *boundedly rational*, and is constrained by limited memory and computational capacity. The decisions made by such humans are, therefore, further constrained by a great variety of social settings and emotional drives. Based on these developments, Thaler (2000), a leading behavioral economist, predicted that, in the world of economics, *Homo economicus* would eventually "evolve" into *Homo sapiens*. Since more than 15 years have now passed since Thaler made his insightful prediction, it seems to be a good time to evaluate whether Thaler's prediction is correct in the sense that the research paradigm characterized by *Homo economicus* has been gradually replaced by the paradigm characterized by *Homo sapiens.*

The underlying assumption of this study is that the changes in these two research paradigms are reflected by a change in the language used by economists. Research articles are an important genre of the economists' discourse, and this era of big data has bombarded us with a huge number of readily available research articles. When "manpower" is too limited to surf over such an extensive "ocean" of studies, adopting an automated or computational approach toward a corpus built from a myriad of research articles has become increasingly practical (Biber et al. 1998). A corpus is a collection of textual data, and it appears in the form of either written languages or spoken languages. Written language can comprise texts retrieved from online websites, like Twitter, Amazon reviews, eBay product descriptions, etc., or it can be retrieved from research articles, as in our case (Knight et al. 2015).

Many humanists have used the techniques of corpus linguistics to gain knowledgeable insights into a particular domain that interests them. Corpus linguistics was initially used by linguists to study language patterns, but it was then adopted by scholars for different areas of investigation. Scholars can apply the corpus linguistic approach to study any section of the research articles, such as the abstract, discussion sections, and methodology sections (Flowerdew 2015). The approach has also been

used to study the changes in concepts and ideas. For example, Pumfrey et al. (2012) analyzed the changes in the meaning of the word "experiment" from early English books, using both the corpus linguistics approach and the manual approach, i.e., a traditional close reading of randomly selected texts. It was found that the former approach proved to be more efficient than the latter.

This chapter is, to the best of our knowledge, a pioneering application of corpus linguistics to economics. We shall first address the issue of how to build a corpus that can represent the economics literature and introduce a tool for the corpus linguistic analysis. Second, to have a concrete illustration, we shall use the stemmed word, "cognit*," to demonstrate the analysis conducted in this paper. The same kind of analysis will be applied to other 100 keywords, which are picked by the machine and a human expert. Third, we further divide the whole corpus into two periods, namely before and after the year 2000, and carry out a co-word network analysis to examine any discernible changing pattern. Finally, we summarize the main findings of our study by remarking on its current limitations and the opportunities for future research.

## 5.2 Methods

To trace the evolution of a field, one approach that economists are very familiar with is bibliometrics, a methodology that has long been used to trace the specific dynamics of a discipline (De Bellis 2009). The approach demonstrated herein, however, is atypical; it is motivated by the recent applications of corpus linguistics and network analysis in studying the history of ideas. Since this approach is not familiar to most economists, providing a brief background should be useful.

To begin with, we wish to point out that neither *Homo economicus* nor *Homo sapiens* refers to a well-defined research discipline or methodology. This lack of articulation makes the conventional bibliometric approach hardly applicable to tracing the development of the two paradigms. In fact, what really interests us is whether the idea of *Homo sapiens* has spread widely in different territories of economics, and is not just confined to some specific fields, such as behavioral economics, health economics, or labor economics, or to a specific methodology, such as experimental economics, computational economics, or neuroeconomics. In other words, the framework that appears to be more suitable for us is the *history of ideas* (Lovejoy 2011).

There is a well-established area called the history of economic ideas, also known as the history of economic thought, to which little attention has so far been paid (Blaug 2001). As in the study of other histories, the methodology used for the study of the history of economic ideas is mostly narratives-oriented. It is usually based on the study of a number of magnum opuses and influential articles. Normally, we can learn from these painstaking studies and infer when an idea of interest was first introduced, why it was needed, how it was formulated, and how it then evolved, with each incidence or change supported by major references. Needless to say, this classical narrative approach will continue, since, as one may rightly argue, only man

can read and think. However, these studies are time-consuming, and researchers using this approach will find it harder to keep up with the rapid growth of the literature. Furthermore, no one can be assured of what may be missing given the voluminous amount of literature.

The recent information and communication technology revolution provides us with a new division of labor between humans and machines that makes various novel forms of cooperation between humans and machines possible. One concrete example is digital humanities, the application of digital tools to studying humanities (Schreibman et al. 2008). Not only has the second wave of the Industrial Revolution provided us with new methods, but it has also given us new ontologies and epistemologies. We might then ask: what is the history of ideas, or even more fundamentally, what actually is an idea?

Without losing generality, let us assume that an idea can be presented in the form of written texts, which are in turn collections of words or symbols. Of course, this collection is not just a monotone enumeration of words or a fragmental list of symbols; rather, these words and symbols are embedded within a specific structure, so that they are coherently connected. This description puts the history of ideas in a representation of evolving networks. In this representation, ideas may behave as a biological entity and can be treated as a research object in a complex adaptive system (Simon 1962). A machine may not be able to read and think on the works and ideas found in Shakespeare's plays, but it can help us to dig out the possible networks (structures, patterns) within those works. These extracted patterns or structures may not be complete, but humans can decide how incomplete they are and how they can be made complete. This human–machine interaction can, therefore, fundamentally change our perception of the history of ideas, from its ontology, to its epistemology and methodology.

This chapter, to the best of our knowledge, is the first study in the history of economic ideas that combines the corpus linguistics approach with the co-word network analysis. In the following, we will begin with the basic level of analysis, i.e., a word. A word serves as a node in a network. Hence, being a fundamental unit, it should be treated as the first step toward a fully-fledged study of the history of ideas. We then trace words in a network to see their relationships.

### *5.2.1 Building a General Economics Corpus*

In corpus-based research, the primary task is to build a corpus that is sufficiently representative of a research field that is of interest. As mentioned above, since our interest is to examine how the idea of *Homo sapiens* has been accepted by economists at the global (the mainstream) level, the corpus of this study should be built upon the articles published in mainstream economics journals. Characterizing what mainstream journals are is straightforward. The criterion that we have adopted here is based on ratings. There are many different ratings available. However, since they do tend to overlap quite substantially, we can simply follow one of them. In this study, we follow the rating provided by Kalaitzidakis et al. (2011).

Kalaitzidakis et al. (2011) analyzed the number of citations for the previous 10 consecutive years for articles published between 2003 and 2008 in a number of journals. We pick this rating because their survey covers a longer horizon, as opposed to other rankings, such as the rating in Thomson Reuter's Journal Citation Report, which is based on the Web of Science database[1] and covers only 2 years of citations, and the SCImago journal ranking, which is based on the Scopus database[2] and also covers only 2 years of citations. Since the mainstream should be reasonably enduring, we therefore choose a ranking that is less sensitive to time.

We took the list of the 50 top-ranking journals. Most of these journals, such as the *American Economic Review*, are relevant to almost all major fields in economics, but some are either too domain-specific or too technical. The latter kind of journals may have little relevance to either of the paradigms, be it *Homo economicus* or *Homo sapiens*. As including the journals of the latter kind could blur our research focus, we decided to remove those journals. Since this decision inevitably involved a degree of discretion, we only took a mild cut, i.e., 8 journals out of the 50. Table 5.1 presents a list of the 42 journals selected.

**Table 5.1** The 42 mainstream economic journals

| Index | Journal | Index | Journal |
|---|---|---|---|
| 1 | Am Econ Rev | 22 | J Econ Growth |
| 2 | Qjecon | 23 | J Hum Resour |
| 3 | Econometrica | 24 | J Econ Dyn Control |
| 4 | J Polit Econ | 25 | J Econ Behav Organ |
| 5 | Rev Econ Stud | 26 | J Health Econ |
| 6 | J Econ Theory | 27 | J Appl Econom |
| 7 | J Public Econ | 28 | Brookings Pap Eco Ac |
| 8 | Econ J | 29 | World Bank Econ Rev |
| 9 | J Econ Perspect | 30 | Econ Theor |
| 10 | J Int Econ | 31 | Scand J Econ |
| 11 | J Econ Lit | 32 | Oxford Econ Pap |
| 12 | J Financ Econ | 33 | Can J Econ |
| 13 | Eur Econ Rev | 34 | Econ Inq |
| 14 | Rand J Econ | 35 | Econ Policy |
| 15 | Int Econ Rev | 36 | Int J Ind Organ |
| 16 | J Eur Econ Assoc | 37 | Public Choice |
| 17 | Game Econ Behav | 38 | J Law Econ |
| 18 | Econ Lett | 39 | World Dev |
| 19 | J Dev Econ | 40 | J Law Econ Organ |
| 20 | Rev Econ Dynam | 41 | J Ind Econ |
| 21 | J Labor Econ | 42 | Labour Econ |

[1] http://wokinfo.com/products_tools/analytical/jcr/

[2] http://www.scimagojr.com/journalrank.php?area=2000

**Table 5.2** Number of abstracts and word counts of the sub-corpus in each year

| Year | No. of articles | No. of abstracts | Word counts |
|---|---|---|---|
| 1992 | 2546 | 1260 | 118,795 |
| 1993 | 2459 | 1293 | 125,616 |
| 1994 | 2444 | 1604 | 152,960 |
| 1995 | 2453 | 1686 | 163,976 |
| 1996 | 2715 | 1805 | 176,908 |
| 1997 | 2603 | 1794 | 177,326 |
| 1998 | 2676 | 1876 | 193,881 |
| 1999 | 2485 | 1838 | 193,951 |
| 2000 | 2575 | 1886 | 201,736 |
| 2001 | 2714 | 2017 | 218,614 |
| 2002 | 2708 | 2087 | 223,984 |
| 2003 | 2836 | 2238 | 241,386 |
| 2004 | 2821 | 2262 | 245,601 |
| 2005 | 2734 | 2283 | 251,377 |
| 2006 | 2760 | 2386 | 261,590 |
| 2007 | 3039 | 2645 | 297,015 |
| 2008 | 3291 | 2899 | 322,744 |
| 2009 | 3042 | 2657 | 308,338 |
| 2010 | 3046 | 2701 | 308,201 |
| 2011 | 3115 | 2718 | 317,283 |
| 2012 | 3528 | 3150 | 350,411 |
| 2013 | 3490 | 3135 | 374,583 |
| 2014 | 3397 | 3065 | 371,259 |
| Total | 65, 477 | 51, 285 | 5,597,535 |

We then built a corpus of over five million words using the self-developed Python code to retrieve the abstracts of research articles published from 1992 to 2014 for the selected 42 mainstream economics journals. These included the *American Economic Review, European Economic Review, Quarterly Journal of Economics*, etc. We downloaded the metadata of those articles from the *Web of Science* database. Table 5.2 provides the size of our corpus year by year, from 1992 to 2014. Over the years, the annual corpus size has increased from 118,795 words in 1992 to 371,259 words in 2014.

Thaler made his prediction about *Homo sapiens* in 2000. That year is roughly in the middle between 1992 and 2014, so the time frame employed in this study captured 8 years prior to his prediction and 14 years after his prediction. We picked the year 1992 as a starting point instead of earlier years for a precise reason: the abstracts of most journals were not available in the Web of Science database until 1992. For example, in 1990, the 42 journals of choice published a total of 2169 articles, but only five had abstracts deposited in the database. In 1992, these 42 journals produced 2546 articles, and 1269 of them had abstracts. In total, from 1992 to 2014, the 42 journals produced 65,477 articles, and 51,285 of them had abstracts.

For the numbers of published articles, their available abstracts, and the word counts of the sub-corpus for each year, the reader is referred to Table 5.2.

To facilitate a diachronic study of the language patterns in the abstracts of these research articles, we grouped the abstracts published in the same year into a single file to have a total of 23 files, with each file corresponding to one of the years between 1992 and 2014. We then analyzed the frequency of words in each year.

### 5.2.2 *Analyzing a Selected Word List Using WordSmith*

After building the corpus, we identified a set of keywords to study. A paradigm shift can involve the deletion of some old ideas (represented by words) and the addition of some new ideas. The purpose of having a set of keywords is to identify the deleted and the added ideas. The practice of obtaining keywords needs to be either manually provided by experts or to be automatically generated by computer or by both. We followed the third route by first having the computer search for a list of frequently appearing words. WordSmith is one of the mostly widely used tools in corpus linguistics (Scott 2001).

We first used the WordSmith *Keywords* tool to identify a list of keywords. These words automatically satisfy the necessary condition of being the key, but they are not sufficient. As one could well expect from Zipf's law, many very frequently appearing words are just articles, pronouns, and modifiers (Zipf 1949). Hence, in the second stage, we removed these "nuisances," and used our domain knowledge to select a list of 101 words as the keywords. These are presented in Table 5.3.

**Table 5.3** The 101 words that are related to the two paradigms

| Category | Words or stemmed words |
|---|---|
| *Homo sapiens* (48) | Adaptive, affect*, agent-based, ambiguity, anthropolog*, attitude*, behavior*, bias*, chaotic, cognit*, complexity, cooperat*, cultur*, darwin*, decentraliz*, donor*, emergence, emotion*, ethnic*, evolut*, experiment*, happiness*, heterogen*, imperfect, incentive*, instabilit*, interact, intuition, laboratory, neighbor*, network*, norm, optimism, overconfidence, psycholog*, religi*, satisfaction*, selection*, social*, socio*, stimulus*, subjective, trust, unbiased, uncertaint*, uninformed, well-being*, wisdom |
| *Homo economicus* (53) | Algorithm*, anticipated, approximation*, axiom*, centralization*, complementarity, computational, consensus*, convergence*, counterfactual, difficult*, efficien*, equilibr*, expect*, experience*, feedback, first-best, first-order, forecast*, free-riding, habit, idiosyncratic*, informational, intertemporal, leader*, logic, maxim*, minim*, model*, motivated, noncooperative, normative, optimal*, optimiz*, optimum, perceived, perfect, plausible, predict*, preference*, profitability, random*, rational*, reason*, satisfi*, search, selfish, simulat*, stationary, steady-state, tractable, tradeoff, utilit* |

It is hard to give a specific account for how each of these words was selected. By and large, we roughly taxonomized the words into three groups: the first group of words facilitated the expression of the idea, *Homo economicus*; the second group of words facilitated the expression of the idea, *Homo sapiens*; and the third group of words was neutral to both paradigms. Two remarks need to be made here, however. First, needless to say, this taxonomy makes a bold assumption for language, since neither meaning nor function is context free. The justification that we make is at best a first-order approximation, and bringing the context and embeddedness together is necessary if our initial analysis can indicate that this is an interesting direction for future research. Second, even though we can assume such a taxonomy, any manual classification of the words into the above three categories might still suffer from a certain degree of arbitrariness. In other words, different "experts" can come up with different taxonomies. We certainly cannot exclude such possibilities, but we can assume that such differences are mostly of a secondary or minor degree and hence will have little effect on the results.

With the above two qualifications in mind, we identified the first and the second group of words. Of the 101 words we picked, 48 words were related to the *Homo sapiens* paradigm (the first group) and 53 words were related to the *Homo economicus* paradigm (the second group). A quick look at Table 5.3 shows that many of the words in the former relate to psychology, sociology, or other social sciences. For example, "cognit*" and "emotion*" were chosen in light of Thaler (2000): "Economists will study human *cognition*," and "*Homo economicus* will become more *emotional*" (Ibid, p. 137 and p. 139, respectively).

Note that in Table 5.3, for some words, we used the stemmed word and placed a * symbol after the stemmed word. The * symbol enables WordSmith to search for any derivatives of the stemmed word. If we take "cognit*" as an example, WordSmith will count the occurrences of "cognitive," "cognition," "cognitively," "cognitivity," etc. By taking "anthropolog*" as another example, WordSmith searches not only for "anthropology," but also for "anthropological," "anthropologist," etc.

We used the WordSmith *Concord* tool to analyze the frequencies of these 101 words year by year from 1992 to 2014. *Concord* indicates all references for each single word in our corpus and can show the *concordance lines*, which enables us to look closely at how each word is used in a sentence. It can also show the frequencies of a word in any given text (Scott 2001). Each of our documents is indexed by year, and the frequencies of a word across all documents can be arranged to appear in the form of a time series.

After a set of keywords was determined, the application of the corpus linguistics to the study of the paradigm shift in economics could be formulated into a *trending hypothesis*. The trending hypothesis indicated that the frequencies of words that are relevant to the paradigm of *Homo economicus* have a tendency to decline over time, whereas the frequencies of words relevant to the paradigm of *Homo sapiens* tend to increase over time.

### 5.2.3 A Linear Regression Model to Analyze Word Frequencies

A standard way to see whether a word has a trend is to run a simple linear regression and to regress the word frequency (the regressor) against time (the regressand) (Bianchi et al. 1999). The trend can then be determined by the slope of the regression line. A positive slope indicates an upward trend, while a negative slope indicates a downward trend. At the same time, the R-squared value is also calculated for each word, which means that the proportion of variation explained by the model, or simply, how good the model fits the real data, can also be known. We used Python codes to construct the regression model.

In this study, we also checked the *p*-values for the trend. A rule of thumb applied here is that the trend is statistically significant if the *p*-value is lower than 0.05. Words without a significant trend were removed from the modified word list, and our subsequent analysis only focused on words with a statistically significant trend. For those words which had a *p*-value lower than 0.05, we plotted the estimated trend (their slope) on the *x*-axis and the corresponding R-squared value on the *y*-axis (see Figs. 5.4 and 5.5 below).

### 5.2.4 Co-word Network Analysis of the Economics Literature Before 2000 and After 2000

We are not only interested in the trend of the individual words, but also in how the relationships among words changed. We carried out a co-word network analysis based on the abstracts published before 2000 and the abstracts published after 2000, respectively, and compared how the two networks differed. The co-word network analysis can be used to discover the relationship between words by exploring which words co-occur with other words (He 1999). We used the tool referred to as *ConText* to implement the co-word network analysis (Diesner 2014), and examined the relationships among the 101 words mentioned above.[3]

## 5.3 Results

In this section, we first use the stemmed word "cognit*" as an illustration to show how each stemmed word is analyzed in this chapter (Sect. 5.3.1). Then, in Sect. 5.3.2, we show the overall frequency trend for both the *Homo sapiens* words and *Homo economicus* words. Finally, in Sect. 5.3.3, we offer a microscopic

[3] Since only WordSmith can recognize stemmed words, a.k.a., the "*" symbol, and *ConText* cannot, we removed the * symbol and changed the stemmed words to normal words when using the tool *ConText*.

interpretation that explains why some words exhibit an upward trend and some words a downward trend. We propose four reasons that could explain the upward trend observed for some *Homo sapiens* words.

### 5.3.1 Concordance for the Stemmed Word "Cognit*"

A typical result is demonstrated below using the stemmed word "cognit*." Figure 5.1 shows the first 25 of 456 sentences that include the derivatives of the word "cognit*." This figure provides us with a close examination of the context in which this keyword of interest is situated. For example, the second entry involves Frederick's (2005) discussion of the relationship between cognitive reflection and decision-making.

WordSmith can show how many hits a word has in each year and the normalized frequency of that word in every 1000 words. Figure 5.2 visually shows that the derivatives of "cognit*" have only a few occurrences (see the column "Plot") before the year 2000, but more occurrences in later years.

Figure 5.3 is a visual representation of the normalized frequency of "cognit*." The trendy property of this stemmed word is represented by a fitted regression line as shown in Fig. 5.3. Despite a degree of fluctuation, there is clearly an upward trend in terms of its use. The accompanying regression line shows that this linear model

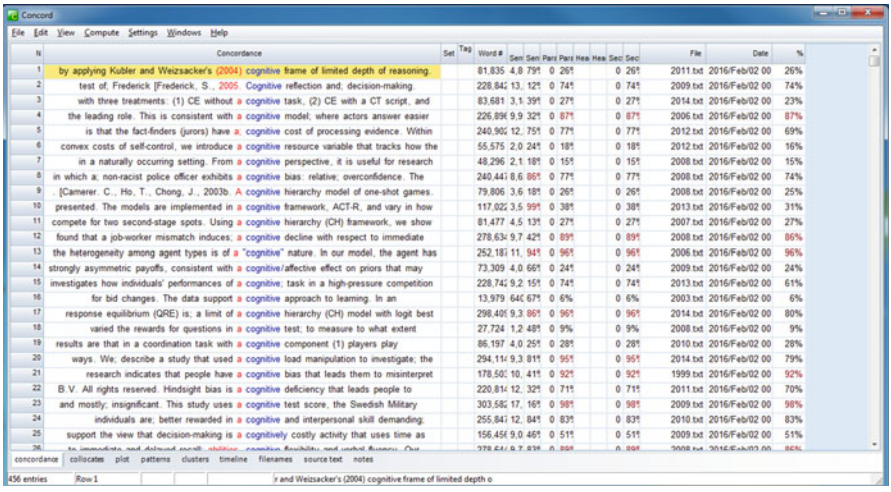



**Fig. 5.1** The concordance lines for the stemmed word "cognit*." This page was generated by the WordSmith *Concord* tool. The first column is the index of the concordance line. The second column is the concordance line, which shows the sentence that includes the derivatives of the stemmed word "cognit*." In the bottom left corner of the figure, it is stated that there were 456 entries, meaning that in our corpus, the derivatives of the stemmed word "cognit*" appeared 456 times

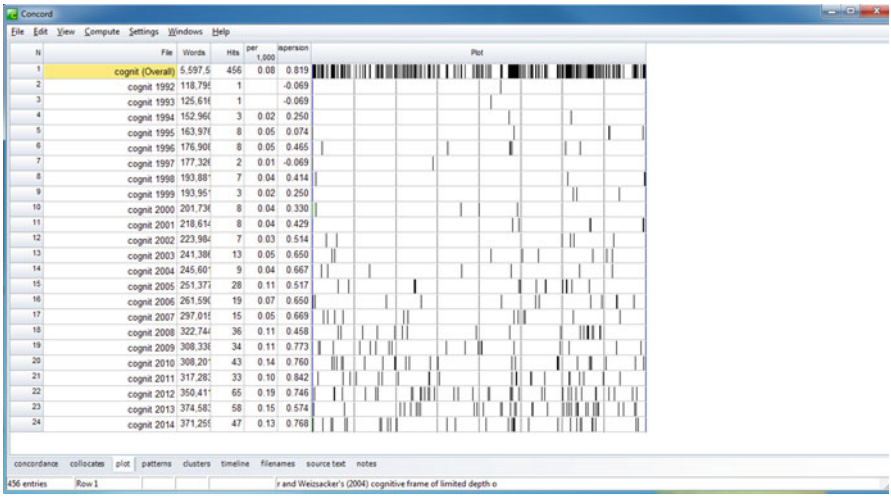



**Fig. 5.2** The concordance plot for the derivatives of "cognit*." This figure was also generated by the WordSmith *Concord* tool. The "File" column denotes the files corresponding to each year. For example, "Cognit 1992" refers to the file which includes all abstracts published in 1992. The "Words" column shows how many words there are in the respective file. The "Hits" column shows how many times the derivatives of "cognit*" appear in a file. The "Per 1000" column shows the normalized frequency of the derivatives by averaging their frequencies of appearance for every 1000 words. The "Plot" column shows the position where the derivatives appear in a file by reconfiguring the text as a strip

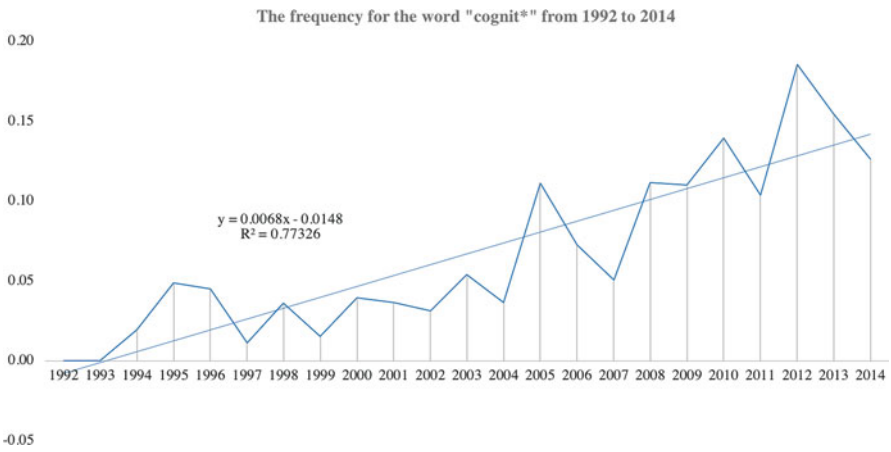



**Fig. 5.3** The normalized frequency of the derivatives of "cognit*" increases with time (the curved line) and the fitted regression (the straight line). The *x*-axis denotes the year, and the *y*-axis denotes the normalized frequency

is a good fit because the fitted line has an *R*-squared value of 0.773. The slope of the linear regression line, 0.0068, is positive, and has a *p*-value of 3.30453E-08, much smaller than 0.05, which indicates that the trend is both statistically significant and positive.

### *5.3.2 A Macroscopic Examination of the Trendy Keywords*

The linear regression analysis as illustrated for the stemmed word "cognit*" was applied to all the other 100 keywords. For brevity, in the rest of this chapter we use "word" to indicate not only a word but also a stemmed word. Out of a total of 101 words, 12 were not found to be trendy (the p-value of the estimated slope being higher than 0.05). They are therefore not considered in our further analysis. The remaining 86 words are demonstrated in an *x*–*y* plot in Figs. 5.4 and 5.5, where the *x*-axis denotes the trending coefficient and the *y*-axis denotes the associated R-squared value.

To make sense of these results, we first considered an extreme, but ideal, pattern of the trending hypothesis. Specifically, we expected to see that those words associated with the *Homo sapiens* paradigm had a positive slope, while those words associated with the *Homo economicus* paradigm had a negative slope. However, this is a rather ideal situation, as the methodological restrictions or the violations of the simplified assumptions that occurred previously could all cause some degree of deviation. Nevertheless, we were still able to ask whether the majority of words did in fact behave in line with our expectations.

Table 5.4 shows that, for the *Homo sapiens* words, this was roughly the case: 81.4% (the majority) of the words (35 out of 43) had an upward trend, and only 18.6% (the minority) of the words (8 out of 43) had a downward trend. For the *Homo economicus* words, we see a similar, but a less pronounced, pattern: 65.2% (a mild majority) of the words (30 out of 46) had a downward trend, while 34.8% (a mild minority) of the words (16 out of 46) had an upward trend.

Hence, from a macroscopic viewpoint, the paradigm of *Homo sapiens* did gain momentum over time, and that gain, to some extent, can be translated into the gradual decline of the paradigm of *Homo economicus*. From this observation, we can conclude that Thaler's prediction is largely correct. In fact, considering that our period of observation is not sufficiently long, one may expect this predicted shift to be still ongoing.

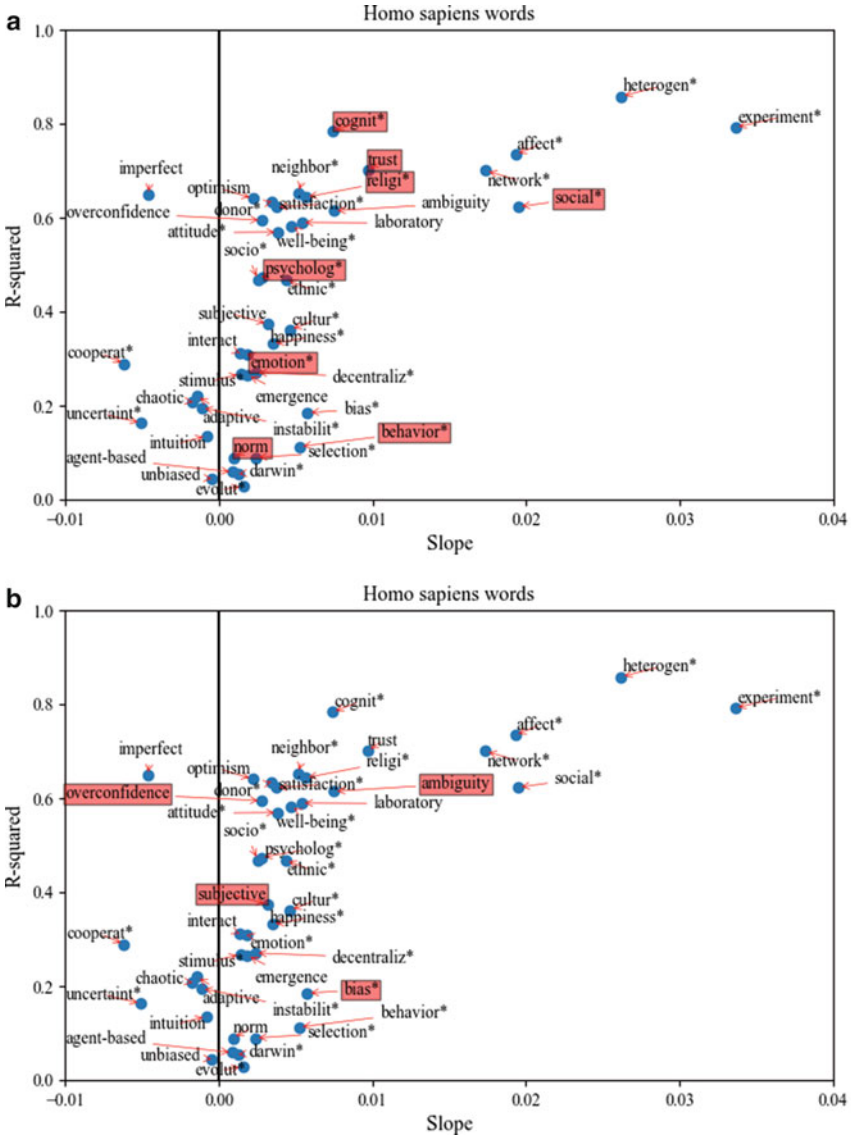



**Fig. 5.4** (**a**) The trending coefficient (*x*-axis) and the R-squared value (*y*-axis) of the first group of words that relate to the paradigm of *Homo sapiens*. (**b**) The trending coefficient (*x*-axis) and the R-squared value (*y*-axis) of the second group of words that relate to the paradigm of *Homo sapiens*. (**c**) The trending coefficient (*x*-axis) and the R-squared value (*y*-axis) of the third group of words that relate to the paradigm of *Homo sapiens*. (**d**) The trending coefficient (*x*-axis) and the R-squared value (*y*-axis) of the fourth group of words that relate to the paradigm of *Homo sapiens*

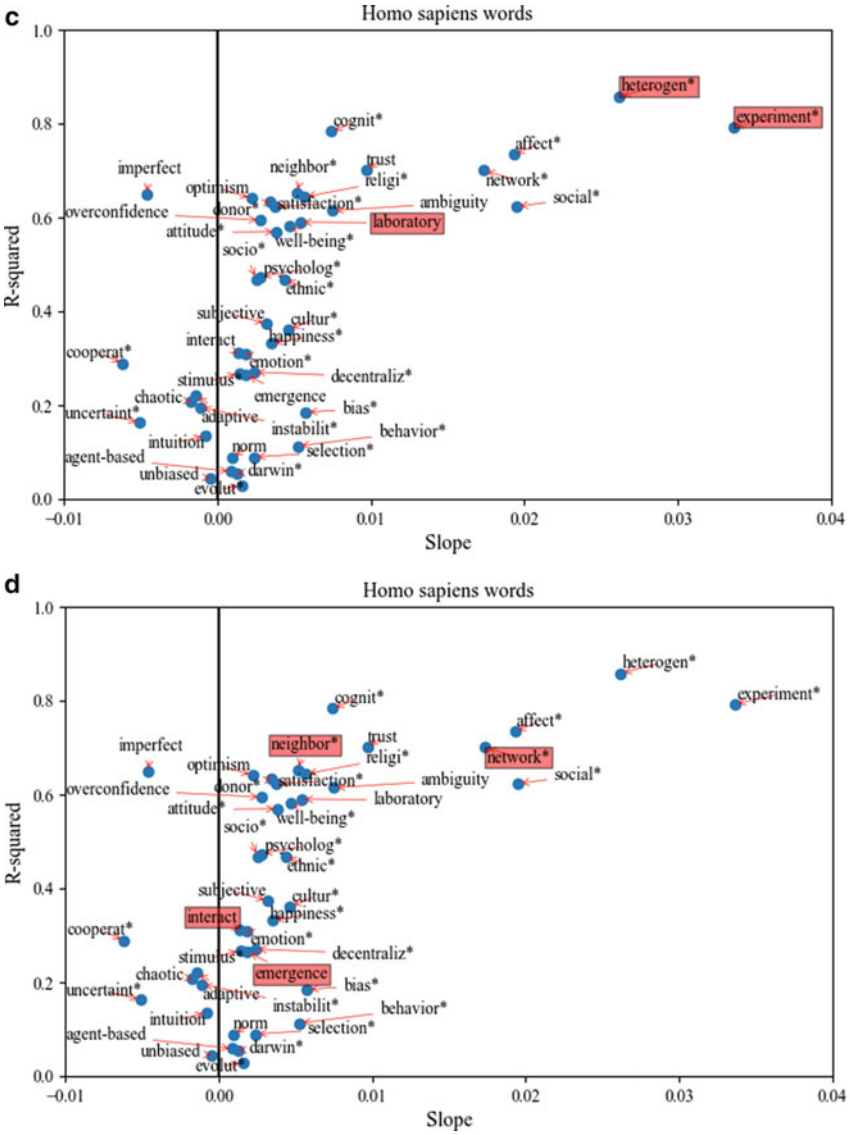



**Fig. 5.4** (continued)

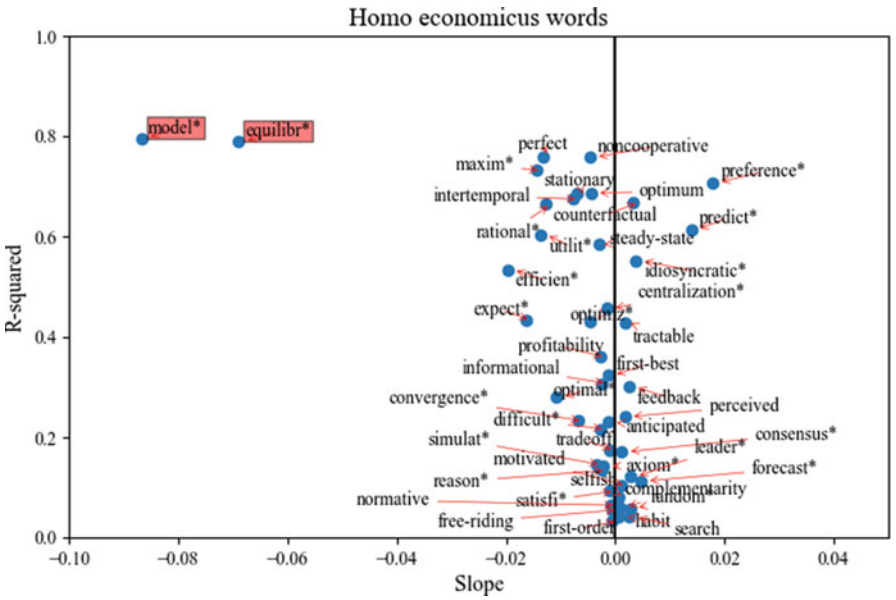



**Fig. 5.5** The trending coefficient and R-squared value of words related to the *Homo economicus* paradigm

**Table 5.4** The words with a positive slope and a negative slope

| | *Homo sapiens* | *Homo economicus* |
|---|---|---|
| Total number of words | 48 | 53 |
| Words with a $p$ value <0.05 | 43 | 46 |
| Words with a positive slope | 35 | 16 |
| Words with a negative slope | 8 | 30 |

### 5.3.3 *A Microscopic Examination of the Trending of Words in the Entire Corpus*

In addition to the macroscopic perspective, it is interesting to provide a microscopic examination of words. We first, in Sect. 5.3.3.1, discuss the words related to the *Homo sapiens* paradigm, and then, in Sect. 5.3.3.2, the words related to the *Homo economicus* paradigm. We are not only interested in words that fit our hypothesis, but are also interested in the words that deviate from our expectations.

#### 5.3.3.1 *Homo sapiens* Words

In this section, we first examine the 35 *Homo sapiens* words that have a positive slope. It will, however, be hard and even fragmental to elaborate on them individu-

ally. Some words were conceptually connected, so we grouped them and elaborated on a group of words. We identified four groups of words that had a positive slope. We then addressed how each of these groups can help *Homo sapiens* get established.

The first group of words, highlighted in Fig. 5.4a, is related to other sister disciplines in the social sciences, for example, "psychology*," "cognit*," "emotion*," "social*," "trust," "norm," "religi*," "behavior*," etc. All these words, when put together, indicate that the idea of *Homo sapiens* is the consequence of long-standing interdisciplinary integration, by which economists have accepted ideas from psychology, sociology, anthropology, ethnics, and cultural studies.

Thaler (2000) predicted that "*Homo sapiens* will begin losing IQ" and "will be a slow learner" (Ibid, p. 134, 135, respectively). The second group of words, highlighted in Fig. 5.4b, including words such as "bias*," "ambiguity," "overconfidence," and "subjective", are words that represent the (cognitive) constraints of humans, in terms of IQ, which may in turn influence their decision-making. "Overconfidence," for example, has a steep slope, which basically reflects increasing general concerns that economists have for their economic man: economic miscalculation is not an exception, but a rule.

The third group of words, highlighted in Fig. 5.4c, is related to the heterogeneity of humans, as exemplified in the following words: "heterogen*," "laboratory," and "experiment*." The paradigm of *Homo sapiens* assumes that agents are non-trivially different or that they are heterogeneous. To harness their behavior, one cannot just count on the analytical models, but should use the "laboratory" approach by running "experiments." The third group of words supports this claim.

The fourth group of words, highlighted in Fig. 5.4d, are "network*," "neighbor*," "interact," and "emergence." These words are all concepts related to the complexity of economic behaviors. Complexity science studies complex systems, in which many parts interact with each other and conceptually form a network. During the past three decades, complexity economics has emerged as a field that treats economic agents as constantly interacting with each other and changing their behavioral rules or strategies along the course of these interactions (Arthur 2014). The macroscopic patterns of humans or a complex system are not just a linear summing up of individual components, but often have properties not seen at the individual level (the emergent properties).

After going through the four groups of well-expected or justifiable trendy words for the paradigm of *Homo sapiens*, it is also interesting to see those *Homo sapiens* words that actually exhibit an opposite trend. There are seven such stemmed words. Several of them, including "adaptive," "chaotic," "imperfect," and "uncertain*," are words that were already used by heterodox economists well before the 2000s, when such economists attempted to challenge the stringent assumptions of neoclassical economics. For example, they used adaptive behavior to question or challenge rational behavior, used chaotic dynamics to challenge the trivial and uninteresting linear dynamics, and used the environment characterized by true uncertainty and the resultant imperfect information to address the difficulty of rational calculation and the implausibility of an expected utility maximization framework. What, then,

caused the importance of these words to decline over time? We believe that this is an open question that can only be answered by further research.

#### 5.3.3.2 *Homo economicus* Words

In Sect. 5.3.2, we have seen the increasing frequencies for the majority of the *Homo sapiens* words, and the decreasing frequencies for the (mild) majority of the *Homo economicus* words. After analyzing the positive trend of the former, in this subsection, we examine the negative trend of the latter. There are 30 *Homo economicus*-related words with a negative slope. In Fig. 5.5, we have highlighted the two that noticeably stand out, namely "model*" and "equilibr*." These two, as compared to many others, have both a steep negative slope (a sharp declining rate) and a larger *R*-squared value.

The sharp decline in the normalized frequency of the word "equilibrium" is interesting but not surprising. The concept of equilibrium is a centerpiece of the paradigm of *Homo economicus*, and it goes hand in hand with "rationality." *Homo economicus* is assumed to be rational as if it is mathematically "optimizing" a well-specified objective function. Not only does this well-articulated behavioral formulation help characterize what an equilibrium is, but it also helps provide a solution to the model, which is normally characterized as the "steady state(s)" of the model. Once the steady state is determined, the remaining issue left for the dynamic analysis will be the path "converging" to the steady state.

*Homo sapiens,* on the other hand, is assumed to be boundedly rational. The mathematical description of the behavior of *Homo sapiens* is far from just optimizing and, worse, is normally not homogeneous and unique. Putting all of these together makes the model-solving a daunting task if not impossible. Under these circumstances, the equilibrium is no longer operational for the model. Neither is the "steady state," nor is "convergence". In Fig. 5.5, we see that in addition to "equilibrium," "rational," "optimal," "steady-state," and "convergence" all exhibit a declining normalized frequency, although to a much milder degree.

The sharp decline of the keyword "model" is intriguing and probably more mythical. One possible conjecture is that under the paradigm of *Homo sapiens*, the pure analytical model may tend to be less useful or relevant given the explanation above; hence, it drives economists to find other ways of handling the uncertainty of a theoretical world, for example, by means of simulation, laboratory experiments, field experiments, naturally occurring experiments, or even a model-free data-driven approach. This by no means implies the extinction of the models; as a matter of fact, in some disciplines of economics, models are still very much alive, but despite this being so more space needs to be left for other accompanying approaches.[4]

[4]In a sense, this simply means that it is becoming increasingly difficult for the pure theoretical model to be accepted and published.

### 5.3.4 *Co-word Network Analysis*

The co-word network based on abstracts both before 2000 and after 2000 is shown in Figs. 5.6 and 5.7, respectively. The size of a word is scaled according to betweenness centrality, which is a network parameter that indicates how central a node is in a network (Borgatti 1995). We can observe that before 2000, the first tier of central words includes "model" and "equilibrium," followed by words such as "social," "uncertain," "behavior," and "optimal."

We observe a change in network structure after 2000. The co-word network shows that the centrality of "equilibrium" and "model" declined substantially. The first tier of words includes "model," "behavior," and "social," followed by "optimal," "equilibrium," and "rational." It is another way to show that the *Homo economicus* paradigm is gradually changing to the *Homo sapiens* paradigm.

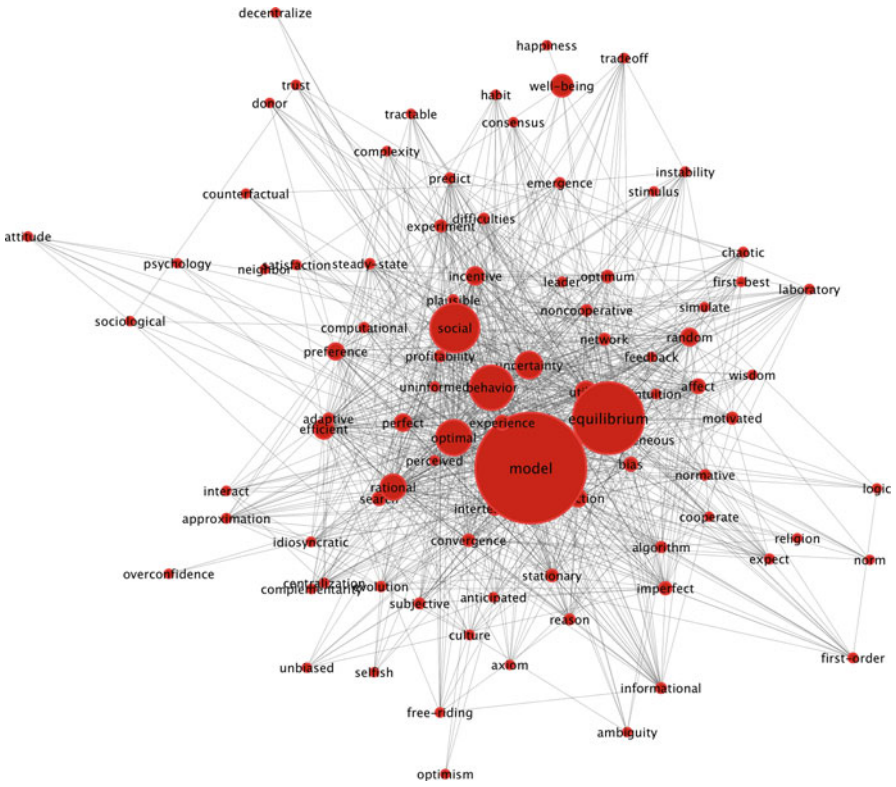



**Fig. 5.6** Co-word network of economic corpus: 1992 to 1999

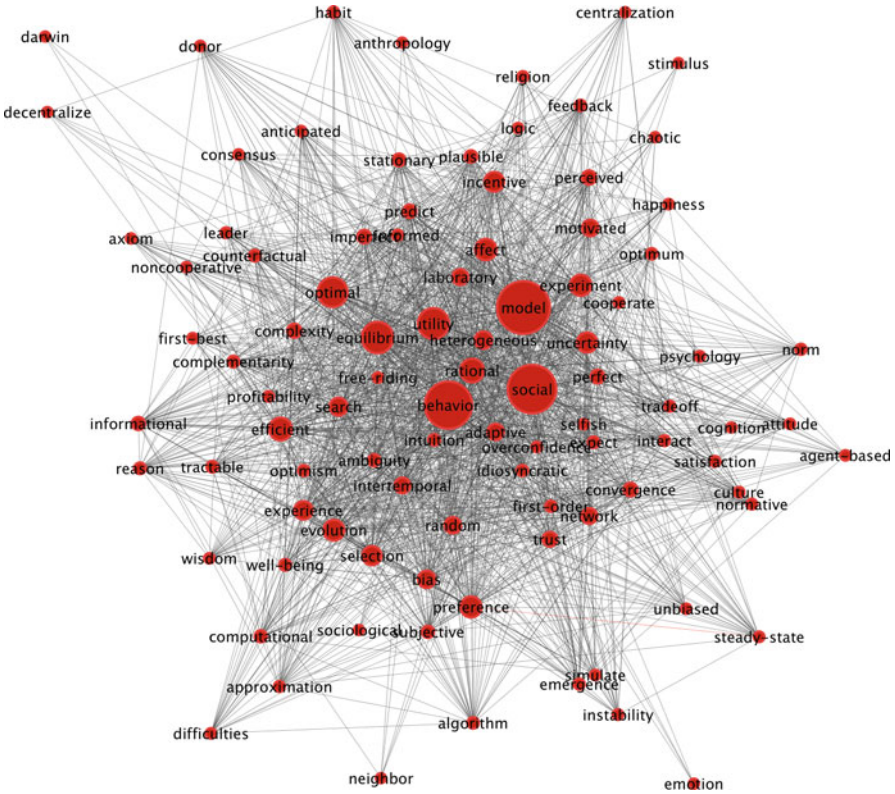



**Fig. 5.7** Co-word network of economics corpus: 2001 to 2014

## 5.4 Concluding Remarks

In this chapter, we have applied the corpus linguistic approach to address an issue related to the paradigm shift. The historian and philosopher of science Kuhn (1962) invented the term "paradigm shift" to describe the change in the set of core concepts or the set of operating routines belonging to a scientific discipline. This term is not restricted to the sciences and was introduced to the social sciences when a similar change was experienced. While a paradigm shift can be clearly recognized after the end of the shift, recognizing it at the initial or the middle stage has not been that straightforward. For quite some time, the document-based approach has already been applied to study the process of the shift, again usually much later than when the shift settled. Nowadays, with the advances in ICT and various automated information extraction algorithms, machines can be assigned to this laborious work. Using this modern technology, we can detect or predict a shift through the constant online monitoring of the texts, posts, and human conversations in social media networks. In this way, we can know the time, the scale, and the milestones of a

paradigm shift. Hence, we can study the paradigm shift not as a subject of history, but as its contemporaneity. Our study of the paradigm shift from *Homo economicus* to *Homo sapiens*, which is still ongoing, provides a concrete illustration of this promising research methodology.

From an economics viewpoint, we can consider that keywords (symbols) are both competing and cooperating with each other to gain limited human attention. In this regard, the paradigm shift is equivalent to the shift in human attention. The keywords are, on the one hand, competing as if they want to draw a certain share of human attention. On the other hand, to achieve the above purpose, cooperating with other keywords and forming a co-word network is also important. With this conceptual framework, in this chapter, we have presented both the normalized frequency (market share) and the co-word network of keywords. The former allows us to distinguish momentum-gaining keywords from momentum-losing keywords, and the latter allows us to know the "major players" or "hubs" of a "syndicate." These two analyses together essentially provide technical characterizations of a "paradigm," which not only governs the use of our attention resource, but also dictates the organization of the attention modularized by the keywords. Hence, under this framework, a paradigm shift not only means that we pay attention to different things, but we also consider how these different kinds of attention are organized together. In this chapter, we have been able to find both characterizations, hence suggesting that a paradigm shift has occurred from *Homo economicus* to *Homo sapiens*.

There is a fundamental limitation which we do not intend to leave unnoticed. As we have mentioned before, our approach based on machines should at best read as an "assistant's job," which does not intend to replace the role that a historian can play. In fact, the foundation of this study is identifying keywords and their classifications. A machine can help us perform the first task very powerfully, but for the second one there is simply no theoretical justification on which we can rely. To some extent, this latter task is still very subjectively performed in this part, and can be problematic. Therefore, we should keep the question, regarding the soundness of the two sets of keywords established in this study, open.

Despite this possible limitation, what is found in this paper is generally insightful. On the one hand, we see the declining tendency of keywords such as "rationality," "equilibrium," and "optimal". Can this evidence alone be a sign for the decay of *Homo economicus*? On the other hand, we also see the increasing tendency of keywords such as "psychological," "emotion," "cognitive," "culture," "social," and "heterogeneity". Can this evidence alone confirm our feeling that economics has become increasingly pluralistic and has no longer carried her crown of "economic imperialism"? We do not have a definite answer, but our evidence prompts us to throw the questions out, and we believe that when digital social sciences or humanities becomes more advanced we may one day have the answer too, of course, under the joint efforts with humans.

**Acknowledgements** The second author is grateful for the research support in the form of *Ministry of Science and Technology (MOST) Grants*, MOST 106-2410-H-004-006-MY2.

## References


Arthur, W. B. (2014). *Complexity and the economy*. Oxford: Oxford University Press.

Bianchi, M., Boyle, M., & Hollingsworth, D. (1999). A comparison of methods for trend estimation. *Applied Economics Letters, 6*(2), 103–109.

Biber, D., Conrad, S., & Reppen, R. (1998). *Corpus linguistics: Investigating language structure and use*. Cambridge: Cambridge University Press.

Blaug, M. (2001). No history of ideas, please, we're economists. *Journal of Economic Perspectives, 15*(1), 145–164.

Borgatti, S. P. (1995). Centrality and AIDS. *Connections, 18*(1), 112–114.

De Bellis, N. (2009). *Bibliometrics and citation analysis: From the science citation index to cybermetrics*. Metuchen, NJ: Scarecrow Press.

Diesner, J. (2014). ConText: Software for the integrated analysis of text data and network data. In *Social and semantic networks in communication research*. Seattle, WA: Conference of the International Communication Association (ICA).

Dohmen, T. (2014). Behavioral labor economics: Advances and future directions. *Labour Economics, 30*, 71–85.

Flowerdew, L. (2015). Using corpus-based research and online academic corpora to inform writing of the discussion section of a thesis. *Journal of English for Academic Purposes, 20*, 58–68.

Frederick, S. (2005). Cognitive reflection and decision making. *The Journal of Economic Perspectives, 19*(4), 25–42.

Frey, B. S. (1997). *Not just for the money*. London: Edward Elgar Publisher.

He, Q. (1999). Knowledge discovery through co-word analysis. *Library Trends, 48*(1), 133.

Kalaitzidakis, P., Mamuneas, T. P., & Stengos, T. (2011). An updated ranking of academic journals in economics. *Canadian Journal of Economics/Revue canadienne d'économique, 44*(4), 1525–1538.

Knight, D., Walsh, S., & Papagiannidis, S. (2015). I'm having a Spring Clear Out: A corpus-based analysis of e-transactional discourse. *Applied Linguistics, 38*(2), 234–257. https://doi.org/10.1093/applin/amv019.

Kuhn, T. S. (1962). *The structure of scientific revolutions*. Chicago: University of Chicago Press.

Lovejoy, A. O. (2011). *The great chain of being: A study of the history of an idea*. Piscataway, NJ: Transaction Publishers.

McDonald, I. M. (2009). The global financial crisis and behavioural economics. *Economic Papers: A Journal of Applied Economics and Policy, 28*(3), 249–254.

Mill, J. S. (1874). On the definition of political economy, and on the method of investigation proper to it. In *Essays on some unsettled questions of political economy* (2nd ed.). London: Longmans, Green, Reader & Dyer.

Nowak, M., & Highfield, R. (2011). *Super cooperators: Altruism, evolution, and why we need each other to succeed*. New York: Simon and Schuster.

Pumfrey, S., Rayson, P., & Mariani, J. (2012). Experiments in 17th century English: Manual versus automatic conceptual history. *Literary and Linguistic Computing, 27*(4), 395–408.

Schreibman, S., Siemens, R., & Unsworth, J. (Eds.). (2008). *A companion to digital humanities*. New York: Wiley.

Scott, M. (2001). Comparing corpora and identifying key words, collocations, and frequency distributions through the WordSmith tools suite of computer programs. *Small Corpus Studies and ELT*, 47–67.

Simon, H. A. (1962). The architecture of complexity. *Proceedings of the American Philosophical Society, 106*(6), 467–482.

Thaler, R. H. (2000). From homo economicus to *Homo sapiens*. *Journal of Economic Perspectives, 14*(1), 133–141.

Zipf, G. K. (1949). *Human behavior and the principle of least effort*. Boston: Addison-Wesley.

# Chapter 6
# Big Data and FinTech


**Jia-Lang Seng, Yao-Min Chiang, Pang-Ru Chang, Feng-Shang Wu, Yung-Shen Yen, and Tzu-Chieh Tsai**

## 6.1 Introduction

With the prevalence of digital convergence, mobile communication, Big Data analytics, cloud computing service, and artificial intelligence, the digital financial industry is experiencing a revolutionary trend. In essence, the widespread popularity of smartphones, the intelligent computing, and the ubiquity of mobile cloud services have revolutionized the face of business worldwide, especiallyin the financial

J.-L. Seng (✉)
Department of Accounting, National Chengchi University, Taipei, Taiwan
e-mail: seng@nccu.edu.tw

Y.-M. Chiang
Department of Finance, National Taiwan University, Taipei, Taiwan
e-mail: yaominchiang@ntu.edu.tw

P.-R. Chang
Department of Risk Management and Insurance, Shih-Chien University, Taipei, Taiwan
e-mail: brchang@g2.usc.edu.tw

F.-S. Wu
Graduate Institute of Technology, Innovation and Intellectual Property Management, National Chengchi University, Taipei, Taiwan
e-mail: fswu@nccu.edu.tw

Y.-S. Yen
Department of Computer Science and Information Management, Providence University, Taichung, Taiwan
e-mail: ysyen@pu.edu.tw

T.-C. Tsai
Department of Computer Science, National Chengchi University, Taipei, Taiwan
e-mail: ttsai@cs.nccu.edu.tw



S.-H. Chen (ed.), *Big Data in Computational Social Science and Humanities*, Computational Social Sciences, https://doi.org/10.1007/978-3-319-95465-3_6

services industry, also known as financial technologies (FinTechs). The market changes to platform and infrastructure mean that data analytics related or other real-time computing requirements affect the value chain in FinTechs which in turn determines the market share in the world. The financial industry must be encountering the global changes by actively seeking for destructive innovation ideas and escalating their customer experience and service quality. With the sponsorship of Taiwan Ministry of Science and Technology's 3-year multidisciplinary research project, "Innovative and Mobile Financial Technologies," this chapter aims to document the research findings and outcome of the research project where we try to integrate the financial intelligence of structured and unstructured data, create a cloud-based mobile computing platform, and build a two-way decision support prototype system. This integrated approach adopts the academic studies and advanced technical development. The main research results documented include the news media database, rule base, asset pricing model, multi-case study series, empirical research model, data analytics algorithm development, sentiment analysis and opinion mining application, and mobile cloud computing platform (Deloitte 2016, pp. 1–5; KPMG 2015a, pp. 10–13; WEF 2015, pp. 1–3; WEF 2017).

With the emergence of service innovations and innovation integration in clusters, the focus has shifted to the development of new financial services with data science and analytics, mobile technologies, and cloud computing that diversify and customize the new functionalities available to business clients and customers. Most worldwide financial holdings corporations have faced major challenges resulting from dynamic business environments, diverse telecommunication situations, and heterogeneous data sources that must be marshaled in an integrated, real-time, and seamless manner. Today, businesses interact with heterogeneous business models, business processes, and workflows. The results have been upheaval, chaos, and disruption. Therefore, it is crucial to develop comprehensive, cross-disciplinary, and in-depth resolution of services, models, and technologies to address the above issues. Facing the massive amount of dynamically changing and heterogeneous data sources, financial advisors, managers, investors, and other stakeholders need real-time, accurate, comprehensive, high-frequency, and interactive decision support systems to make investment decisions as services. The goal is to increase the wealth of clients and to improve the corporations' competitive edge. In this chapter, we present four main research results in the following sections. They are (1) intelligent investment models in the capital market, (2) text analytics and sentiment analyses of financial news, (3) innovative financial service strategies, and (4) mobile cloud technologies to implement these proposed solutions.

In Sect. 6.2, we incorporate news sentiment into an asset pricing model to examine various views of stock price reactions to investment, macroeconomic, and political news in 2013–2014. The results show that there is positive relationship between investment news sentiment and TW50 return that is there is the statistical significance. However, regarding the effect of macroeconomic news on stock prices, these results are very different from the results of previous articles in the literature. Political news sentiment has statistically significant negative effects on certain financial stocks listed on the TW50 index.

In Sect. 6.3, we focus on the financial news sentiment analysis conducted by using software. We analyze the content of the news, develop various customized word lists, build a dictionary, and refine scoring rules that evaluate news items in the contexts of various research topics. The result is statistically significant but the result may be affected by the subjective judgment of readers. Our dictionary captures positive words and negative words, but it has limited grasp of what the words in news articles convey.

In Sect. 6.4, we investigate how securities firms and information and communication technology (ICT) development firms respond to the threats and opportunities of coming era of FinTechs. From the results, we thought that securities firms may need to rethink the way they divine customers' needs and interact with customers and building their own in-house innovative capabilities in order to be capable of quickly responding to customers' needs and to social and economic changes, and the ICT development firms may need to reposition their products and services for the financial services industry and may need to update their ICT technologies. In this section, we also find that the relationship between securities firms and ICT development firms may be changing as the securities firms appear to intend to develop mobile systems and service innovations by themselves.

In the final section, we propose a framework that can handle complex data structures and large amounts of data in a short time using cloud computing. The result shows the response times of different task assignment algorithms for different intensities of burst traffic.

## 6.2 Text Mining News for Stock Price Predictions

### *6.2.1 Introduction*

The efficient market and rational investor hypotheses in the literature imply that the price of a security reflects the information available to investors concerning the value of that security. Security pricing is a topic that is still debated among academics and practitioners, and the efficient market hypothesis has been re-examined by numerous researchers (Tetlock 2007, pp. 1139–1168; Fama and French 2015, pp. 1–22). In the traditional asset pricing model, researchers use a single factor or multiple factors to predict actual asset price, such as market premium, economic factors, firm size, and book-to-market ratio (Fama and French 1992, pp. 427–465; Fama and French 2015, pp. 1–22). However, Baker and Wurgler (2006, pp. 1645–1680), Brown and Cliff (2005, p.405–440), and Kumar and Lee (2006, pp. 2451–2486) show that other factors such as investor sentiment can predict stock returns in the cross-section. Therefore, in this study, we add news sentiment as another factor to construct a new security pricing model to examine the effect of news sentiment on stocks.

Some recent studies suggest that media coverage may have statistically significant relationships to asset prices even when it does not involve hard, breaking

news (Tetlock 2007, pp. 1139–1168; Kearney and Liu 2014, pp. 171–185; Fang and Peress 2009, pp. 2023–2052) and report that firms with less media coverage have higher required rates of return. Firms that experience an exogenous reduction in analyst coverage have higher required rates of return as well as less efficient pricing and lower liquidity. Another study uses the textual sentiment to price actual security assets to show that textual sentiment can produce statistically significant abnormal returns (Loughran and McDonald 2011, pp. 35–65). Therefore, we include news sentiment into an asset pricing model. We argue that using media coverage as a proxy for investor sentiment in an asset pricing model may improve the performance of that asset pricing model.

## *6.2.2 Literature Review*

### 6.2.2.1 Asset Pricing

Asset pricing was first explored by Sharpe (1964, pp. 425–442), who uses a Markowitz model to estimate average returns and risks of market portfolios; this is called a capital asset pricing model (CAPM). This single-factor model suggests that the security risk expected return depends on beta and the risk premium of the market portfolio. Apart from the market premium, multifactor models add other factors such as business-cycle risk, monetary policy, and the ratio of economic growth to the price security, as arbitrage pricing models (APMs) do. In 1993, Fama and French design a three-factor model that considered market premium, firm size, and book-to-market equity ratio to price securities. In 2015, to improve their model's performance regarding price-risky assets, Fama and French add another two factors to their three-factor model, namely operating profitability and investment.

### 6.2.2.2 News Sentiment

Recent studies find that media coverage and investors sentiment are substantially related to asset prices. Models of investor sentiment predict that low sentiment will generate downward price pressure, and unusually high or low values of sentiment will generate high volume. Antweiler and Frank (2004, pp. 1259–1294) study messages in Internet chat rooms focused on stocks, characterizing the content of the messages as "buy," "sell," or "hold" recommendations to investigate the relationship between the message and return. Tetlock (2007, pp. 1139–1168) suggests that measures of media content serve as a proxy for investor sentiment or noninformational trading; his empirical results show that media volume (media coverage) has positive relationships with idiosyncratic volatility, past-year return, and past-year absolute return but a negative relationship with the book-to-market ratio. Kurov (2010, pp. 139–149) also finds that monetary policy shocks strongly influence investor sentiment in bear market periods. Garcia (2013, pp. 1267–1300) uses the

proportions of positive and negative words to examine the effect of sentiment on asset prices and finds that daily news content helps predict stock returns, particularly during recessions. Kearney and Liu (2014, pp. 171–185) indicate that researchers use positive and negative verbal information based on textual sentiment, which may lead to notable differences in risk-adjusted stock returns for complementary econometric modeling of financial market effects. They further claim that textual sentiment or the tone of qualitative information has noteworthy effects on stock prices and returns and that negative sentiment has the strongest influence.

### *6.2.3 Methodology*

We construct a multifactor model to examine the effects of news on stock prices. News events are gauged by collecting 59,223 investment articles, 20,090 macroeconomic articles, and 49,848 politics articles from January 1, 2013, to December 31, 2014, and stock market returns from January 2, 2013, to December 31, 2014, as independent variables. We adopt the stock returns of TSEC-FTSE Taiwan 50 Index (TW50) listed companies as dependent variables. The model can be expressed as follows:

$$r_i = \beta_0 + \beta_1 r_m + \beta_2 \Delta s_{\text{inv}} + \beta_3 \Delta s_{\text{mac}} + \beta_4 \Delta s_{\text{poli}} + \varepsilon_i$$

where $r_i$ is the return on the price of stock $i$, $r_m$ is the stock market return, $\Delta s_{\text{inv}}$ is the difference of investment news score, $\Delta s_{\text{mac}}$ the difference of macroeconomic news score, and $\Delta s_{\text{poli}}$ is the difference of political news score.

### *6.2.4 Empirical Results*

Table 6.1 provides an overview of three group's news sentiment scores and the stock market index. The market index changed from 7616 to 9569 in 494 business days. In Table 6.2, the TW50-listed stock returns have strong, statistically significant positive correlations with stock market returns. The positive relation of investment news sentiment to TW50 return is statistically significant. However, regarding the effect of macroeconomic news on stock prices, these results are very different from the results of previous articles in the literature. Political news sentiment has statistically significant negative effects on certain financial stocks listed on the TW50 index, namely stocks 2883 (China Development Financial), 2884 (E.SUN Financial Holding Company), and 2891 (CTBC Financial Holding Co., Ltd.).

**Table 6.1** Summary statistics of news observations

| | Market index | Investment news | Politics news | Macroeconomic news |
|---|---|---|---|---|
| *N* | 494 | 59,223 | 20,090 | 49,848 |
| Mean | 8544.21 | 0.38 | 0.34 | 0.04 |
| Median | 8466.38 | 0.00 | 0.00 | 0.00 |
| SD | 523.13 | 2.04 | 4.92 | 1.08 |
| Min | 7616.64 | −38.00 | −56.00 | −28.00 |
| Max | 9569.17 | 36.00 | 54.00 | 36.00 |

We select news from three categories—investment, politics, and macroeconomic news, and stock market index—from January 2, 2013, to December 31, 2014.

### *6.2.5 Conclusion and Discussion*

This research examines various views of stock price reactions to investment, macroeconomic, and political news in 2013–2014. The TW50-listed stock prices are highly related to the stock market index and to the investment news, but only the stock prices in the banking industry relate to political news. In future work, we can classify various stocks according to the industries used by the Taiwan Stock Exchange Corporation (TWSE) and calculate the cross-sectional price return for each day.

This table shows the TW50 stock alphas, coefficients, *t*-values, and adj-$R^2$ values from January 2, 2013, to December 31, 2014, in a time series regression. The model is

$$r_i = \beta_0 + \beta_1 r_m + \beta_2 \Delta s_{\text{inv}} + \beta_3 \Delta s_{\text{mac}} + \beta_4 \Delta s_{\text{poli}} + \varepsilon_i$$

where the right-hand variable is stock price return and the left-hand variables are four-factor differences of investment news, macroeconomic news, political news, and stock market returns.

## 6.3 Financial News Sentiment Analysis and Application

### *6.3.1 Introduction*

This section explores and consolidates valuable information from analysis of financial news based on sentiment analysis and opinion mining. We analyze sentiments in news content and apply this sentiment analysis. We extract subjective vocabulary that reflects emotion or attitude. Moreover, we develop a dictionary and propose a method for calculating a score from specific emotionally subjective vocabulary used in the news.

We focus on using our content analysis techniques to extract information from a news database and construct an automatic news-scoring application based on our

**Table 6.2** TW50 stocks 4-factor regression

| Stock id | Intercept | $\Delta s_{inv}$ | $t$-value | $\Delta s_{mac}$ | $t$-value | $\Delta s_{poli}$ | $t$-value | $r_m$ | $t$-value | Adj-$R^2$ |
|---|---|---|---|---|---|---|---|---|---|---|
| 1101 | 0.053 | 0.632*** | (3.99) | −0.003 | (−0.07) | 0.209 | (0.80) | 0.618*** | (10.73) | 0.19 |
| 1102 | 0.039 | 0.276** | (2.51) | 0.013 | (0.50) | 0.091 | (0.51) | 0.387*** | (9.68) | 0.15 |
| 1216 | 0.030 | 0.450** | (2.65) | −0.025 | (−0.61) | −0.100 | (−0.36) | 0.541*** | (8.77) | 0.13 |
| 1301 | 0.005 | 0.234* | (1.80) | 0.036 | (1.14) | −0.193 | (−0.90) | 0.541*** | (11.46) | 0.21 |
| 1303 | 0.050 | 0.283* | (1.65) | 0.017 | (0.41) | 0.585** | (2.08) | 0.632*** | (10.17) | 0.17 |
| 1326 | 0.000 | 0.519** | (3.63) | 0.040 | (1.17) | 0.307 | (1.31) | 0.623*** | (11.99) | 0.22 |
| 1402 | 0.018 | 0.262** | (2.09) | 0.018 | (0.61) | −0.315 | (−1.53) | 0.435*** | (9.56) | 0.16 |
| 1476 | 0.290 | 0.581 | (1.53) | −0.094 | (−1.04) | −0.788 | (−1.27) | 0.494*** | (3.59) | 0.03 |
| 2002 | 0.010 | 0.374** | (3.79) | 0.038 | (1.60) | 0.162 | (1.00) | 0.436*** | (12.16) | 0.23 |
| 2105 | 0.045 | 0.512** | (3.19) | 0.055 | (1.42) | 0.065 | (0.25) | 0.434*** | (7.45) | 0.10 |
| 2207 | 0.181 | 0.773** | (3.05) | −0.003 | (−0.05) | 0.121 | (0.29) | 0.983*** | (10.66) | 0.18 |
| 2227 | 0.112 | 0.256 | (1.08) | −0.009 | (−0.16) | −0.092 | (−0.24) | 0.466*** | (5.40) | 0.05 |
| 2301 | 0.018 | 0.598** | (3.15) | 0.028 | (0.61) | 0.545* | (1.75) | 0.482*** | (6.99) | 0.09 |
| 2303 | 0.074 | 0.508** | (2.52) | 0.110* | (2.27) | −0.294 | (−0.89) | 0.768*** | (10.48) | 0.18 |
| 2308 | 0.144 | 0.477** | (2.15) | −0.033 | (−0.62) | −0.194 | (−0.53) | 0.585*** | (7.26) | 0.09 |
| 2311 | 0.107 | 0.327 | (1.61) | 0.042 | (0.86) | −0.349 | (−1.05) | 0.618*** | (8.38) | 0.12 |
| 2317 | 0.056 | 0.452** | (3.27) | 0.017 | (0.52) | 0.141 | (0.62) | 0.588*** | (11.69) | 0.21 |
| 2325 | 0.120 | 0.409** | (2.06) | 0.051 | (1.08) | −0.022 | (−0.07) | 0.614*** | (8.50) | 0.12 |
| 2330 | 0.093 | 0.731*** | (4.70) | 0.072* | (1.92) | −0.323 | (−1.27) | 0.845*** | (14.98) | 0.32 |
| 2354 | 0.021 | 0.418** | (2.69) | 0.033 | (0.88) | −0.349 | (−1.37) | 0.506*** | (8.98) | 0.14 |
| 2357 | 0.057 | 0.246 | (1.14) | 0.043 | (0.83) | −0.347 | (−0.98) | 0.525*** | (6.70) | 0.08 |
| 2382 | 0.069 | 0.760** | (3.82) | 0.075 | (1.57) | 0.349 | (1.07) | 0.601*** | (8.30) | 0.12 |
| 2395 | 0.185 | 0.332 | (1.29) | 0.121* | (1.95) | 0.116 | (0.28) | 0.694*** | (7.45) | 0.10 |
| 2408 | 0.301 | 0.218 | (0.46) | −0.019 | (−0.17) | −0.527 | (−0.68) | 0.650*** | (3.80) | 0.02 |
| 2409 | 0.061 | 0.392 | (1.47) | 0.041 | (0.63) | −0.074 | (−0.17) | 0.759*** | (7.84) | 0.11 |
| 2412 | 0.023 | 0.105 | (1.49) | −0.006 | (−0.36) | 0.039 | (0.33) | 0.186*** | (7.23) | 0.09 |
| 2454 | 0.098 | 0.463** | (2.20) | 0.066 | (1.30) | −0.220 | (−0.64) | 0.607*** | (7.96) | 0.11 |

(continued)

**Table 6.2** (continued)

| Stock id | Intercept | $\Delta s_{inv}$ | $t$-value | $\Delta s_{mac}$ | $t$-value | $\Delta s_{poli}$ | $t$-value | $r_m$ | $t$-value | Adj-$R^2$ |
|---|---|---|---|---|---|---|---|---|---|---|
| 2474 | 0.141 | 0.404 | (1.55) | −0.044 | (−0.71) | −0.655 | (−1.54) | 0.769*** | (8.15) | 0.12 |
| 2801 | 0.056 | 0.420** | (3.84) | 0.033 | (1.25) | 0.110 | (0.61) | 0.473*** | (11.89) | 0.22 |
| 2880 | 0.043 | 0.217** | (2.23) | 0.008 | (0.33) | −0.186 | (−1.17) | 0.454*** | (12.90) | 0.25 |
| 2881 | 0.097 | 0.548** | (3.63) | 0.003 | (0.07) | −0.274 | (−1.10) | 0.676*** | (12.33) | 0.24 |
| 2882 | 0.124 | 0.349** | (2.29) | 0.021 | (0.58) | −0.186 | (−0.74) | 0.698*** | (12.61) | 0.24 |
| 2883 | 0.079 | 0.481** | (3.44) | 0.033 | (0.99) | −0.524** | (−2.28) | 0.533*** | (10.49) | 0.19 |
| 2884 | 0.091 | 0.250** | (1.90) | 0.003 | (0.11) | −0.430** | (−1.99) | 0.433*** | (9.07) | 0.15 |
| 2885 | 0.036 | 0.602*** | (4.08) | 0.050 | (1.42) | 0.230 | (0.95) | 0.715*** | (13.34) | 0.26 |
| 2886 | 0.040 | 0.281** | (2.32) | 0.009 | (0.32) | −0.248 | (−1.25) | 0.492*** | (11.20) | 0.20 |
| 2887 | 0.077 | 0.374** | (3.06) | 0.028 | (0.94) | −0.059 | (−0.29) | 0.486*** | (10.94) | 0.19 |
| 2890 | 0.056 | 0.376** | (3.14) | 0.034 | (1.19) | −0.437** | (−2.22) | 0.499*** | (11.47) | 0.22 |
| 2891 | 0.078 | 0.579*** | (4.61) | 0.022 | (0.73) | −0.065 | (−0.31) | 0.578*** | (12.67) | 0.24 |
| 2892 | 0.048 | 0.203** | (2.32) | −0.003 | (−0.15) | −0.161 | (−1.13) | 0.445*** | (14.06) | 0.29 |
| 2912 | 0.114 | 0.343* | (1.91) | 0.070 | (1.62) | −0.129 | (−0.44) | 0.440*** | (6.74) | 0.08 |
| 3008 | 0.254 | 0.723** | (2.30) | −0.054 | (−0.72) | −0.050 | (−0.10) | 0.835*** | (7.33) | 0.10 |
| 3045 | 0.030 | 0.527** | (3.36) | 0.035 | (0.92) | 0.330 | (1.28) | 0.397*** | (6.96) | 0.09 |
| 3474 | 0.581 | 0.674 | (1.52) | −0.049 | (−0.46) | −0.743 | (−1.02) | 0.496*** | (3.08) | 0.02 |
| 3481 | 0.018 | 0.707** | (2.46) | 0.103 | (1.49) | −0.627 | (−1.33) | 0.810*** | (7.77) | 0.11 |
| 4904 | 0.027 | 0.156 | (0.98) | 0.014 | (0.35) | 0.510* | (1.95) | 0.295*** | (5.11) | 0.05 |
| 4938 | 0.172 | 0.238 | (0.94) | 0.064 | (1.06) | 0.113 | (0.27) | 0.486*** | (5.31) | 0.04 |
| 5880 | 0.033 | 0.169** | (2.35) | 0.003 | (0.19) | −0.062 | (−0.53) | 0.305*** | (11.68) | 0.21 |
| 6505 | −0.034 | 0.300** | (2.04) | −0.070* | (−1.98) | −0.001 | (0.00) | 0.589*** | (11.01) | 0.20 |
| 9904 | 0.077 | 0.748** | (3.41) | 0.040 | (0.77) | 0.543 | (1.51) | 0.624*** | (7.82) | 0.11 |

***, **, and * are respectively the statistical significance levels at 1%, 5%, and 10%

scoring rules. In addition, on the basis of an analysis of commonly used words in the news, our study develops various customized word lists of terms oriented toward macroeconomics, politics, and investment.

### *6.3.2 Literature Review*

Data analytics is a process in which computer programs apply mathematical and statistical methods to extract new and previously unknown information from textual data. Unlike numerical and financial data, textual data contains not only the effect, but also the possible causes of the event. The ability to exploit textual information successfully could increase the quality of the input data and improve the understanding of issues.

Content analysis[1] is a wide and heterogeneous set of manual or computer-assisted techniques for contextualized interpretations of documents produced by communication processes (in the strict sense of that phrase) or by signification processes, of which the ultimate goal is the production of valid and trustworthy inferences. Tetlock (2007, pp. 1139–1168) examines investor sentiment by measuring the pessimism index from the GI dictionary. Tetlock et al. (2008, pp. 1437–1467) find that negative words in firm-specific news stories predict low firm earnings. Past studies are also commonly known as document-level sentiment classification because the whole document is considered as a basic information unit.

Hu et al. (2012, pp. 674–684) indicate that research on sentiment analysis uses an automatically generated sentiment lexicon, in which a list of seed words is used to determine whether a sentence contains positive or negative connotations. Then, the polarity (positive or negative) of an opinion is determined on the basis of the words in the document. Xu et al. (2011, pp. 743–754) classify technologies of sentiment analysis into two categories: unsupervised approaches and supervised approaches. The unsupervised approaches usually create a sentiment lexicon and determine a document's polarity by counting that document's positive and negative phrases. The supervised approaches use labeled data to train certain classifiers to predict unlabeled data. The semantic orientation approach (Zhang et al. 2013, pp. 851–860) performs classification on the basis of positive or negative sentiment words (or phrases) contained in each evaluated item (this operates on several levels: document, sentence, or attribute). The lexicon is crucial to the semantic orientation approach. However, the speed at which vocabulary items are collected is less than the speed at which vocabulary is generated daily on the Internet. In addition, constructing different vocabularies for different domains is indeed challenging because of the polysemous nature of the terms. Therefore, the task of vocabulary construction is difficult because of the different areas of domain knowledge. Four challenges are encountered when creating domain-specific texts (Ittoo and Bouma 2013, pp. 2530–2540): silence, the absence of knowledge resources, complex terms, and noise (informal or ungrammatical language). Domain-specific texts are sparse and do not provide sufficient statistical evidence to facilitate the detection of terms. This phenomenon, whereby infrequent but vital terms are rejected or missed, known

[1] https://en.wikipedia.org/wiki/Content_analysis.

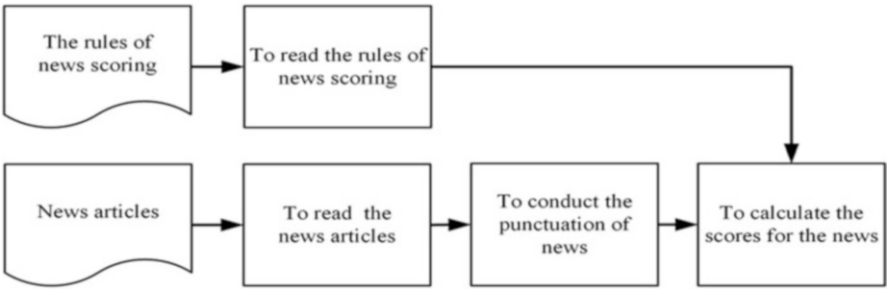



**Fig. 6.1** News scoring process

as silence, affects the recall of term extraction. Most extraction techniques fail to identify long phrases or expressions. Furthermore, ambiguity arises when a domain-specific corpus is expressed in a terse language. Another common type of linguistic incoherence is the omission of punctuation symbols, such as periods (".") to indicate the end of sentences. As mentioned previously, the lexicon is pivotal and must include comprehensive, domain-specific, and multiword forms, enabling expanded research on financial news (Fig. 6.1).

## *6.3.3 Methodology*

### 6.3.3.1 Data Collection

Our news data source contains a total of 129,161 news items from a news database.[2] For classifying each news item on the basis of its content, we choose some news seed words that relate to the macroeconomics, politics, and investment. For each news item in the corpus of news from the news database,[3] we calculate the optimal group for the news item using category seed words and a vector space model (Manning et al. 2008, pp. 2–4). Each news article is placed into the category with the least distance between the seed words and the news text. Concurrently, we calculate the relevant scores for each news article based on similarity with news category seed words. The higher an article's scores, the more relevant that article is to the category (Fig. 6.2).

### 6.3.3.2 Research Design

We use content analysis techniques to extract information from the news data. Initially, we select 1535 news articles from the news database.[4] We separate those news articles into three subsets as follows: 1092 news articles related to

[2]Knowledge Management Winner, http://kmw.chinatimes.com/.

[3]Knowledge Management Winner, http://kmw.chinatimes.com/.

[4]Knowledge Management Winner, http://kmw.chinatimes.com/.

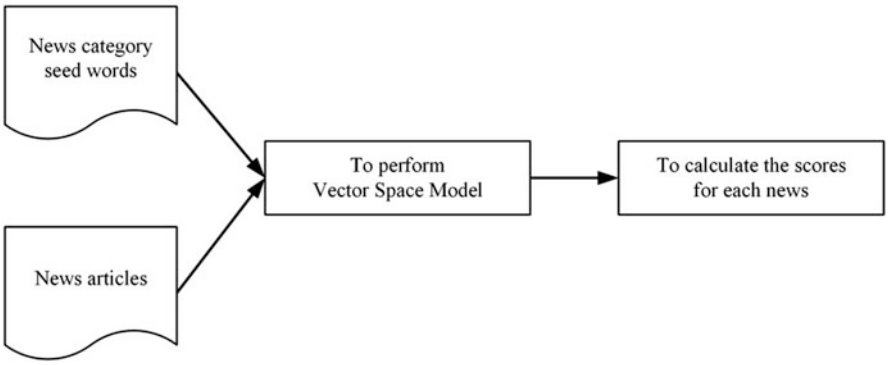



**Fig. 6.2** Calculating relevant scores for each news item

macroeconomics, 129 news articles related to politics, and 314 news articles related to investment. Furthermore, we select the most relevant 130 news articles from these three types of news; we apply a computer program to read the news artificially, and use our dictionary to count the positive words and negative words objectively. The system develops various customized word lists and creates scoring rules by dividing the scale into ten deciles: five negative and five positives. The score interval for a news article is configured between the most negative decile and the most positive decile. We filter news through the three basic news categories at first and then analyze the features of each news category to build the sentiment lexicon.

Finally, the system applies the rules to compute news scores automatically for each news article. The application automatically reads both news and scoring rules from the data source. Thus, through content analysis, we analyze the content of sentences according to the scoring rules and score each news article.

### *6.3.4 Research Results*

By analyzing the news data through sentiment analysis, we build the scoring rules and construct an automatic news-scoring application. Following (Liu 2012, pp. 1–167), we use a numeric score expressing the strength (i.e., intensity). The maximum score is $+5$ and the minimum score is $-5$; news content indicating a preferable situation receives a positive score, whereas news content indicating an undesirable situation receives a negative score.

For example, if the news refers to positive GDP growth, the macroeconomic news score is $+2$; if the news mentions a plunge of stock prices, the macroeconomic news score is $-5$. If a political news article refers to the Sunflower Student Movement of Taiwan, the political score for that article is $-4$; by contrast, an article that mentions economic integration is assigned a score of $+1$. A news item that mentions the reappearance of the Golden Cross is scored at $+5$ on the investment scale, whereas

an article about a plunge of stocks at the close of a trading session receives a −5 investment news score. The tone of a news item is the affect or emotional feeling that the news item communicates regarding macroeconomic news, political news, and investment news.

The score interval is divided into ten segments, five of which are below zero and five above; the scoring of each news article depends on the amounts of positive and negative words.

### *6.3.5 Conclusion and Discussion*

This section focuses on the financial news sentiment analysis conducted by using software. We analyze the content of the news, develop various customized word lists, build a dictionary, and refine scoring rules that evaluate news items in the contexts of various research topics. These rules are applied by an automatic news-scoring application. Social media influences human behavior and causes fluctuations in the market. The price changes in financial instruments (such as stocks, bonds, or mutual funds) are consequences of actions taken by investors, reflecting their perceptions of events commented on in social media. By use of our lexicon dataset and our scoring rules, we can investigate whether the emotional messages in the social media context correlate with the price changes of financial instruments, and make it possible to forecast stock price fluctuations in advance. The empirical result is statistically significant but the result may be affected by the subjective judgment of readers. Our dictionary captures positive words and negative words, but it has limited grasp of what the words in news articles convey. Consequently, future research can expand the quantity of data that is considered and the range of years from which news is sampled to gain additional insights.

## 6.4 Development of Mobile Banking Service Innovation and Its Effects on Securities Firms

### *6.4.1 Introduction*

Banks increasingly struggle to enhance their mobile banking capabilities as they confront growing pressure from customers, emerging technological innovations, and growing competition from new market entrants (KPMG 2015a, pp. 1). Recently, new types of financial technologies (abbreviated as FinTechs) have been receiving avid attention from all over the word (Chuen and Teo 2015, pp. 24–37). The concept of FinTechs emphasizes firms' usage of new technologies to provide new financial services in a more effective and efficient way. Banks and other financial services companies attach great importance to momentous innovations from FinTechs and

attempt to comprehend threats and find responses to some fundamental questions (Gulamhuseinwala et al. 2015, pp. 16–23).

The study mainly focuses on two problems. First, we investigate the effects of FinTechs on both securities firms and information and communication technology (ICT) development firms in Taiwan. Second, we address how these securities firms and ICT development firms endeavor to innovate and change their business models to adapt to the new age of FinTechs.

## *6.4.2 Literature Review*

### 6.4.2.1 Global Development Trend of Mobile Banking

In the East, giants in the internet industry, such as Alibaba and Tencent, are rising to become providers of banking services with branchless banks such as Ant Financial and WeBank. These communication technologies not only enhance the financial services sector but also provide wider access to banking and financial services (Chuen and Teo 2015, pp. 24–37).

KPMG (2015b, pp. 1–10) indicates that in the early twenty-first century, smartphones with WAP support enable the use of the mobile web. Various mobile devices and tablets can access bank websites and services. After 2010, mobile banking apps capitalize on the burgeoning success of the iPhone and the speedy growth of Android smartphones. Bank customers are directed to mobile-based websites or apps. After initially offering basic portfolio of banking through mobile, mobile banking has evolved from basic service to include a broad, rich set of capabilities.

### 6.4.2.2 FinTech Development and Strategy from a Global Perspective

FinTech firms combine innovative business models and technologies to enable, enhance, and disrupt financial services (Gulamhuseinwala et al. 2015, pp. 16–23). FinTechs also involve innovative financial services or products delivered through new technology. Consumer expectations alter with advances in technology (particularly advances in mobile and Internet technologies). Customers expect benefits from widespread technologies that have been globally adopted (Chuen and Teo 2015, pp. 24–37). FinTechs must meet customers' expectations efficiently; customers must find FinTechs easy to use.

Even though FinTech development is in an early stage, successful FinTech development is obviously difficult in the current situation of extreme competition. Financial service firms face challenges from competitors from both inside and outside of the securities industry (Chuen and Teo 2015, pp. 24–37). Regarding new entrants to the FinTech market, these start-ups appear to have begun to "unbundle" banking services and carve out business in some of the established banks' most profitable business lines (KPMG 2015c, pp. 1–33). The Internet Finance Guidelines

indicate that China is creating both a financial market infrastructure and a regulatory framework that is specific to FinTechs. In fact, FinTechs-related services are booming in China with numerous peer-to-peer lending providers (Arner and Barberis 2015, pp. 78–91). From the global perspective, most young banking customers appear to be willing to use new mobile banking services; the average age of mobile banking users is in the mid-1930s (KPMG 2015a, pp. 1–8; KPMG 2015b, pp. 1–10).

Some traditional companies are beginning either to cooperate with partners or to implement outright acquisitions to respond to FinTechs. The money transfers and payment services provided by FinTechs are also an essential part of the customer journey for numerous popular e-commerce sites, which are designed to eliminate conflicts and enhance conversion rates at the purchase stage (Gulamhuseinwala et al. 2015, pp. 16–23).

Taiwan's 2016 Financial Supervisory Commission (FSC) "FinTechs Development Strategy White Paper" (FSC 2016, pp. 109) recommends the Taiwanese securities industry to pursue goals such as raising online order rates to 70%, improving automated trading mechanisms (such as robo-advisors and online fund sales platforms), strengthening cloud services for securities and futures, and deepening Big Data applications.

## *6.4.3 Methodology*

### 6.4.3.1 Research Framework

The study investigates how FinTechs affect both securities firms and ICT development firms and how these firms change in response to the maturation of FinTechs. Thus, a conceptual framework for this study is proposed as follows (Fig. 6.3).

### 6.4.3.2 Research Approach and Research Subjects

The development of FinTechs is still in its infancy. Additionally, very few studies in this area have been completed. Thus, the use of a case study approach for obtaining a clear understanding is appropriate (Yin 1994).

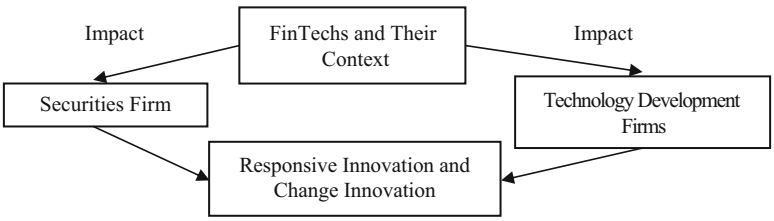



**Fig. 6.3** Conceptual framework for the study

We eventually chose one top securities firm, JihSun Securities, and two ICT development firms, Mitake Corporation and SysJust Corporation, as the major companies to be investigated. The latter two firms occupy more than 95% of the market for special ICT and financial service software in Taiwan. We select nine managers from these three firms for our in-depth interviews.

## *6.4.4 Research Results*

### 6.4.4.1 Effects of FinTech

The interviewee from JihSun mentions that FinTechs have numerous effects on the JihSun company. For example, they may face competition from new entrants that might be considered outsiders, such as Alibaba. Additionally, a securities firm faces more pressure from customers who are equipped with advanced handheld devices and who request prompt feedback regarding securities. Furthermore, as the consumers and marketing actions become more individualized, the firm needs to find an acceptable way to identify particular needs. Finally, the interviewee pointed out that government policies are also crucial influencing factors; financial services firms ask for help from government because FinTechs intensify global competition. The interviewees from the ICT development firms indicated their originally dominated markets appear to be affected because financial services companies plan to build internal innovative capabilities to deal with the threats and opportunities from FinTechs. Additionally, they must update their relevant technological capabilities, such as online transaction platforms and Big Data, which are changing much faster than before. The major effects of FinTechs on both securities firms and ICT development firms are listed in Table 6.3.

**Table 6.3** Effects of FinTech

| | Impacts |
|---|---|
| FinTechs and its context | *Securities firm*<br>• More competitions from new entrants and global competitors<br>• Deregulation in areas of online account, online payment, online order, personal information protection, prediction risk notice, forbiddance in investment suggestions, etc.<br>• Provision of new business development fund and consulting services from government to support firms to develop new services<br>• Increased customers' bargaining power<br>*ICT development firms*<br>• Gradually losing customers from and markets of financial service sector<br>• Facing new and broader technologies, such as online transaction platform, APP technology, robo-advisor, Big Data, and cloud service |

#### 6.4.4.2 Responsive Innovations and Changes

Facing the challenges from FinTech, both financial holdings and ICT companies are innovating technologies, services, and products supporting their digital strategy. The interviewee from the securities firm mentioned that his company is putting more resources into the new areas of robo-advisor Big Data, apps, and third-part payments. He further indicates that they must analyze customers' behaviors more quickly and accurately by using these new technologies to provide superior services. To reach the aforementioned goals, the company has tried hard to establish internal innovative capabilities. For instance, they hire some new employees with particular expertise in new areas. The securities firm also puts great effort into training its employees. They attempt to cooperate with several universities either to conceive new service ideas or to develop new technologies in an effective and efficient way. However, the ICT development companies seem to realize the changes experienced by their customers. They must upgrade their technological capabilities as FinTechs generate some new technologies, which are quite different from the traditional "information" technologies. Furthermore, they must reposition their products and markets as both technologies and customers undergo tremendous changes. The detailed results from the case studies are listed in Table 6.4.

**Table 6.4** Responsive innovations and changes

| Securities firm | ICT development firms |
|---|---|
| *New product and service innovation*<br>**New products**<br>• Currency trading, funds<br>**New services**<br>• Online orders, TSM service (for credit card), VIP service fee discount, or investment services seminar, openness of information about investment targets; e.g., managed funds-push notifications, personalized advice recommend, third-part payment<br>*Technology and equipment resources*<br>• Investment on software, such as APP, Big Data analysis, and robo-advisor system<br>• Investment on ICT infrastructure<br>*Human resource*<br>• Hiring and training the human talents with skills in APP technology<br>*External cooperation for innovation*<br>• Cooperating with universities | *New product and service innovation*<br>**New products**<br>• Security mobile order platform, database, cloud testing centers, third-part payment system<br>**Service-mobile security**<br>• Message order, e.g., LINE, voice order, mobile banking order, web banking order; APP order, e.g., Mitake Estock, cloud service, analysis of big data; e.g., customer behavior, investment, customized service<br>*Technology and equipment resources*<br>• Investment on ICT infrastructure, IT resource for APP<br>*Human resource*<br>• Hiring and training human talents in APP technology and Big Data.<br>*External cooperation for innovation*<br>• Cooperate with securities firms to design the online platform |

### 6.4.5 Conclusion and Discussion

This study investigates how securities firms and ICT development firms respond to the threats and opportunities of coming era of FinTechs. From the results of preliminary qualitative case studies, several conclusions can be drawn: (1) Securities firms may need to rethink the way they divine customers' needs and interact with customers because those consumers are equipped with much more advanced and fast-changing handheld devices than earlier consumers had. (2) Securities firms may need to consider building their own in-house innovative capabilities in order to be capable of quickly responding to customers' needs and to social and economic changes. (3) The relationship between securities firms and ICT development firms may be changing as the securities firms appear to intend to develop mobile systems and service innovations by themselves. (4) The ICT development firms may need to reposition their products and services for the financial services industry and may need to update their ICT technologies.

## 6.5 Real-Time Financial Service Framework

### 6.5.1 Introduction: Creative Financial Services

The advances of cloud computing and mobile devices have changed human lifestyles in numerous respects. We propose a framework that can handle complex data structures and large amounts of data in a short time using cloud computing. To provide ubiquitous mobile financial services, we develop an adaptive task assignment algorithm on our framework to overcome the instability of wireless mobile networks so that the financial system can deliver essential messages to a large number of customers instantly.

### 6.5.2 Methodology and Research Results

#### 6.5.2.1 Cloud Computing and Big Data

Cloud Computing and Big Data Platform Comparison

There are numerous Big Data platforms based on the MapReduce model (Dean and Ghemawat 2008, pp. 107–113), such as Hadoop, Disco, and Spark. Although Spark is a heavyweight framework, Spark is a practical choice to manipulate Big Data because it is known for in-memory computing. In this chapter, we propose a real-time framework for solving stock market prediction algorithms, as shown in Fig. 6.4.

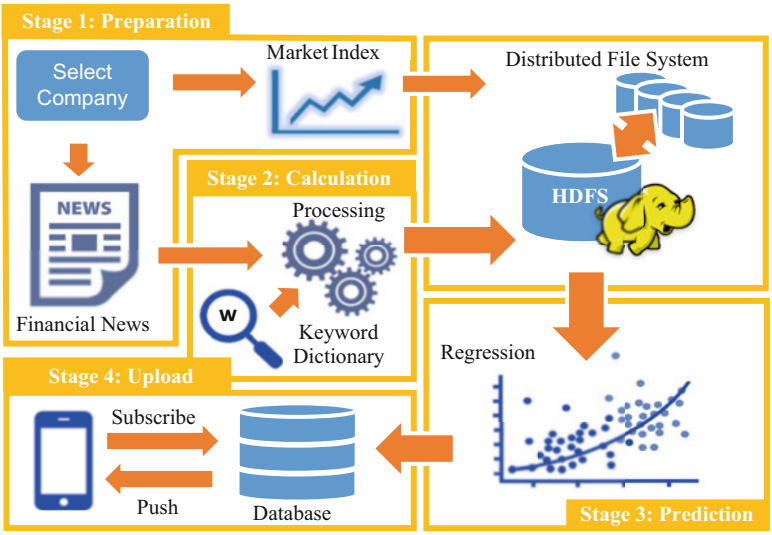



**Fig. 6.4** Four stages of stock market prediction algorithms

**Stage 1: Preparation** To make a precise prediction, we collect data from the Internet through the web crawler. The data includes news, volumes, stock quotes, price–book ratios (PBRs), price–earnings ratios (PERs), and market indexes.

**Stage 2: Calculation** The key point of the model is news-processing algorithms. First, we filter the news by topics appropriate to finance. Second, we segment compound Chinese sentences into Chinese words. Third, we search through the keywords in a news file. Keywords are divided into two categories: positive words and negative words. According to the keywords in the content, we score each news item.

**Stage 3: Prediction** A regression analysis is used for our stock market prediction model. In our case, we try to discover the relationship between news and stock trends.

**Stage 4: Upload** The prediction results for stock trends are saved in our database. We build a back-end web service to send this information to users instantly.

## Simulation Result of Platforms Comparison

In the simulation, we provide results using a Python language model for normal Python, Disco, and Spark. We focus on the calculation step, because the other steps only take few seconds to complete processing. In our MapReduce strategy, each news item is processed line by line and in one job.

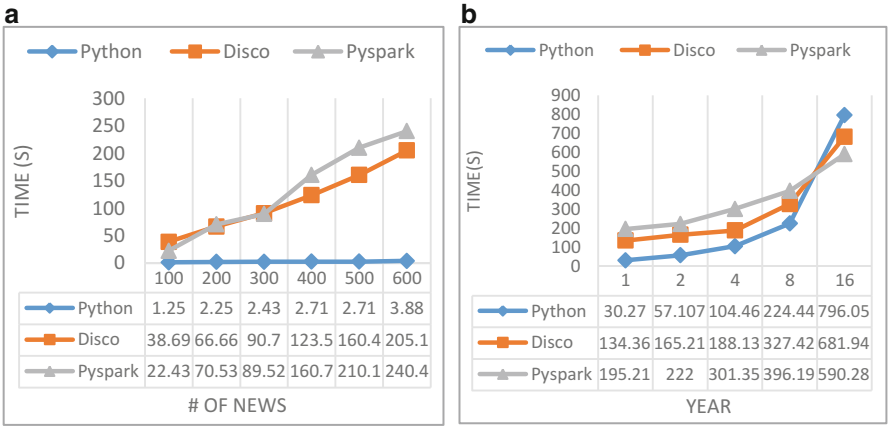



**Fig. 6.5** (**a**) Calculation time for real cases (600 news/MB). (**b**) Calculation time for Big Data sets (60,000 news/year)

Figure 6.5a shows the real calculation times of news-processing steps on different platforms. In our experiment, we found that we collected only 600 news items per day performed by the financial news web crawler. On the Python platform, it takes less than 5 s to finish this step; however, on the Disco and Spark platforms, it takes more than 200 s to finish the job. First, the heavy-duty Big Data platforms (Spark and Disco) require long calculation times because our news datasets are not big enough to fit the MapReduce model. A typical new item has only 30 lines and produces a small dataset that performs badly in the MapReduce model.

Second, launching Spark and Disco applications incurs several seconds of start-overhead. As mentioned earlier, each new job is considered as a single task, and when the datasets are small, numerous instances of start-overhead add up into a major performance issue.

To determine the effects of expanding the news dataset in our model, we combine an entire year's worth of news into one news file. In Fig. 6.5b, the Big Data platform (Spark and Disco) curves show slow growth, and the Python curve shows a linear relationship. However, Python shows positive exponential growth when the resource limit is reached.

According to the simulation results, it is crucial for researchers to choose an appropriate data processing platform. For example, if the dataset is too small to be computed, using regular high-end programming languages is the optimal choice. There are various practical applications that fit the MapReduce models. MapReduce is a flexible, highly fault-tolerant, and distributed processing framework, which can process massive data efficiently. We compare the performance between Pyspark and Disco. The module of Pyspark uses a Java Virtual Machine (JVM), and the module of Disco uses Erlang. If you want to implement your project on Big Data platform, you must consider the sizes of datasets and their effects upon start-overhead. In our case, the start-overhead of Pyspark is higher than Disco.

#### 6.5.2.2 Mobile Messaging System to Overcome Bursty Traffic and Network Instability

To deliver numerous instant messages to customers, we develop an adaptive task assignment algorithm on our service framework to overcome bursty traffic and mobile network instability.

System Architecture

The delivery service is built on the top of a proven and widely used open source XMPP project, Openfire, with a Connection Manager (CM) Module that can be horizontally scaled to meet increasing demands, as shown in Fig. 6.6.

System Behavior and Bottleneck

After the Load Balancer selects a CM to build the keep-alive connection with an incoming user, the messages sent to that user go through the same connection path until that user goes offline. Because each CM queue is a single FIFO queue and the mobile network is unstable, when bursty traffic happens, performance is greatly affected by how the Load Balancer selects the CM to build connections with users. Thus, our goal is to predict the network delay for an adaptive task assignment algorithm to reduce the average response time when the system handles bursty traffic in an unstable network environment.

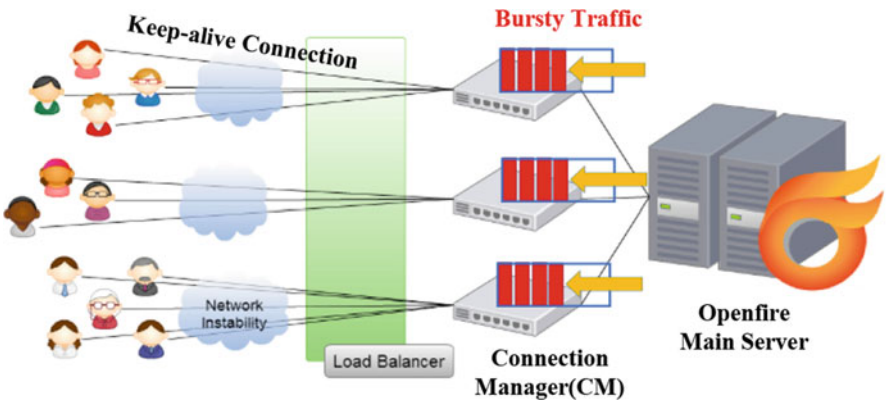



**Fig. 6.6** System architecture

Adaptive Task Assignment Algorithm

We predict future network delays through short-term historical data by using Queueing Theory (Harchol-Balter 2013) to model our system and to predict the system performance. The task assignment algorithm that we propose can predict future network delays from historical data; it is named SeeFuND. Consider that there is an interval of time between the algorithm processing time and the actual arrival time. In the interval of the same $\boldsymbol{CM_k}$ queue, there is an average probability of 0.5 that messages for other existing users might cut-in in front of the incoming message. We denote that probability as $\boldsymbol{q_{cut\_in^k}}$. Given this phenomenon, a correction term is added in the equation. Therefore, the waiting time $\boldsymbol{W^k_{new}}$ is predicted as follows: $\boldsymbol{W^k_{new} = \left(q^k_{now} + q_{cut\_in^k}\right) \times \overline{S}_k + W^k_0}$ where $\boldsymbol{q_{cut\_in^k} = \frac{user^k_{now}}{2}}$ and $\boldsymbol{W^k_0}$ is the remaining service time.

The SeeFuND task assignment algorithm calculates the expected waiting time $\boldsymbol{W^k_{new}}$ of each CM queue, then the queue with the shortest expected waiting time is selected to connect to incoming user.

Emulation Results

We compare the performance of different task assignment policies like Round-Robin (RR) (Xu and Huang 2009), Random (Buot 2006, pp. 395–396), Shortest Queue First (SQF) (Teo and Ayani 2001, pp. 185–195), and Least User (Connection) First (LUF) (Choi et al. 2010, pp. 127–134). Virtual servers are provided by Digital Ocean's IaaS; user robots are deployed evenly in different datacenter regions in several countries. Each host is run on an Ubuntu14.04 Linux server with 1 GB RAM, a single-core processor, and 30 GB of disk storage. Other factors in the emulating environment are the same as before.

In Fig. 6.7, the result shows the response times of different task assignment algorithms for different intensities of bursty traffic. The nonbursty traffic state is such that the arrival of messages to each queue follows a Poisson process. The arrival rate of low-intensity bursty traffic is 40 messages/s per queue, and the arrival rate of high-intensity bursty traffic is 150 messages/s per queue.

We found that when the intensity of bursty traffic increases from 40 messages/s per queue to 150 messages/s per queue, the SeeFuND algorithm shows a minimally increasing response time (Seng et al. 2017).

### *6.5.3 Implementation: Mobile Application*

For combining mobile devices and cloud computing, we develop an Android application to demonstrate our research. Users can connect to our cloud server and

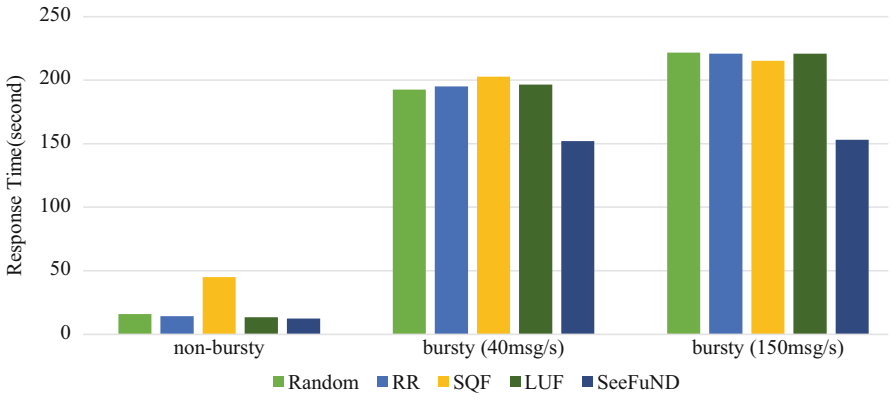



**Fig. 6.7** Response times of different task assignment algorithms for different intensities of bursty traffic

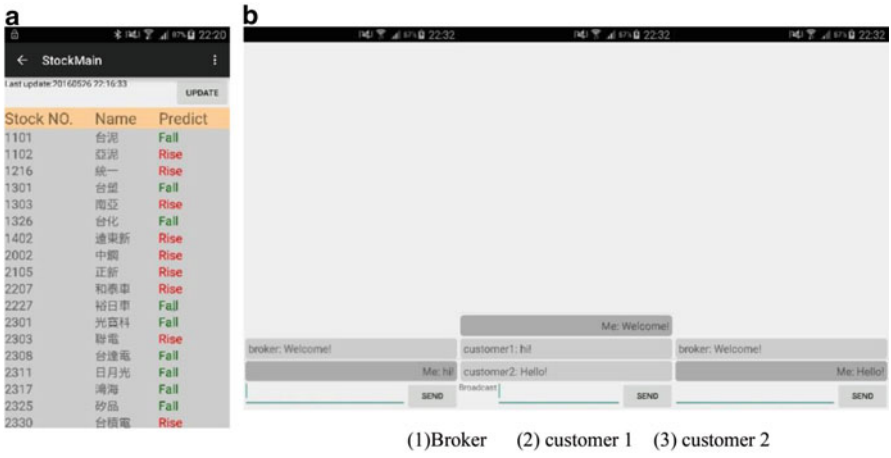



**Fig. 6.8** (**a**) Application for forecasting results. (**b**) Application for chat function which displays broadcast and unicast communications

get the service by this application. We display our forecast result, which predicts Taiwan's top 50 stocks, in Fig. 6.8a.

We provide the latest formal corporate announcements on a real-time platform named Market Observation Post System. The second part is a chat function which lets users can communicate with others directly without switching applications, as shown in Fig. 6.8b.

## 6.6 Concluding Remarks

Big Data, Artificial Intelligence (AI), Internet of Things (IoT), and Cloud Computing's widespread popularity and ubiquity have revolutionized the face of business in the world, especially for the financial service industry, such as FinTech (financial technology). With the emergence of service innovation and innovation integration, the focus has shifted to the development of novel financial services with data analytics, business intelligence, mobile technologies, and cloud computing that diversify and customize the new functionalities available to business clients and customers. Most of the worldwide financial holdings corporations have faced the biggest challenges resulting from the dynamic business environments, diverse telecommunication settings, and heterogeneous data sources that must work together in an integrated, real-time, and seamless manner. Today, businesses have accumulated large numbers of online services and data sources that run and reside on a variety of environments.

Furthermore, they have heterogeneous business models, business processes, and workflows to interact with. The results have been upheaval, chaotic, and disruptive. Therefore, it is crucial to develop a comprehensive, cross-disciplinary, and in-depth resolution of services, methods, and technologies to address the above issues. In this chapter, a set of cross-disciplinary sections are written up and devoted to describe and discuss the data analytics, service innovation, mobile technology in the financial services. We consider the next generation of innovation and integration of financial services in the financial service industry. We propose to investigate the key research issues of the innovation, intelligence, integration of new financial services over mobile technologies, and cloud computing information systems. We conduct extensive literature reviews to develop novel research models, to collect domain databases, to perform live case studies and experiments, and to create prototyping systems and evaluation. Empirical and econometric studies, data and text mining algorithms, selection and interaction models, real life case studies, mobile prototyping, and performance evaluation are performed.

## References

Antweiler, W., & Frank, M. Z. (2004). Is all that talk just noise? The information content of internet stock message boards. *The Journal of Finance, 59*(3), 1259–1294.

Arner, D. W., & Barberis, J. (2015). FinTech in China: From the shadows? *The Journal of Financial Perspectives, 3*, 78–91.

Baker, M., & Wurgler, J. (2006). Investor sentiment and the cross-section of stock returns. *The Journal of Finance, 61*(4), 1645–1680.

Brown, G. W., & Cliff, M. T. (2005). Investor sentiment and asset valuation. *The Journal of Business, 78*(2), 405–440.

Buot, M. (2006). Probability and computing: Randomized algorithms and probabilistic analysis. *Journal of the American Statistical Association, 101*(473), 395–396.

Choi, D., Chung, K. S., & Shon, J. (2010). An improvement on the weighted least-connection scheduling algorithm for load balancing in web cluster systems. In *Grid and distributed computing, control and automation* (pp. 127–134).
Chuen, D. L. K., & Teo, E. G. S. (2015). Emergence of fintech and the LASIC principles. *The Journal of Financial Perspectives, 3*, 24–37.
Dean, J., & Ghemawat, S. (2008). MapReduce: Simplified data processing on large clusters. *Communications of the ACM, 51*(1), 107–113.
Deloitte. (2016). *Perspectives: Banking and securities outlook 2017*. http://www2.deloitte.com/us/en/pages/financial-services/articles/banking-industry-outlook.html
Fama, E. F., & French, K. R. (1992). The cross-section of expected stock returns. *The Journal of Finance, 47*(2), 427–465.
Fama, E. F., & French, K. R. (2015). A five-factor asset pricing model. *Journal of Financial Economics, 116*(1), 1–22.
Fang, L., & Peress, J. (2009). Media coverage and the cross-section of stock returns. *The Journal of Finance, 64*(5), 2023–2052.
Financial Supervisory Commission. (2016). *Fintech-development strategy white paper*. Taipei: Financial supervisory commission.
Garcia, D. (2013). Sentiment during recessions. *The Journal of Finance, 68*(3), 1267–1300.
Gulamhuseinwala, I., Bull, T., & Lewis, S. (2015). FinTech is gaining traction and young, high-income users are the early adopters. *The Journal of Financial Perspectives, 3*, 16–23.
Harchol-Balter, M. (2013). *Performance modeling and design of computer systems: Queueing theory in action*. Cambridge: Cambridge University Press.
Hu, N., Bose, I., Koh, N. S., & Liu, L. (2012). Manipulation of online reviews: An analysis of ratings, readability, and sentiments. *Decision Support Systems, 52*(3), 674–684.
Ittoo, A., & Bouma, G. (2013). Term extraction from sparse, ungrammatical domain-specific documents. *Expert Systems with Applications, 40*(7), 2530–2540.
Kearney, C., & Liu, S. (2014). Textual sentiment in finance: A survey of methods and models. *International Review of Financial Analysis, 33*, 171–185.
KPMG. (2015a). *Mobile banking is a key selling point for a growing number of customers* (pp. 1–8). Amstelveen: KPMG International Cooperative.
KPMG. (2015b). *Digital offerings in mobile banking—The new normal* (pp. 1–10). Amstelveen: KPMG International Cooperative.
KPMG. (2015c). *Mobile banking-global trends and their impaction banks* (pp. 1–33). Amstelveen: KPMG International Cooperative.
Kumar, A., & Lee, C. (2006). Retail investor sentiment and return comovements. *The Journal of Finance, 61*(5), 2451–2486.
Kurov, A. (2010). Investor sentiment and the stock market's reaction to monetary policy. *Journal of Banking & Finance, 34*(1), 139–149.
Liu, B. (2012). Sentiment analysis and opinion mining. *Synthesis Lectures on Human Language Technologies, 5*(1), 1–167.
Loughran, T., & McDonald, B. (2011). When is a liability not a liability? Textual analysis, dictionaries, and 10-Ks. *The Journal of Finance, 66*(1), 35–65.
Manning, C. D., Raghavan, P., & Schütze, H. (2008). Scoring, term weighting and the vector space model. In *Introduction to information retrieval* (Vol. 100, pp. 2–4).
Seng, C., Wu, T., & Chang, Y. (2017). *Innovative, integrated, mobile financial services* (MOST three-year final research project report, MOST 104-2627-E-004-001).
Sharpe, W. F. (1964). Capital asset prices: A theory of market equilibrium under conditions of risk. *The Journal of Finance, 19*(3), 425–442.
Teo, Y. M., & Ayani, R. (2001). Comparison of load balancing strategies on cluster-based web servers. *Simulation, 77*(5–6), 185–195.
Tetlock, P. C. (2007). Giving content to investor sentiment: The role of media in the stock market. *The Journal of Finance, 62*(3), 1139–1168.
Tetlock, P. C., Saar-Tsechansky, M., & Macskassy, S. (2008). More than words: Quantifying language to measure firms' fundamentals. *Journal of Finance, 63*(3), 1437–1467.

World Economic Forum. (2015). *Global agenda council on the future of financing & capital: The future of FinTech a paradigm shift in small business finance*. Retrieved from http://www3.weforum.org/docs/IP/2015/FS/GAC15_The_Future_of_FinTech_Paradigm_Shift_Small_Business_Finance_report_2015.pdf

World Economic Forum. (2017). *Beyond Fintech: How the successes and failures of new entrants are reshaping the financial system*. Part of the Future of Financial Services series | Prepared in collaboration with Deloitte, August 2017. http://www3.weforum.org/docs/Beyond_Fintech_-_A_Pragmatic_Assessment_of_Disruptive_Potential_in_Financial_Services.pdf

Xu, K., Liao, S. S., Li, J., & Song, Y. (2011). Mining comparative opinions from customer reviews for competitive intelligence. *Decision Support Systems, 50*(4), 743–754.

Xu, Z., & Huang, R. (2009). *Performance study of load balancing algorithms in distributed web server systems* (CS213 Parallel and Distributed Processing Project Report, 1).

Yin, R. K. (1994). *Case study research: Design and methods* (2nd ed.). Newbury Park: Sage Publications.

Zhang, Y., Dang, Y., & Chen, H. (2013). Research note: Examining gender emotional differences in web forum communication. *Decision Support Systems, 55*(3), 851–860.

# Chapter 7
# Health in Biodiversity-Related Conventions: Analysis of a Multiplex Terminological Network (1973–2016)


**Claire Lajaunie, Pierre Mazzega, and Romain Boulet**

## 7.1 Introduction

In 2015, the World Health Organization (WHO) and the Secretariat of the Convention on Biological Diversity (CBD) decided to sign a Memorandum of Understanding to strengthen their collaboration notably in raising awareness of the complex linkages between biological diversity, ecosystems, and human health. The same year they published a joint report entitled *Connecting global priorities: biodiversity and human health: a state of knowledge review* highlighting the fact the biodiversity loss constituted a '*fundamental risk to the healthy and stable ecosystems that sustain all aspects of our societies*' (WHO - CBD 2015). In a previous work we decided to examine in detail how this awareness of the interrelations between ecosystems, animal health, and human health evolved by tracking the emergence of health issues in biodiversity-related conventions. Encompassing human health, the One Health approach has been acknowledged by the CBD as it integrates '*the complex relationships between humans, microorganisms, animals, plants, agriculture, wildlife and the environment*' (Lajaunie and Mazzega 2016).

C. Lajaunie (✉)
INSERM/CERIC, UMR DICE 7318, CNRS et Aix-Marseille Université,
Aix-en-Provence Cedex 1, France
e-mail: claire.lajaunie@inserm.fr

P. Mazzega
GET Géosciences Environnement Toulouse UMR5563, CNRS/IRD/Université de Toulouse,
Toulouse, France
e-mail: pierre.mazzegaciamp@get.omp.eu

R. Boulet
Université de Lyon, Jean Moulin, Institut d'Administration des Entreprises de Lyon,
Centre de Recherche Magellan, Lyon, France
e-mail: romain.boulet@univ-lyon3.fr



S.-H. Chen (ed.), *Big Data in Computational Social Science and Humanities*,
Computational Social Sciences, https://doi.org/10.1007/978-3-319-95465-3_7

Then in a following study, we analyzed the transmission, circulation, and persistence of health issues into three international conventions related to biodiversity, the Convention on Biodiversity, the Convention on Migratory Species (CMS), and the Convention on International Trade in Endangered Species of Wild Fauna and Flora (CITES). The results have shown the centrality of the Convention on Biodiversity as a source of terms related to health and the environment (Lajaunie and Mazzega 2017). Interested by the emergence and integration of health issues, in their various dimensions, into the biodiversity-related conventions, we wanted to show the complexity and the interdependence of those issues and aimed at clarifying their multi-dimensional aspect.

Having realized the mining of the textual corpus associating the text of the CBD, the CMS and the CITES conventions themselves, and the decisions or resolutions of their respective Conferences of the Parties (COPs) from the first ones (1973) to June 2016, we obtain more than 22,172 complex nominal terms among which 213 are related to Health. Those terms are organized hierarchically into "micro-ontologies" or concepts, specific to each concept linked to Health (e.g., biodiversity, disease, health, pathogen, safety and security, warning, etc.) as explained in Sect. 7.2. We thus analyze how micro-ontologies are used in a complete or partial form in each COP and how they are transmitted between COPs through a multiplex network (Sect. 7.3): each type of link of the network corresponds to a concept. Then, we identify the most central COPs and their gathering into communities in the process of Health issues emergence.

Our aim is to study the dynamic of health within the COPs of the Convention on Biodiversity and to capture simultaneously the dynamic of the importance of COPs in the diffusion of health issues and the dynamic of health themes within the biodiversity-related conventions (Sect. 7.4). The common dynamic shown, thanks to the use of multiplex analysis combined with text mining used in a big data perspective, helps to understand how each concept contributes to the building of an integrative and multi-dimensional approach of Health issues within the CBD, CMS, and CITES (Sect. 7.5). The main conclusions are then presented in Sect. 7.6.

## 7.2 Data Acquisition and Pre-Processing

We examine here three international environmental agreements that are considering human, animal, or ecosystem health issues: the CBD, CMS, and the CITES. For each of them, the COP is the governing body gathering representatives of governments and regional organizations that have ratified the Convention. Each COP convenes its members at regular intervals to discuss possible measures to be taken to update the implementation of the Convention and publish on their respective official website a series of decisions (CBD) or resolutions (CMS, CITES).

### 7.2.1 The Conventions

Opened for signature in June 1992 (Earth Summit; Rio de Janeiro, Brazil) and entered into force in 1993, the Convention on Biological Diversity promotes the conservation of biological diversity, considering many related issues like the sustainable use of biodiversity, or the equitable sharing of benefits (Morgera 2015) resulting from the utilization of genetic resources, in particular through the work of an intergovernmental advisory body and several specialized working groups.[1] We analyze the CBD's decisions because, as we have shown, this convention plays a central role in the emergence of health and environmental themes. The first COP was held in Nassau, Bahamas, in 1994, and has been held every 2 years since 1996. The last COP met in December 2016 in Cancun, Mexico, and resulted in decisions that are not included in our data set, as they are not available in their final form while we finalize this chapter. In total, we will analyze, using text mining, all the decisions available at the beginning of June 2016, i.e., a total of 367 decisions covering all the areas of competence of the COPs as well as the decisions concerning the governance of the CBD.[2]

The Convention on the Conservation of Migratory Species of Wild Animals is presented as a "*global platform for the conservation and sustainable use of migratory animals and their habitats*".[3] It was opened for signature in 1979 and entered into force in November 1983. Among the interactions between environmental change and health, the emergence or re-emergence of infectious diseases is mainly related to zoonoses, carried not only by domestic animals but also by wildlife, including some migratory species. Since about a decade, CMS has taken these dynamics into account in the work of its bodies. In addition to the Conferences of the Parties, the governance of the CMS involves a Standing Committee and a Scientific Council, both meeting about once a year. The COP publishes sets of resolutions on a period of 3 years (except the time interval 1997–1999 between the fifth and sixth COPs) since the first COP held in Bonn, Germany, in 1985. In this study we use the texts of the 175 resolutions published along the 11 COPs of the CMS held between 1985 and 2014.

We analyze also 89 resolutions published along the 16 COPs of the Convention on International Trade in Endangered Species of Wild Fauna and Flora[4] held between 1976 (1st COP) and the Sixteenth meeting of the Conference of the Parties in Bangkok, Thailand in March 2013. The data from the very recent 17th COP (Johannesburg, South Africa; 24 September—04 October 2016) are not included in our network analysis. The illegal trade in wild animals can be a significant vector for the spread of diseases affecting human health, especially across borders. The CITES

[1] For example, in 1999, the Open-ended Ad Hoc Working Group on Biosafety submitted a draft text of the Cartagena Protocol on Biosafety to the Convention on Biological Diversity.

[2] See the Convention's website at https://www.cbd.int/ (accessed December 19th, 2016)

[3] See the Convention's website at http://www.cms.int/en/ (accessed December 19th, 2016)

[4] See the Convention's website at https://www.cites.org/ (accessed December 19th, 2016)

itself was opened to signature in 1973 and entered into force 2 years later. The text of the three conventions (CBD, CMS, and CITES) is included in our textual corpus for mining.

### *7.2.2 Terminology Extraction*

The textual corpus analyzed thus includes all the decisions and resolutions of the COPs of the CBD, CMS, and CITES published until June 2016, plus the text of the conventions themselves. First, we extracted from this corpus all the "complex terms," that is, noun-phrases (word or group of words that functions in a sentence as subject, object, or prepositional object) composed of several simple terms (for example, a name and an adjective such as "biological diversity" or "ecosystem approach," or noun, preposition, adjective and noun like "emergence of infectious disease", and so on). Simple terms are usually too generic and cannot be linked to a defined topic—such as human or animal health. We thus obtain more than 22,000 distinct complex terms.

From this list we remove general terms (e.g., "international organization," "general provision," "accessible checklist," "consideration of relevant question," etc.), expressions referring to objects or to a context that is too local or specific (e.g., "fourth meeting," "next fiscal year," etc.), the sequences of characters corresponding to a spelling error or nonsense itself (e.g., numbering of text sub-sections like paragraphs or articles), and so on. This filtering is carried out by constituting a first list of undesirable terms obtained by reading a list of complex terms extracted from a group of decisions or resolutions. The second group is sifted through this filter and new but undesirable terms are identified manually and added to the filter as updated. The process is applied until the initial complete list of terms is exhausted and the final filter is once again applied to the whole corpus. This simple approach combines the efficiency of automatic iterative filtering and the accuracy of a choice of undesirable terms by one or more experts in the field. At the end of this filtering process, we obtained a list of 8867 separate and complex terms used in the CBD texts, 2565 terms from CMS, and 3450 terms from CITES.

This new list was then independently re-read by two people with independent academic backgrounds and scientific cultures (legal studies versus modeling), to retain only those terms that are more or less directly related to health and environment issues in a broad conception of Health, as conveyed by the One Health (WCS 2004; Hall and Coghlan 2011) or EcoHealth (Zinsstag et al. 2005; Brown 2007) movements, in particular (Zinsstag 2012), or in some of the most recent interdisciplinary works (Walther et al. 2016; Morand et al. 2015; Lajaunie et al. 2015). Indeed, the choices made depend on the current view of what are the links between health and environment in a rapidly evolving international legal and political context (Lajaunie 2016). We obtain a list containing 72 complex terms or expressions related to the theme "health–environment" in the texts of the CBD, 91 in the texts of CMS, and 50 terms in CITES texts (see Lajaunie and Mazzega 2017

for details). It should be noted that approximately 95% of the terms were retained by the two persons separating them independently from one another, and the remaining 5% were treated without dispute, with a final common decision to include them or not after a brief exchange of arguments.

### *7.2.3 Micro-Ontologies for Text Mining*

The analysis of the emergence of the "health and environment" theme in the conventions on biological diversity requires the use of specific terms over time. Some terms, such as "risk" or "threat," for example, can of course be used without reference to health issues, and then gradually see their regular use in relation with this topic. Thus the conceptual field to which a given term refers is not fixed but on the contrary evolves with the development of knowledge, the legal frameworks of public action, and political consciousness. The terms are here conceived as kinds of tracers—certainly indirect and subject to interpretation—of these evolutions.

The follow-up of the use of a given term or expression is partly random: according to the drafters of decisions or resolutions and also according to the derivatives of the terminology used at a given period in a particular context (e.g., the meetings of the COPs), an expression may be imposed gradually or on the contrary be set back against a relatively synonymous expression that might be covering a wider field (see the case of "One Health" in Lajaunie and Mazzega 2016). To overcome this difficulty, we group together a set of terms under the same label: the occurrence of a term in this list is counted as a reference to the list itself, or equivalent to the concept designated by this label. This procedure is the one developed in text mining (Feldman and Sanger 2007), the concept being defined as a set of terms. The use of any of the terms in the set is considered an occurrence of the concept. This approach has two advantages: (a) terms with similar meanings can be grouped together under the same concept, thus improving the precision and recall of the text mining procedure (e.g., Hassanpour and Das 2011); (b) the relative importance of each concept in the production of meaning is reinforced for concepts that contribute more broadly to the semantics of the textual segments considered (Durga et al. 2012).

We use the following 13 concepts: biodiversity, disease, health, impact, knowledge, mortality, "pathogen," process, resource, risk or threat, security, technology, and warning. They cover all the terms related to the theme "health–environment" derived from our textual corpus,[5] each term being linked to a single concept. Let us note that we have verified that the "label" used for these concepts functions in a way like the lemmatized form in our corpus: for example, the term "pathogen" also

[5] In other words, this list of concepts is adapted to the corpus that we analyze here. The analysis of another corpus or an extended body (for example, epidemiology, medicine, or veterinary science) would probably lead us to modify this list.

makes it possible to find the form "pathogenic," "health," also the form "healthy," and so on. The terms members of a concept (seen as a set) were then organized hierarchically for other applications. We call these hierarchies "micro-ontologies"—(an example of which is given in Lajaunie and Mazzega 2017, Fig. 7.2)—because, on the one hand, relations between terms of the same concept are mainly of type "*is a*" or "*is a part of*" (Grüber 1995), but, on the other hand, they cover only the terms appearing in our specific corpus and for the precise purpose of our analysis (hence the preposition "micro-"), but not a domain of knowledge or fundamental concepts (like "space," "time," "relation," etc.) as in a core-ontology (Guarino and Musen, 2015).

For each of these 13 micro-ontologies, we count the number of occurrences of the corresponding "concept" (occurrence of any of its constituent terms) in the biodiversity-related conventions and decisions or resolutions of the COPs. Thus we obtain the data used in the multiplex analysis of Conventions (and their COPs) and concepts in health–environment.

## 7.3 Multiplex Network Analysis

Graph theory is the mathematical foundation of network analysis. A graph is defined by a set of vertices (or nodes) and a set of edges (or links). An edge can be seen as a couple of nodes. This structure of simple graph can be enriched by adding information on the edges like a direction, a weight, a color, or a label. Thanks to network analysis, graphs can be seen as an interdisciplinary tool between mathematics and social sciences or humanities. We can think, for example, about sociology with the study of social networks but complex legal structures have also interacted with mathematics through graph theory at the national (e.g., Bommarito and Katz 2010; Boulet et al. 2010; Boulet et al. 2011) or international (e.g., Kim 2013; Boulet et al. 2016) levels.

### *7.3.1 Network Modeling: From a Bipartite Graph to a Multiplex Network*

Our data on concepts (see Sect. 7.2) occurring in the legal corpus is modeled by a bipartite graph that is a graph with two kinds of vertices (COPs and terms), with edges between vertices of different kinds, and with no edge between two vertices of the same kind. Here the first kind of vertices is constituted by the 27 COPs and the second kind of vertices by the 13 concepts (labels of the micro-ontologies). There is a link between a COP and a concept if the concept is used in a decision or resolution of the COP. The resulting graph has 108 edges, representing 31% of all possible links (there are $27 \times 13 = 351$ possible links). Some COPs do not use any concept

of our list of terms in "health–environment" and therefore will no longer appear in the representations of the graphs.[6]

From this bipartite graph we can perform a projection on the COP vertices: we create a new graph the vertices of which are the COPs and there is an edge between two COPS if they are linked (in the bipartite network) to a same term (or concept). However in order to make this projection more relevant and not to lose information we keep information on the edges: the resulting graph has 13 kinds of edges which can be represented as 13 different colors of edges, each color corresponding to a term. Such a network, with different kinds of edges, is called a multiplex network. This graph has 27 vertices and 620 edges (there can be several edges of different colors between two COPs because two COPs can have several terms in common).

### *7.3.2 Dynamics of Centralities*

Centralities in network analysis are measures on vertices of a graph aiming at assessing the position of vertices in the graph that is the "role" of a node in a network. Originally used in sociology (Freeman 1979) when studying social networks, notably to measure the power of a person, the use of these measures has become generalized to other types of networks and relations. The degree centrality is defined by the number of links of a node, indeed the degree of a vertex on a graph is the number of incident edges to this vertex. The underlying idea of degree centrality is that a person sharing many social links is popular and can directly influence many people. This measure is widely used in online social networks: we count how many friends we have on Facebook, how many followers on Twitter, etc.

Another way to have power in a network is to be a key intermediary between other nodes or between parts of the network. This property is measured by the betweenness centrality. Betweenness centrality of a vertex is computed by counting the number of times that vertex belongs to a shortest path between two other vertices (this count is weighted in case of multiples shortest paths between two given vertices). Thus, a vertex with a high betweenness centrality is a vertex which is often on shortest paths, playing an important role in the connectivity of the network.

In this section we compute and analyze degree and betweenness centralities for COPs and for terms. For this purpose we consider the bipartite networks (presented in Sect. 7.3.1) at different dates (the years in which one or several COPs were held). Since the degree of a COP does not vary with time (the number of concepts used being determined once the decisions or resolutions are published), the dynamics analysis is performed for the betweenness centrality of the COPs and the degree and betweenness centralities of the terms. Therefore we could not only know which COPs or concepts are currently more central but also when they appear as essential in the emergence of health–environment issues (as detailed in Sect. 7.4).

[6]These are the CMS COPs number 2, 4, 5, and 6, and CITES COPs number 1–8 and 14–16.

### 7.3.3 Community Detection

A community in a network is a group of vertices (a group of individuals) more densely connected than the overall density of the network. In other words, there are (in proportion) many more links within a community than in the global network. This joins the intuitive idea of a social community seen as a group of individuals sharing many social ties. The division of a network into communities is not unique. Indeed a vertex may belong to several communities, for instance, and the number of communities is itself unknown. There are various algorithms to detect communities in a network some of these methods being presented in Fortunato (2010).

Extracting communities of COPs from the bipartite graph can be done by performing a hierarchical clustering on a dissimilarity matrix. The Dice dissimilarity is well adapted to reveal the communities of a network (Kuntz, 1992). With this dissimilarity, two COPs will be similar if they both use many common terms and thus two COPs are dissimilar if they use few common terms. The Dice dissimilarity is obtained by computing for each pair of COPs the ratio between the number of terms that are not common to the two COPs and the sum of the degrees of the two COPs. Once we have a dissimilarity matrix between the COPs we can gather the COPs by performing a hierarchical clustering. The result of this clustering is presented in Table 7.1 and discussed in the next section.

Let us now detail the method of spectral clustering, von Luxburg (2007) providing a good survey on these methods. We can associate several matrices to a graph G on N vertices:

- The adjacency matrix, denoted A, which is a matrix with N rows and N columns where the input (i, j) equals 1 if there is an edge between the vertices i and j, and equals 0 otherwise;
- The diagonal matrix of degrees D which ith diagonal term is equal to the degree of the vertex i (and the non-diagonal terms are zero);
- The Laplacian matrix, denoted L, which is given by $L = D - A$ and normalized Laplacian matrix given by $D^{-1/2}LD^{-1/2}$. These matrices are diagonalizable and can be written as $P\Lambda P^{-1}$, P being the matrix of the eigenvectors and $\Lambda$ the diagonal matrix of eigenvalues.

Spectral clustering consists first in a spectral embedding: the graph is embedded in a Euclidean space of dimension k (the k coordinates of the vertices are given by the k eigenvectors associated with the k smallest eigenvalues) and then on a clustering of these points in the Euclidean space (usual clustering methods like k-means can be used for this purpose). A spectral clustering performed on the simple graph resulting of the projection of the bipartite graph on the term-vertices gives two communities of terms given in Table 7.2 (see next section).

Based on both the spectral partitioning method and statistical factor analysis methods, we propose the following method for extracting communities from our multiplexed network. We first define a matrix B with 620 columns (each column represents an edge in the multiplex graph resulting from the projection of the

**Table 7.1** The four robust communities obtained by clustering the multiplex network of COPs and terms (concepts related to health–environment issues)

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Community cM1 | B03_1996 | B05_2000 | B06_2002 | B07_2004 | B08_2006 | B09_2008 | B012_2014 | | |
| | M08_2005 | M09_2008 | M10_2011 | M11_2014 | | | | | |
| | T09_1994 | T10_1997 | T11_2000 | T12_2002 | | | | | |
| | Diseases | Health | Impact | Knowledge | Mortality | Pathogen | Resource | Technology | Warning |
| Community cM2 | B00_1992 | B02_1995 | B04_1998 | B10_2010 | B11_2012 | | | | |
| | M01_1985 | M03_1991 | | | | | | | |
| | T013_2004 | | | | | | | | |
| | Biodiversity | Risk and threat | | | | | | | |
| Community cM3 | M00_1979 | M07_2002 | | | | | | | |
| | Process | | | | | | | | |
| Community cM4 | B01_1994 | T00_1973 | | | | | | | |
| | Security | | | | | | | | |

In the COP labels the letter indicates the Convention (B = CBD, M = CMS, T = CITES) followed by the COP number (0 for the Convention itself) and the year it was hold (e.g., B03_1996 is the 3d CBD COP held in 1996)

**Table 7.2** The four communities of COPS obtained from a hierarchical clustering on the Dice dissimilarity matrix (see text)

| | | | | | | |
|---|---|---|---|---|---|---|
| Community cC1 | B03_1996 | B05_2000 | B06_2002 | B07_2004 | B08_2006 | B12_2014 |
| | M08_2005 | M09_2008 | M10_2011 | | | |
| | T09_1994 | T10_1997 | | | | |
| Community cC2 | B02_1995 | B04_1998 | B09_2008 | B10_2010 | B11_2012 | |
| | M11_2014 | | | | | |
| | T11_2000 | T12_2002 | T13_2004 | | | |
| Community cC3 | B01_1994 | | | | | |
| | T00_1973 | | | | | |
| Community cC4 | B00_1992 | | | | | |
| | M00_1979 | M01_1985 | M03_1991 | M07_2002 | | |

In the COP labels the letter indicates the Convention (B = CBD, M = CMS, T = CITES) followed by the COP number (0 for the Convention itself) and the year it was hold

bipartite graph on the COP-vertices) and 40 rows (each row represents a vertex in the bipartite graph: a COP or a term that can also be seen as a color of an edge of the multiplex graph). Since each edge in the multiplex graph is defined by the two COPs it links and by the color associated with a term, the entries of the jth column of B (associated with the jth edge of the multiplex graph) are defined this way: B[i,j] = 1 if i is either a COP incident to the edge or i represents the color of the edge, and B[i,j] = 0 otherwise. Then we define the matrix E to be a diagonal matrix where the ith diagonal element equals the ith row sum of B and we define the matrix Q by the following product: $Q = E^{-1/2}(B^tB)E^{-1/2}$. Since the matrix Q is symmetric, it is diagonalizable and we perform a spectral embedding of the vertices on the Euclidean space for which the coordinates of the vertices are given by $P\Lambda^{1/2}$ where P is the matrix of eigenvectors and $\Lambda$ is the diagonal matrix of eigenvalues.

Once in a Euclidean space, we can perform a clustering by using classical tools like k-means or hierarchical clustering. The aim of the k-means method is to assign an individual to the class it is closest to and the aim of the hierarchical clustering is to successively merge the two closest classes. These two methods can give different results and, for robustness purposes, we carry out these two clustering and then extract robust communities that is groups of vertices which are gathered together whatever the clustering method used. These robust communities are given in Table 7.1. This method has the advantage of not only extracting communities of COPs but also assigning health–environment terms or concepts to each community on the basis of the corpus mining results.

## 7.4 Dynamic of Health Issues Within the COPs

The multiplex analysis provides a view of the dynamic of health issues in the COPs of the three biodiversity-related conventions with a fine-grained semantic resolution through the joint interactions between concept-uses and COPs (bipartite COPs—Terms graph). It also allows the detection of different communities, whether they

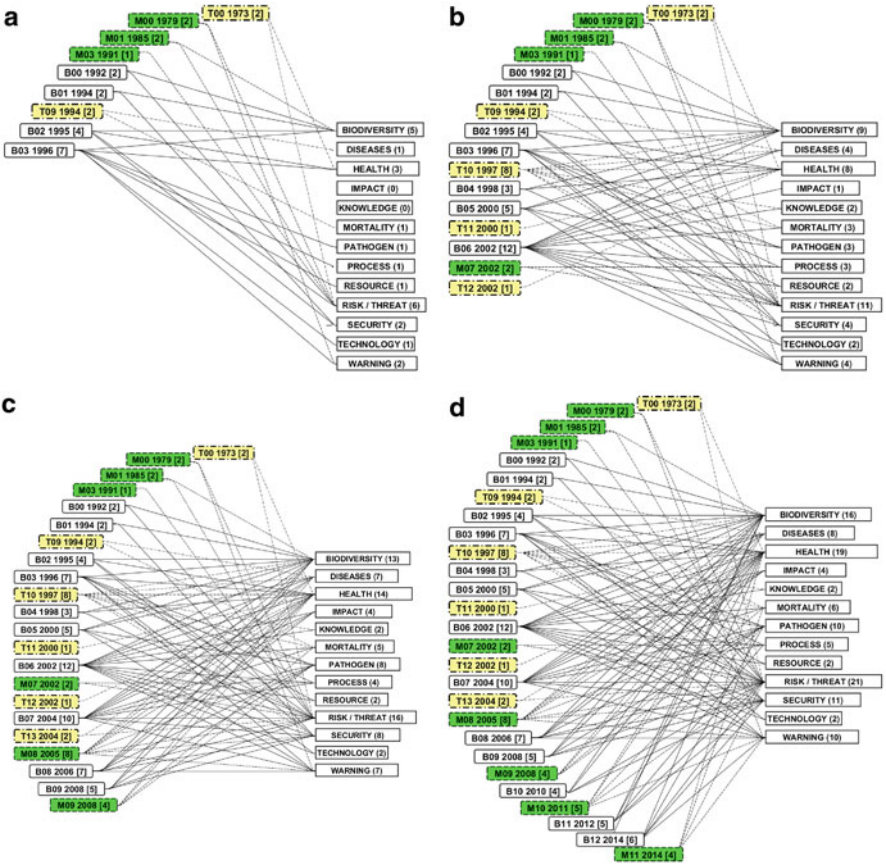



**Fig. 7.1** Snapshots of the cumulative temporal evolution of the bipartite graph including COPs (ellipses, left) and terms (rectangles, right). (**a**) Period 1973–1996; (**b**) Period 1973–2002; (**c**) Period 1973–2008; (**d**) Period 1973–2014. In the COP labels the letter indicates the Convention (B = CBD, M = CMS, T = CITES) followed by the COP number (0 for the Convention itself), the year it was hold. The degree centrality of each COP is given in brackets. The width of the rectangles containing the terms is increasing with the degree centrality (number of links) of the term (number in parenthesis)

group COPs and terms or they consider them separately. These communities are reflecting the dynamics emerging from the process of publication of COPs decisions and show how they express their relation to the terms.

From Fig. 7.1, we can detect that the highest degree centrality scores concern the COPs of the CBD, respectively, COP03, COP06, and COP07, then CITES, COP10 and CMS, COP08. In order to understand these results, it is worth noting that COP03 of 1996 contains the first decision intending to link and increase the relationship of the CBD with the Commission on Sustainable Development

and biodiversity-related conventions as well as other international agreements or institutions (Decision 3/21). As such, it insists on the formalized process of cooperation between the CBD, CITES ,and CMS. Then, COP06 of 2002 endorses the joint programme of CBD and CITES and emphasizes the need of increased coordination of activities between CMS and CBD (Decision 6.20). It also contains the Strategic Plan for the Convention on Biological Diversity (Decision 6.26) whose first goal is "*The Convention is fulfilling its leadership role in international biodiversity issues*". In 2002, COP07 considers many biodiversity areas (agricultural biodiversity, forest biodiversity, biodiversity of inland water ecosystems, or marine and coastal biodiversity) and decisions encompassing broad topics such as climate change, ecosystem approach, tourism, and assessments but, above all, it creates the Biodiversity Liaison Group to enhance the coherence and cooperation in their implementation (Decision 7.26).

In 1997 the COP10 of the CITES calls for increased cooperation and collaboration between CITES and CBD and considers the synergy between the two conventions as an opportunity particularly on the issue of introductions of invasive species (Decision 10.54, d). This synergy between biodiversity-related conventions should be developed further (Decision 10.63) as well as the way to harmonize their reporting requirements (Decision 10.10).

As for the CMS, COP05 in 2005 adopts a resolution on Migratory Species and Highly Pathogenic Avian Influenza which underlines the fact that "*some diseases have the potential to reduce biodiversity*" (Resolution 8.27). In line with the decision 7.26 of the COP07 of the CBD, COP05 also adopts a resolution for an enhanced cooperation among the biodiversity-related conventions (resolution 8.11) and a resolution to assess the Contribution of the CMS in Achieving the 2010 Biodiversity Targets (resolution 8.07) and to ensure that on-going and future CBD programmes of work appropriately integrate migratory species at the global level (resolution 8.18).

The semantic evolution in relation to the micro-ontologies or concepts developed in Sect. 7.2 shows, for instance, the importance of the concept of risk and threat from 1973 in the CITES Convention itself. Indeed in the Convention the term of risk was used in relation to the transportation of living specimens: the Parties must take measure to "*minimize the risk of injury, damage to health or cruel treatment*". Here the term risk is from the beginning used in relation to health issues. The Convention on Migratory Species of 1979 prefers the term "threat" which appears in relation to species at threat whether it is direct threat or indirect threat (reduction of breeding success due to pesticide contamination). The reduction of breeding success is not expressed in term of health of species nevertheless we can underline that it is a health issue in itself (Res. 1.5, §3). The habitat loss or fragmentation identified as another threat for migratory species could be seen as a health threat as well (Res. 3.2, §4).

From the beginning the Convention on Biodiversity of 1992 considers the adverse environmental impacts as "*risks to human health*" or study the risk related to the use of living modified organisms. The term of threat is also used in association

to species, ecosystems, or biological diversity. This semantic use of risk on the one hand and threat on the other hand will persist in the decisions of the COPs of the CBD: risk will also appear in the expressions "risk management" and "risk assessment." The concept of risk and threat is the most frequently used from the beginning (Fig. 7.1a) and persists as such over time (Fig. 7.1b–d). This example gives various pieces of information about the considered concept or micro-ontology; for instance, it indicates the degree of concern over time and a detailed analysis shows that different Conventions can have preferences for a term rather than another and association of terms can appear and evolve through time. From Fig. 7.1, we can see the continuous increase of the degree centrality of the terms "risk and threat" (21) and "health" (19), and starting later on, the degree centrality of the term security (11).

In Table 7.1, communities 3 and 4 are of little importance. There is a clear separation between the terms Disease and Health (belonging to the community cM1) and biodiversity and risk/threat (Community cM2). Let us note that risk is apart from security. We can underline the fact that the emerging communities are not gathering COPs from the same convention: they merge COPs of different conventions. The communities help to highlight the common interest or issues the conventions are dealing with. cM1 covers a period from 1994 to 2004 while cM2 spans the period 1985–2012. It shows that the two communities co-exist at the same time. The terms Disease and Health are not limited to human health with a turning point in 2004 showing a concern for environmental (ecosystem) and animal health in CITES and CMS (for a complete analysis Lajaunie and Mazzega 2016). The text of CMS and CITES conventions does not concern either biodiversity or health (Conventions M00 and T00 are in communities cM3 and cM4): therefore, the study of communities allows understanding the evolution of concerns of the conventions expressed by the work of the COPs.

## 7.5 Discussion

The importance of COPs in disseminating health issues cannot be measured solely by the number of concepts (and therefore by the degree centrality) used in the decisions or resolutions they publish. In a similar way to what is done in the analysis of social networks, the "betweenness" of the COPs plays an important role in the continuity and the persistence in time of the topics addressed and treated by these conferences. Indeed, a COP has a high betweenness centrality when it binds together, by the terms it conveys, groups of COPs which otherwise would tend to be separated. The temporal evolution of the betweenness centrality of the most central COPs is presented in Fig. 7.2. The curve shape shows the dynamic of biodiversity-related conventions and the COPs having the highest betweenness, the betweenness expressing the role played has a vehicle of transmission of health terms between COPs.

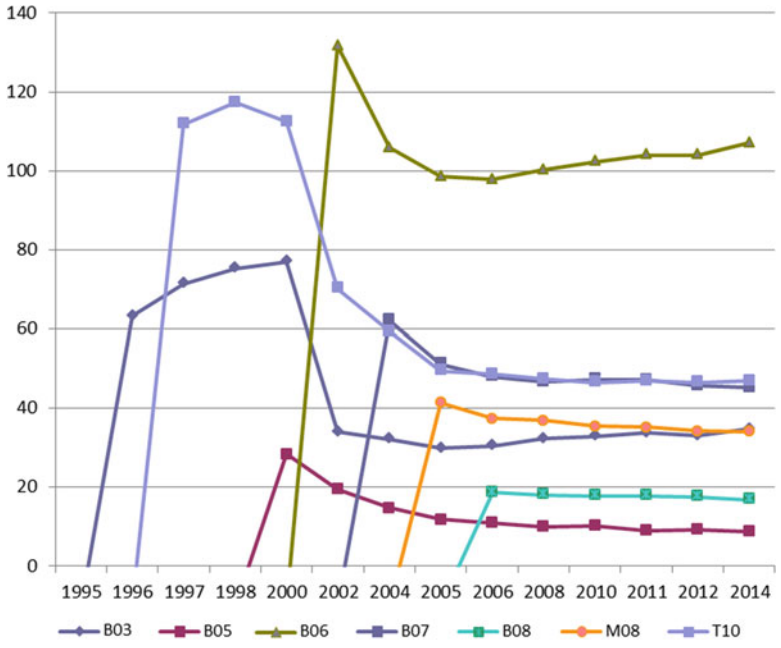



**Fig. 7.2** Time variation of the betweenness centrality (y-axis) associated with the most intermediary biodiversity-related COPs (from the time of their respective holding to 2014). In the COP labels the letter indicates the Convention (B = CBD, M = CMS, T = CITES) followed by the COP number

Almost all the curves obtained show an exponential decrease in the betweenness centrality of a COP, with the highest degree being achieved in the year of the conference. This is the result of an almost mechanical process: the terms used in the decisions or resolutions of a COP **A** are fixed once and for all after their publication (whereas the terms can be used more and more by successive COPs). If subsequent COPs use some of the terms that have provided strong betweenness to **A**, then the betweenness centrality of **A** will decline with time at a rate that decreases with the reduction of its centrality. Only the betweenness centrality of the CBD COP06, after a decrease, grows again from 2006 until today. This is due to the fact that this COP uses 12 (degree centrality) of the 13 concepts we follow, including at least one concept from each of the emerging COPs/terms mixed communities of Table 7.1.

We also note that the COPs with the highest betweenness centralities are linked to each of the three conventions (CBD, CMS, and CITES), which attests to the role played by each of these institutions in turn to bring and relay the themes in Health, particularly in response to global health crises or as a result of the work of ad hoc working groups (see above).

Finally, from Table 7.2 we can see that the community of COPs cC1 uses a high diversity of terms related to health in a strict sense with terms such as disease, health, and pathogens (cf. Lajaunie and Mazzega 2017; Fig. 7.1). The components

of that community are using terms linking many others and we could say that it is a generalist community. The community cC2 concerns more the issues of risk, security, and warning as well as biodiversity and more in relation with ecosystems or animal and plant having an effect on health. This community has a secondary link with health issues. The communities cC3 and cC4 are using groups of concepts that are distinct or dissimilar from the two other communities with a tendency to treat topics not addressed by the other communities.

As previously, in addition to the degree centrality (number of COPs having used a given term, see Fig. 7.2), the betweenness centrality of each term is another measure of its importance in taking into account the stakes of Health by the COPs. The temporal evolution of the degree centrality between the most central terms at present (2014–2016)—risk and threat, health, safety, biodiversity, diseases—is presented in Fig. 7.3. A term has a high degree centrality when it binds, via the COPs mentioning it, groups of terms that would otherwise tend to be separated. Since year 2000 the betweenness centrality of "health" is more important than that of the terms "security" or "biodiversity."

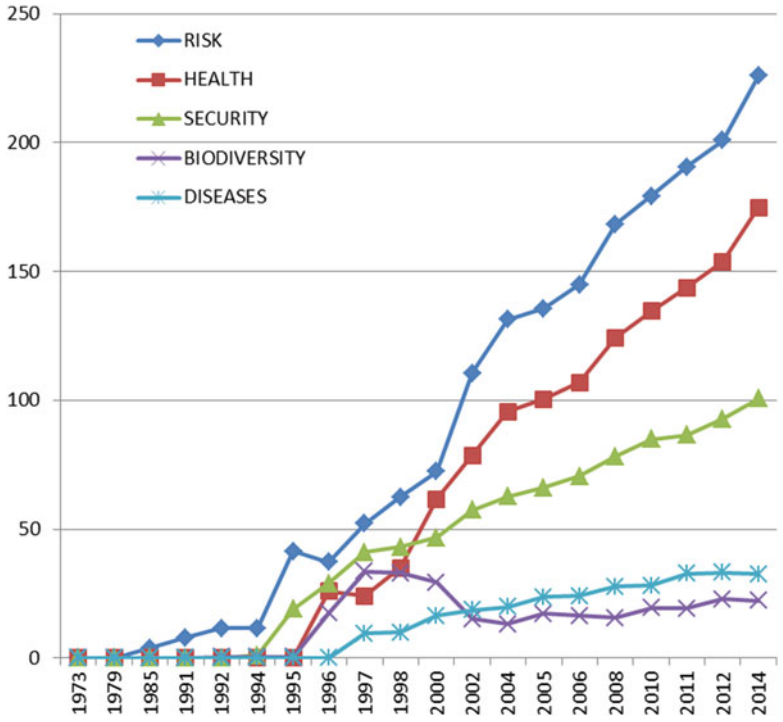



**Fig. 7.3** Time variation of the betweenness centrality (y-axis) associated with the most intermediary health–environment concepts (from 1973 to 2014)

## 7.6 Conclusion

On a legal science perspective, the dynamic of law and the way international environmental law integrates health issues in relation to other related themes constitutes a useful insight of the transformation of environmental law. Indeed, it is very interesting to analyze this evolution within the framework of complex systems, including the environment. We can work with a big amount of data (here the Conventions and decisions), thanks to the combination of methods such as text mining, graph theory, and multiplex analysis. The multiplex method enables a meticulous examination of terminology and COPs.

We focus on a limited text corpus but in the future we could broaden the study to the production of scientific groups attached to each convention whether they are scientific councils, specific working groups, or task-force to get a better comprehension of the issues at stake and the envisioned solutions or working avenues.

We have shown the interest of COP/term communities detection in the context of biodiversity-related conventions. One of the best ways to bring to light the emergence of "health–environment" issues on the 1973–2014 period as well as the legal instruments and the actors championing those issues is the multiplex analysis combined with text mining. On the one hand, it presents in a dynamic way the co-evolution of the biodiversity-related conventions (CBD: Convention on Biological Diversity; CMS: Convention on Migratory Species; and CITES: Convention on International Trade in Endangered Species of Wild Fauna and Flora) shown by the importance of some COPs in the distribution of terms or linking conventions. On the other hand, it highlights the use of specific terms in relation to communities of terms and thus reveals the issues at stake and their evolution over time. The main point to underline here is that the internal system constituted by the COPs has developed itself and does not correspond to an intended result nor it is the effect of a stated willingness of the different parties to the convention. The power is this method is to uncover the evolution of this system which has been constituted over time and in an autonomous manner, with a life on its own.

**Acknowledgments** This work is a contribution to a) Labex OT-Med (n° ANR-11-LABX-0061) and has received funding from Excellence Initiative of Aix-Marseille University—A*MIDEX, a French "Investissements d'Avenir" programme; b) the ANR Project FutureHealthSEA (n° ANR-17-CE35-0003-02) "Predictive scenarios of health in Southeast Asia: linking land use and climate changes to infectious diseases" (PIs: S. Morand and C. Lajaunie). The Ecology and Environment Institute of the National Center for Scientific Research (InEE CNRS, France) supports the International Multidisciplinary Thematic Network "Biodiversity, Health and Societies in Southeast Asia," Thailand (PI: S. Morand, CNRS/CIRAD) to which this study contributes.

## References


Bommarito, M. J., & Katz, D. M. (2010). A mathematical approach to the study of the United States code. *Physica A: Statistical Mechanics and its Applications, 389*, 4195–4200.

Boulet, R., Barros-Platiau, A. F., & Mazzega, P. (2016). 35 years of multilateral environmental agreements ratification: A network analysis. *Artificial Intelligence and Law, 24*, 133–148.

Boulet, R., Mazzega, P., & Bourcier, D. (2010). Network analysis of the French environmental code. In P. Casanovas, U. Pagallo, G. Sartor, & G. Ajani (Eds.), *AI approaches to the complexity of legal systems: Complex systems, the semantic web, ontologies, argumentation, and dialogue, LNAI* (Vol. 6237, pp. 39–53). Berlin: Springer.

Boulet, R., Mazzega, P., & Bourcier, D. (2011). A network approach to the French system of legal codes part I: Analysis of a dense network. *Artificial Intelligence and Law, 19*, 333–355.

Brown, V. A. (2007). Principles for EcoHealth action: Implications of the health synthesis paper, the millennium ecosystem assessment, and the millennium development goals. Workshop group, EcoHealth ONE, Madison, Wisconsin, October 2006. *EcoHealth, 4*(1), 95–98.

Durga, J., Sunitha, D., Narasimha, S. P., & Sunand, B. T. (2012). A survey on concept based mining model using various clustering techniques. *International Journal of Advanced Research in Computer Science and Software Engineering, 2*(4), 408–411 Avaiable at: www.ijarcsse.com.

Feldman, R., & Sanger, J. (2007). Core text mining operation. In R. Feldman & J. Sanger (Eds.), *The text mining handbook: Advanced approaches in analyzing unstructured data* (pp. 19–56). Cambridge: Cambridge University Press.

Fortunato, S. (2010). Community detection in graphs. *Physics Reports, 486*(3–5), 75–174.

Freeman, L. C. (1979). Centrality in social networks: Conceptual clarification. *Social Networks, 1*(3), 215–239.

Grüber, T. R. (1995). Toward principles for the design of ontologies used for knowledge sharing. *International Journal of Human-Computer Studies, 43*(5–6), 907–928.

Guarino, N., & Musen, M. (2015). Applied ontology: The next decade begins. *Applied Ontology, 10*, 1–4.

Hall, D., & Coghlan, B. (2011). Asia–Europe meeting. Implementation of the one health approach in Asia & Europe: How to set-up a common basis for action and exchange of experience. Preparatory study, European Union, p. 61.

Hassanpour, S., & Das, A. K. (2011). *Ontology-based text mining of concept definitions in biomedical literature*. CSWS2011 proceedings, 40–45. http://ceur-ws.org/Vol-774/das.pdf. Accessed 21 Dec 2016.

Kim, R. E. (2013). The emergent network structure of the multilateral environmental agreement system. *Global Environmental Change*. https://doi.org/10.1016/j.gloenvcha.2013.07.006.

Kuntz, P. (1992). Représentation euclidienne d'un graphe abstrait en vue de sa segmentation. Thèse de Doctorat, spécialité Mathématiques et Applications aux Sciences de l'Homme, École des Hautes Études en Sciences Sociales, Paris.

Lajaunie, C. (2016). The evolution of the link between health and environment in international law: The example of infectious diseases in South-East Asia. Accreditation to supervise research (HDR, in French). Presented on December 13th, 2016, University of Aix-Marseille, France, p. 246.

Lajaunie, C., & Mazzega, P. (2016). One health and biodiversity conventions. The emergence of health issues in biodiversity conventions. *IUCN Academy of Environmental Law eJournal*, (7), 105–121 http://www.iucnael.org/en/e-journal/current-issue.

Lajaunie, C. & Mazzega, P. (2017). Transmission, circulation et persistance des enjeux de santé dans les conventions internationales liées à la Biodiversité et Conventions de Rio. Confluence des Droits n° Special « Diffusion de normes et circulations d'acteurs dans la gouvernance internationale de l'environnement », S. Maljean-Dubois (Dir.), in press.

Lajaunie, C., Morand, S., & Binot, A. (2015). The link between health and biodiversity in Southeast Asia through the example of infectious diseases. *Environmental Justice, 8*(1), 26–32. https://doi.org/10.1089/env.2014.0017.

Morand, S., Dujardin, J. P., Lefait-Robin, R., & Apiwathnasorn, C. (2015). *Socio-ecological dimensions of infectious diseases in Southeast Asia* (Vol. IX, p. 338). Singapore: Springer.

Morgera, E. (2015). Fair and equitable benefit-sharing at the cross-roads of the human right to science and international biodiversity law. *Laws, 4*, 803–831. https://doi.org/10.3390/laws4040803.

von Luxburg, U. (2007). A tutorial on spectral clustering. *Statistics and Computing, 17*, 395–416.

Walther, B. A., Boëte, C., Binot, A., By, Y., Cappelle, J., Carrique-Mas, J. J., Chou, M., Furey, N., Kim, S., Lajaunie, C., Lek, S., Méral, P., Neang, M., Tan, B. H., Walton, C., & Morand, S. (2016). Biodiversity and health: Lessons and recommendations from an interdisciplinary conference to advise southeast Asian research, society and policy. *Infection, Genetics and Evolution, 40*, 29–46.

WCS-Wildlife Conservation Society (2004) The Manhattan principles, 2004.

WHO - CBD (2015) Connecting global priorities: biodiversity and human health: a state of knowledge review, p. 344.

Zinsstag, J. (2012). Convergence of EcoHealth and one health. *EcoHealth, 9*(4), 371–373.

Zinsstag, J., Schelling, E., Wyss, K., & Mahamat, M. B. (2005). Potential of co-operation between human and animal health to strengthen health systems. *Lancet, 366*, 2142–2145. https://doi.org/10.1016/S0140-6736(05)67731-8.

# Chapter 8
# How Does Linguistic Complexity in Shakespeare's Plays Relate to the Production History of a Commercial American Theater?


**Brian Kokensparger**

## 8.1 Introduction

Most of us remember reading at least one of Shakespeare's plays in high school, most likely with fondness (perhaps some less so). Though I had been an English speaker from my toddler days, as a high school student I found myself intimidated by the language, and also found it difficult to work through the play, to figure out what was going on in the story. The language got in the way of the experience. Yet I felt that the effort was worth the gain: a whole new world of literature was opened to me.

In my first day as an undergraduate student at Creighton University, I was participating in acting exercises during an Oral Interpretation of Literature course. A faculty member was staging *The Taming of the Shrew*, and saw me going through my paces. He suggested that I audition for a role. Again, I was intimidated by the language of the play. But fighting the urge to run far, far away, I auditioned and received a major role. I was thrilled, and though the language of the play baffled me as an actor at times, it still provided a rich world of linguistic pleasure that I have never quite gotten out of my system.

It is this push and pull—the richness and beauty of the language versus its complexity and archaic-ness, that greets most Americans who interact in some way with Shakespearean drama. Production companies that stage these plays must try to enhance their endearing qualities—their richness and beauty—and lessen their challenging aspects—by-products of their linguistic complexity. They must therefore make careful choices about which plays they stage, and how they stage them.

B. Kokensparger (✉)
Journalism, Media & Computing Department, Creighton University, Omaha, NE, USA
e-mail: bkoken@creighton.edu



S.-H. Chen (ed.), *Big Data in Computational Social Science and Humanities*,
Computational Social Sciences, https://doi.org/10.1007/978-3-319-95465-3_8

In this chapter, I have attempted to capture these choices and to quantify them to determine if there is a relationship between the linguistic complexity of the plays themselves, and which among them get chosen more frequently for production. To begin to look at this relationship, I first had to consider why linguistic complexity matters in drama in the first place.

## 8.2 Theoretical Framework

Sweller's (1990) and Sweller and Chandler's (1994) Cognitive Load Theory has been accepted in the field of education since the early 1990s, as a way to examine the learning process and to inform course delivery methods. In brief, human beings perceive input through their senses (all senses, not just hearing and seeing), and process that input within the learning context. What each human being already knows about the subject is represented in the phenomenon of "chunking," where connections between various facts and experiences have been made and affirmed through varied and repeated encounters with specific knowledge. Cognitive load refers to the amount of energy (exerted through the construct of "attention") required for the learner to process the new input; too much and the learner gets overwhelmed and focuses her attention elsewhere. Too little and the learner focuses her attention elsewhere as well (due mostly to boredom). There are other nuanced factors that affect cognitive load, some of them learner-specific (such as intrinsic and extrinsic motivation and health factors) and others environmental in source (such as room temperature and noisy hallways). For that reason, the educational field has paid particular attention in the past few decades to varying the modalities of learning and employing other learning methods that allow users to adapt their cognitive loads to their own learning needs.

As human beings encounter more learning situations each day outside the traditional classroom environment, Sweller's Cognitive Load Theory can also be applied to these areas, including on-the-job-training situations, and even arts events. A theatrical production mimics the classroom in a number of ways: the signals are applied in a controlled setting, in multiple modalities (auditory and visual at a minimum), and rely upon the audience member's previous experiences with the subject area, characters and situation to draw meaning from the production. As is true in a classroom presentation, cognitive load must be controlled in a theatrical production for the audience members to remain engaged with the production and not focus their attentions elsewhere. Though there are many ways to attempt to measure cognitive load on an audience member, one way is through quantifying linguistic complexity.

## 8.3 Methodology

Linguistic complexity can be defined as the intrinsic complexity of the language and grammar choices made in composing the written text of a manuscript. Grammatical choices, common diction (word choice), sentence length, paragraph cohesion and transition all affect linguistic complexity in the written word. Students often characterize linguistic complexity as how "hard" a text is to read. Linguistic complexity in the spoken word is similar to that in the written word, with some additional factors, such as the importance of vocal inflection, tonal variation, and rhythm of delivery. Though these factors also occur, to some extent, in the reader's mind as she is reading written text, they are much more pronounced in an oral-only delivery.

There are several ways to measure linguistic complexity (Hendrikse and van Zweel 2010; Juola 1998; McKee et al. 2000; Vulanovic 2007; Warren and Gibson 2002), but here is a simple way. First, count all of the syllables in the text and divide that value by the total number of words. That gives an average syllables per word value. As syllabification affects cognition (the more syllables in a word, the more attention required by the listener to wait to process the entire word), texts with a higher syllables-per-word average could be considered more linguistically complex. Next, count the average words per sentence in a text. Since a sentence is an idea extended over a number of words, the average word count per sentence is a good way to continue quantifying linguistic complexity. There is a difference between determining the average number of syllables per word in the text, and the percentage of "complex words," those words with three or more syllables. This third dimension provides a value for the percentage of complex words in the text. Finally, cognition requires immediate recognition of specific words in a text. An unknown word will require the listener to try to figure out the meaning of that word based upon its context, which requires additional cognitive attention. So a fourth dimension in our simple linguistic complexity measure is the percentage of words that are not in a dictionary common to the listener.

It must be noted here that these measurements were constructed solely by me, based on formulae used for reading level analysis, and the instrument is not validated through any linguistic validation methods. However, as the dimensions are used for comparative purposes over the entire corpus of Shakespeare's plays, any small programming errors or omissions affect all of the measurements equally. In a comparative analysis, these imperfections are minimized since they are applied over the entire corpus. It would be interesting, as further study, to see if other linguistic complexity measurement systems also provide the same results.

## 8.4 Findings

I applied these linguistic complexity dimensions to the plays attributed to Shakespeare. Of course, there is considerable discussion among Early Modern English Literature scholars as to which plays should be attributed to Shakespeare, but the approach adopted by the Folger Shakespeare Library (2017) identifies 38 plays as being either wholly recognized as attributed to Shakespeare, or in large part attributed to him. From hereto forward in this text, when I refer to Shakespeare's plays, I mean those plays commonly attributed to Shakespeare.

### *8.4.1 Linguistic Complexity in Shakespeare's Plays*

Cook (2006) has researched the connection between Shakespeare and cognition. Building upon this work, I applied my simple linguistic complexity analysis to the entire corpus of spoken words in Shakespeare's plays. I decided to use the version of the plays provided on the Folger Digital Texts website (2017). There are other versions of the plays available, but the ones on the Folger Digital Texts website are particularly edited for the American English reader, and are commonly used by production companies in the USA to stage the plays. I also decided to focus in particular on the words that the audiences were meant to hear. So stage directions, character names, act, and scene headings were all left out of the analyzed text. Here are the results of the analysis, ordered from the least linguistically complex of Shakespeare's plays to the most linguistically complex (see Table 8.1).

The top three least linguistically complex plays include a tie for first place between *The Comedy of Errors* and *Romeo and Juliet*, followed by *Much Ado About Nothing*. The bottom three (i.e., the three most linguistically complex of Shakespeare's plays) also includes a tie (between *Henry V* and *Henry VI, Part 1*), followed by *Titus Andronicus*.

By genre, it is no surprise that the comedies rank highly among the least linguistically complex plays, holding 12 of the top 15 places. However, the three most linguistically complex among the comedies include *Troilus and Cressida*, *Love's Labor's Lost*, and *Measure for Measure*.

The histories rank among the most linguistically complex plays, holding six of the eight plays among the highest linguistic complexity. The least linguistically complex of the history plays falls to *Henry IV, Part 2*, which is followed a few places down (thus higher in complexity) by its counterpart, *Henry IV, Part 1*. *Henry V* reigns as the most linguistically complex of Shakespeare's plays, as measured through this simple system.

The tragedies are distributed relatively evenly through the field in terms of linguistic complexity, though those based on classical themes (*Julius Caesar*, *Antony and Cleopatra*, and *Titus Andronicus*) settle to the bottom of the list (and are therefore, some of the more highly linguistically complex plays) while *Romeo*

**Table 8.1** Shakespeare's plays ranked by linguistic complexity summed rankings

| Syl/W | SWRnk | CW | WPS | WPSRnk | CWpW | CWRnk | NinDpW | NinDRnk | TotScore | Play/ Category |
|---|---|---|---|---|---|---|---|---|---|---|
| 1.258 | 1 | 686 | 11.65 | 18 | 0.048 | 1 | 0.019 | 3 | 23 | Err |
| 1.269 | 2 | 1182 | 10.56 | 8 | 0.049 | 3 | 0.023 | 10 | 23 | Rom |
| 1.279 | 6 | 1130 | 11.56 | 17 | 0.054 | 7 | 0.018 | 1 | 31 | Ado |
| 1.276 | 4 | 1013 | 9.17 | 1 | 0.048 | 1 | 0.030 | 26 | 32 | Wiv |
| 1.278 | 5 | 1118 | 10.21 | 6 | 0.057 | 13 | 0.023 | 10 | 34 | TN |
| 1.288 | 16 | 1439 | 9.89 | 5 | 0.057 | 13 | 0.021 | 6 | 40 | Lr |
| 1.283 | 7 | 845 | 10.87 | 13 | 0.051 | 5 | 0.026 | 17 | 42 | Mac |
| 1.284 | 9 | 896 | 10.83 | 12 | 0.055 | 8 | 0.024 | 14 | 43 | Tmp |
| 1.270 | 3 | 1189 | 12.82 | 27 | 0.051 | 5 | 0.022 | 8 | 43 | TNK |
| 1.284 | 9 | 963 | 11.36 | 15 | 0.057 | 13 | 0.022 | 8 | 45 | TGV |
| 1.283 | 7 | 1209 | 12.09 | 23 | 0.056 | 11 | 0.020 | 5 | 46 | AYL |
| 1.287 | 12 | 1003 | 12.54 | 26 | 0.056 | 11 | 0.023 | 10 | 59 | Per |
| 1.284 | 9 | 1118 | 10.83 | 11 | 0.055 | 8 | 0.039 | 33 | 61 | Shr |
| 1.294 | 24 | 1088 | 9.87 | 4 | 0.061 | 23 | 0.026 | 17 | 68 | Tim |
| 1.289 | 17 | 1398 | 11.84 | 20 | 0.062 | 29 | 0.019 | 3 | 69 | AWW |
| 1.293 | 22 | 1430 | 12.43 | 24 | 0.055 | 8 | 0.025 | 16 | 70 | 2H4 |
| 1.287 | 12 | 1542 | 12.95 | 29 | 0.062 | 29 | 0.018 | 1 | 71 | WT |
| 1.287 | 12 | 1610 | 11.99 | 22 | 0.060 | 19 | 0.027 | 19 | 72 | Cym |
| 1.291 | 21 | 1795 | 11.98 | 21 | 0.060 | 19 | 0.024 | 14 | 75 | Ham |
| 1.295 | 25 | 1639 | 9.54 | 2 | 0.063 | 31 | 0.027 | 19 | 77 | Oth |
| 1.287 | 12 | 1222 | 13.24 | 30 | 0.058 | 18 | 0.027 | 19 | 79 | MV |
| 1.289 | 17 | 1373 | 12.84 | 28 | 0.057 | 13 | 0.030 | 26 | 84 | 1H4 |
| 1.293 | 22 | 1636 | 11.75 | 19 | 0.061 | 23 | 0.028 | 22 | 86 | Cor |
| 1.290 | 19 | 979 | 11.54 | 16 | 0.061 | 23 | 0.035 | 29 | 87 | MND |
| 1.290 | 19 | 1152 | 14.1 | 34 | 0.049 | 3 | 0.038 | 31 | 87 | 3H6 |
| 1.300 | 28 | 1431 | 13.5 | 32 | 0.061 | 23 | 0.021 | 6 | 89 | H8 |
| 1.305 | 32 | 1426 | 11.28 | 14 | 0.067 | 37 | 0.023 | 10 | 93 | MM |
| 1.305 | 32 | 1312 | 10.29 | 7 | 0.063 | 31 | 0.029 | 24 | 94 | LLL |
| 1.298 | 26 | 1168 | 10.75 | 10 | 0.061 | 23 | 0.041 | 37 | 96 | JC |
| 1.300 | 28 | 1532 | 9.83 | 3 | 0.064 | 35 | 0.038 | 31 | 97 | Ant |
| 1.299 | 27 | 1170 | 14.42 | 36 | 0.057 | 13 | 0.028 | 22 | 98 | Jn |
| 1.307 | 34 | 1609 | 10.62 | 9 | 0.063 | 31 | 0.037 | 30 | 104 | Tro |
| 1.310 | 36 | 1703 | 12.52 | 25 | 0.060 | 19 | 0.031 | 28 | 108 | R3 |
| 1.303 | 30 | 1477 | 13.42 | 31 | 0.060 | 19 | 0.039 | 33 | 113 | 2H6 |
| 1.304 | 31 | 1337 | 14.73 | 37 | 0.061 | 23 | 0.029 | 24 | 115 | R2 |
| 1.308 | 35 | 1266 | 13.56 | 33 | 0.064 | 35 | 0.039 | 33 | 136 | Titus |
| 1.348 | 38 | 1560 | 14.15 | 35 | 0.076 | 38 | 0.039 | 33 | 144 | 1H6 |
| 1.313 | 37 | 1630 | 15.34 | 38 | 0.063 | 31 | 0.046 | 38 | 144 | H5 |

*and Juliet*, *King Lear*, and *Macbeth* are lowest in linguistic complexity. *Hamlet*, *Coriolanus*, and *Othello* are mid-range tragedies in terms of linguistic complexity. As an offshoot of this study, these data suggest that a case could be made to re-categorize the Shakespearean tragedies into those which are more like history plays, those which are more like comedies, and the true tragedies that hang in the linguistic complexity mid-range.

### 8.4.2 Production History of an American Commercial Shakespearean Theater

The ranking of Shakespeare's plays is interesting as its own exercise, but not particularly helpful in understanding how linguistic complexity affects the listener's engagement and reaction to the script as it is performed before a live audience. One of the major drawbacks of Sweller's Cognitive Load Theory is that it is difficult to measure learning directly—one would need PET scans and other medical equipment to operate during a classroom learning session to visualize results and produce data. Thus it is very difficult to replicate the classroom environment while performing the study, and by extension, very difficult to replicate a Shakespearean performance in a hospital lab setting.

After gathering linguistic complexity data for all of Shakespeare's plays, I wondered if these data have any relationship to success markers, or at least popularity markers, of American theater companies.

One way to measure theatrical popularity is to examine the production history of a commercial theater, where paying customers comprise the largest part of its revenue. As such, the patrons vote with their wallets, casting votes by buying tickets. In Shakespeare's England, of course, the theater-goers, either those in the boxes and seats or the groundlings, had a trained ear for verse and were spoken to in the vernacular. They were also privy to the scripts' inside jokes due to general knowledge about the historical references. So the pre-existing "chunks" in the heads of those theater-goers were quite different than those likely found in the heads of an American audience. The American theater-goer may or may not have the benefit of academic exposure to Shakespearean and Early Modern drama and may or may not have attended a multitude of performances of these plays. It can be assumed, though, that, patron-by-patron, odds are that the American theater-goer would consider this ground relatively new territory, and would therefore be much more prone to cognitive overload due to linguistic complexity.

I chose a specific theater that advertises a close attention to Shakespeare's original text, but is still a commercial theater of many years running. This is important, as every play is produced by using a "director's cutting" of the script, and therefore, an uneven application of the play could make comparison difficult. With the theater's stated close adherence to the script, though some of these variations obviously exist, in this context they are minimalized. It is also difficult to find a theater that does not accept grant money and donations to cover part of its costs. Therefore, finding a totally commercial theater that makes all of its revenue from ticket prices is impossible: there is no such beast, or at least none that rises above the level of coffee house amateurism. There is nothing wrong with this amateurism, by the way. Some of the most creative and courageous productions arise out of this phenomenon, but that's not quite the kind of company we are looking for here. The chosen theater remains unidentified in this text because I wanted a bit of objective distance from the theater, so I could take its production data at face

value, without any explanations or any other input that may have occurred had I developed a scholarly relationship with the theater staff.

Of course other factors play an important role in terms of which plays get produced, like thematic plays offered during holidays (e.g., *Romeo and Juliet* near Valentine's Day, *Macbeth* near Halloween, *A Midsummer Night's Dream* as a midsummer offering, etc.). But why *Romeo and Juliet* for Valentine's Day? Knowing how the plot turns out, if a couple in the audience stays until the end, they may not be in such a romantic mood when they head off to dinner afterwards. For Valentine's Day, why not *All's Well That Ends Well*, which lives up to its name in terms of a romantic evening of theater, or any other of the many Shakespearean plays that have a love element to them? Perhaps linguistic complexity provides the missing link here, the missing explanation.

I collected the production history of several American commercial theaters. Free summer Shakespeare festivals and educational theaters in colleges and universities generally must receive a large amount of their budgets from corporate sponsors and public grants, and in giving back, their choice of plays to stage often reflects missions tied to education and public service. Commercial theaters' missions are normally centered mostly around selling tickets to stay alive in the arts scene. Therefore, one theater was chosen, and its entire production history was aggregated in terms of the number of times it produced a specific Shakespeare play. Each play's frequency was then ranked among the entire corpus. The production data were available for this theater from the early 1990s through the end of the data collection period in May, 2016. Ranking the produced plays by production frequency, there were few surprises in terms of its most and least produced plays (see Table 8.2).

*Romeo and Juliet* topped the production frequency at 19, followed by *A Midsummer Night's Dream* (14) and *Macbeth* (13). *The Taming of the Shrew* and *Much Ado about Nothing* followed with 11 and 10, respectively. At the bottom, again not surprisingly, were the history plays; all three of the *Henry VI* plays, *Henry IV, Part 2*, *Henry VIII*, and *King John* tied in their ranking at the bottom with only 1 production each.

### *8.4.3 Comparison of Linguistic Complexity and Production History*

With these results, we can now compare the production frequency rankings with the linguistic complexity rankings. Since my assertion is that the least linguistically complex plays would be the most popular for a commercial theater to produce, the number 1 ranked play in production frequency was the one that had the highest number of productions in the chosen theater, and the number 1 ranked play in linguistic complexity was the one with the lowest sum of rankings among the linguistic complexity dimensions, i.e., the lowest linguistic complexity.

**Table 8.2** Shakespeare plays ranked by production frequency of the American commercial Shakespearean theater

| Abbrev | Frequency |
|---|---|
| Rom | 19 |
| MND | 14 |
| Mac | 13 |
| Shr | 11 |
| Ado | 10 |
| TN | 9 |
| AYL | 7 |
| Tmp | 7 |
| Wiv | 7 |
| Err | 6 |
| Ham | 6 |
| MV | 6 |
| JC | 5 |
| LR | 4 |
| TGV | 4 |
| Ant | 3 |
| AWW | 3 |
| Cym | 3 |
| Oth | 3 |
| Per | 3 |
| 1H4 | 2 |
| Cor | 2 |
| H5 | 2 |
| LLL | 2 |
| MM | 2 |
| R2 | 2 |
| R3 | 2 |
| Tim | 2 |
| Tit | 2 |
| TNK | 2 |
| Tro | 2 |
| WT | 2 |
| 1H6 | 1 |
| 2H4 | 1 |
| 2H6 | 1 |
| 3H6 | 1 |
| H8 | 1 |
| JN | 1 |

Ranking Shakespeare's plays by how far apart they were in respect to their linguistic complexity and production frequency rankings, there was a definite pattern (see Table 8.3).

**Table 8.3** Ranking differentials between linguistic complexity and production frequency

| Play/Category | LCRank | ProdRank | Diff |
|---|---|---|---|
| Rom | 1 | 1 | 0 |
| TN | 5 | 6 | 1 |
| Tmp | 8 | 7 | 1 |
| AWW | 15 | 16 | 1 |
| 1H4 | 22 | 21 | 1 |
| 2H6 | 34 | 33 | 1 |
| Ado | 3 | 5 | 2 |
| Cym | 18 | 16 | 2 |
| Cor | 23 | 21 | 2 |
| Jn | 31 | 33 | 2 |
| Wiv | 4 | 7 | 3 |
| Mac | 7 | 3 | 4 |
| TGV | 10 | 14 | 4 |
| AYL | 11 | 7 | 4 |
| Per | 12 | 16 | 4 |
| WT | 17 | 21 | 4 |
| Oth | 20 | 16 | 4 |
| 1H6 | 37 | 33 | 4 |
| MM | 27 | 21 | 6 |
| Tim | 14 | 21 | 7 |
| H8 | 26 | 33 | 7 |
| LLL | 28 | 21 | 7 |
| Lr | 6 | 14 | 8 |
| Err | 1 | 10 | 9 |
| Shr | 13 | 4 | 9 |
| Ham | 19 | 10 | 9 |
| 3H6 | 24 | 33 | 9 |
| MV | 21 | 10 | 11 |
| Tro | 32 | 21 | 11 |
| R3 | 33 | 21 | 12 |
| TNK | 8 | 21 | 13 |
| Ant | 30 | 16 | 14 |
| R2 | 35 | 21 | 14 |
| Titus | 36 | 21 | 15 |
| JC | 29 | 13 | 16 |
| H5 | 37 | 21 | 16 |
| 2H4 | 16 | 33 | 17 |
| MND | 24 | 2 | 22 |

In analyzing these results, we see that in a majority of the plays there is generally very little ranking difference between low linguistic complexity and high number of performances. *Romeo and Juliet* ended up ranked first in both categories. In fact, among the ten plays with the lowest differentials, the highest differential value was 2. This list of ten includes five comedies, two tragedies, and three history plays, and is consistent between some of the least linguistically complex plays (such as *Romeo and Juliet*) and some of the most linguistically complex plays (such as *King John*).

Looking at the plays simply by linguistic complexity ranking, among the ten least linguistically complex plays, six also appeared in the top ten list of most performed plays. And also looking at the plays simply by production frequency ranking, all but 1 (*The Two Noble Kinsmen*) of the ten least linguistically complex plays appeared among the top fifteen most performed plays.

The remainder of the list posted generally close differentials between rankings, except for the largest outlier, *A Midsummer Night's Dream*, which was on the higher linguistic complexity side yet was the second most performed play in our study of this commercial theater. There are possible explanations for this anomaly, most notably that for the past 14 years that theater ran a special performance on or around Midsummer Night's Eve. Which other of Shakespeare's plays does one normally think of when given a theme of summer?

## 8.5 Results and Conclusions

Otherwise, there appears to be a strong relationship between a play's measured linguistic complexity and its production history: the lower a play's linguistic complexity, the higher the number of performances that it generally has. From these results it appears that plays with lower linguistic complexity were more likely to be produced in this American commercial theater than plays of higher linguistic complexity.

These results suggest that Cognitive Learning Theory also plays a role in arts and entertainment, where the complexity of the text of a play that commercial theaters choose to produce directly relates to ticket sales, and perhaps the ultimate survival of the theaters themselves.

## References

Cook, A. (2006). Staging nothing: Hamlet and cognitive science. *Substance: A Review of Theory & Literary Criticism, 35*(2), 83–99.

Folger Shakespeare Library (2017). *Folger digital texts*. Retrieved from http://www.folgerdigitaltexts.org/. Accessed 7 Mar 2017.

Hendrikse, R., & van Zweel, H. (2010). A phylogenetic and cognitive perspective on linguistic complexity. *Southern African Linguistics & Applied Language Studies, 28*(4), 409–422.

Juola, P. (1998). Measuring linguistic complexity: The morphological tier. *Journal of Quantitative Linguistics, 5*(3), 206–213.

McKee, G., Malvern, D., & Richards, B. (2000). Measuring vocabulary diversity using dedicated software. *Literary and Linguistic Computing: Journal of the Association for Literary and Linguistic Computing, 15*(3), 323–337.

Sweller, J. (1990). Cognitive processes and instruction procedures. *Australian Journal of Education, 34*(2), 125–130.

Sweller, J., & Chandler, P. (1994). Why some material is difficult to learn. *Cognition and Instruction, 12*(3), 185–233.

Vulanovic, R. (2007). On measuring language complexity as relative to the conveyed linguistic information. *SKY Journal of Linguistics, 20*, 399–427.

Warren, T., & Gibson, E. (2002). The influence of referential processing on sentence complexity. *Cognition, 85*(1), 79–112.

# Chapter 9
# Language Communities, Corpora, and Cognition


**Huei-Ling Lai, Kawai Chui, Wen-Hui Sah, Siaw-Fong Chung, and Chao-Lin Liu**

## 9.1 Introduction

"Language as a complex adaptive system has a fundamentally social function. Processes of human interaction along with domain-general cognitive processes shape the structure and knowledge of language" (Beckner et al. 2009, pp. 1–3). Language data, whether spoken, written, or gestural, can be digitized for analyzing and computing patterns of linguistic form, meaning, and use shaped and reshaped in the interactions of the users in social–cultural contexts in homogeneous or heterogeneous language communities. The investigation of linguistic patterns rests on language corpora that provide rich data produced by different communities of speakers at various points of time. The corpus analysis of contemporary and historical data, together with cross-linguistic comparison, and psycholinguistic and neurolinguistic experimentation, is crucial to understand the linguistic behaviors of a community and the neurocognition of the language users. Further analyses would shed light on the social–cultural dynamics related to the temporal and spatial dimensions.

Acknowledging the situated, dynamic, and emergent nature of language properties, this chapter introduces digitized language data from a variety of corpora established for different language communities, including adults, children, atypical children, dominant language users, and minority language users. The language data, whether in oral, written, or gestural modes, are valuable and indispensable to conduct research on linguistic behaviors involving form, meaning,and use through

H.-L. Lai (✉) · K. Chui · W.-H. Sah · S.-F. Chung
Department of English, National Chengchi University, Taipei, Taiwan

C.-L. Liu
Department of Computer Science, National Chengchi University, Taipei, Taiwan



S.-H. Chen (ed.), *Big Data in Computational Social Science and Humanities*, Computational Social Sciences, https://doi.org/10.1007/978-3-319-95465-3_9

which to understand the linguistic cognition of different language communities. In Sect. 9.2, the rationale and tenets of the corpus-driven paradigm are introduced as both qualitative and quantitative linguistic analyses require rigorous methods for corpus analysis and data analytics. In Sects. 9.3, 9.4, and 9.5, three areas of studies of linguistic patterns and cognition based on corpus data across language communities are presented. The research issues addressed by the studies are: What kinds of corpora data are employed in language studies? How do the corpus data manifest the recurrent patterns of linguistic form, meaning, and use in various social–cultural contexts? How do the linguistic patterns, whether oral, written, or gestural, manifest the linguistic cognition of a language community?

## 9.2 Corpus-Driven Methods: The Tenets of the Paradigm

Since the development of corpus linguistics, scholars have seemed to understand the term corpus-based approach as a general term to refer to all corpus-related studies. Yet, in 2001, Tognini-Bonelli pointed out that a corpus-based approach should be distinguished from a corpus-driven one: the former refers to using the corpus as a database of exemplars, and the latter refers to data-generated results taken from a corpus. However, the distinction between the two approaches is never clear-cut and no consensus on the matter of the name has been reached by all corpus users. Because corpus linguistic research works in-line with either existing programs or scripts written by program writers, it has a strong connection with computational linguistics. Furthermore, as the term *big-data* has appeared in recent years, the corpus-driven approach seems to relate strongly to big-data methodology in linguistics. For instance, there are heavily data-based corpus studies which still prefer to use the term *corpus-based*, while *corpus-driven* has been applied to the minimum use of data. Biber (2010, p. 202) also commented on corpus-based and corpus-driven approaches in The Oxford Handbook of Linguistic Analysis, saying that "corpus-driven methodologies can differ from one study to the next in three key respects:"

- The extent to which they are based on analysis of lemmas vs. each word form;
- The extent to which they are based on previously defined linguistic constructs (e.g., part-of-speech categories and syntactic structures) vs. simple sequences of words;
- The role of frequency evidence in the analysis.

Biber is of the view that a pure corpus-driven study will be likely to treat words as word forms rather than as lemmas. However, as mentioned, extremely strict criteria may still not solve the problem of interchangeable use of the terms by scholars. No matter what the distinction is, both the corpus-based and the corpus-driven approaches share the following aims: First, they both aim to investigate a certain linguistic phenomenon, be it spoken, written, or even gestural, used by one or more communities; second, primacy is accorded to the investigation of recurrent patterns of linguistic data; last, they both aim to examine a certain linguistic theory by observing the linguistic form, meaning, and use.

In general, in language studies, corpora could be of various types to investigate the linguistic behaviors of a community, including:

- Monolingual corpora which are corpora reflecting the usage among language communities that could be measured in the millions (e.g., British National Corpus, The Academia Sinica Balanced Corpus of Mandarin Chinese, etc.), or as small as a language variety (e.g., the International Corpus of English (ICE) is a project that collects corpora of World Englishes).[1]
- Multilingual corpora that could exist in various forms such as parallel (with line-by-line translation in alignment) comparisons (with corpora of two or more languages designed under the same parameters), etc.

Various corpora could serve different purposes for analyzing and computing patterns of linguistic form, meaning, and use. Whether the approach used is corpus-based or corpus-driven, the aim is to inspect linguistic patterns produced by different communities of speakers at various points of time. In metaphor studies, for instance, various types and stages of metaphors can be detected by inputting specific keywords in corpora databases, and, through analyzing the linguistic environment of these keywords and their collocations, meaningful interpretations could then be produced.

## 9.3 Recontextualization of Metaphor from Language Use

### *9.3.1 A Brief Summary of Metaphor Development*

Metaphor has been demonstrated to be pervasive through a wealth of linguistic data, from daily linguistic usage to fixed expressions such as idioms, proverbs, and the like (cf. Lakoff 1993; Radden and Kövecses 1999). The classical approach to metaphor considers it as a poetic and rhetorical device, purely a feature of language alone. In the 1980s, the publication of *Metaphors We Live By* (Lakoff and Johnson 1980) shifted the focus to the cognitive force of metaphor, and inspired a great volume of studies regarding metaphorical thought and language. However, three issues have been raised as difficulties and questions arise when empirical discourse data are examined (cf. Cameron and Deignan 2006; Deignan 2005). First, linguistic metaphors are subject to grammatical and lexical restrictions. Second, mapping gaps and highly specific metaphorical meanings exist when corpus data are examined in detail. Third, metaphorical mappings are highly contingent on cultural differences. These concerns bring to the fore the importance of discourse situation and social–cultural context in the realm of metaphor studies (cf. Deignan 2005, 2008, 2012; Geeraerts 2010; Kövecses 2002; Yu 2008). In the last decade, one of the trends in research on metaphor has been discourse and corpus approach that takes the

[1] http://www.ucl.ac.uk/english-usage/projects/ice.htm

emergent perspective of metaphor, addressing "the effects of metaphors in discourse and the influence of context in their meanings and functions" (Porto and Romano 2013, p. 60). Such an approach considers discourse as deriving from the interaction of complex dynamic systems that consist of interrelated elements which are under constant change as discourse unfolds (Cameron and Deignan 2006). In other words, metaphorical use, in addition to ideational meanings, carries attitude and evaluation.

Corpus-based studies of metaphor have helped compare and contrast the use of metaphor in different text genres. For instance, Boers (1999), inspecting all of the instances of the use of HEALTH metaphors in *The Economist* over a 10-year period, finds that HEALTH metaphors are more frequently used in articles written during the winter months. Boers maintains that the winter months are the time when matters related to physical health are salient topics, leading to their frequent use as a metaphor. Researchers who identify the dynamic features of metaphor in different contexts of use have found that metaphorical expressions can be used and reused by different language users in different discourse situations as "the density, forms and functions of metaphors in language can vary substantially depending on context of use, or more specifically on genre and register" (Semino et al. 2013, p. 41). To illustrate, Porto and Romano (2013) analyzed two metaphors in newspapers—*green shoots* and *ash cloud*—showing how the two expressions are deliberately reused in contexts different from where they had first come from (i.e., economy and geology). In sum, corpus evidence shows that linguistic metaphors are constrained by their co-text; the topic and purpose of particular text genres influence metaphor choice; and different languages with different cultural factors can affect different choices of metaphor to describe the same topic.

### *9.3.2 A Case of Metaphor Recontextualization in Taiwan Hakka*

Based on the corpus data from the online data news of four major newspapers in Taiwan: Udndata, Knowledge Management Winner, Liberty Times Net, and Apple Daily, a lexicalized item <*nganggiang* stiff neck 硬頸> with a metonymy-based metaphor is investigated in Lai (2017). The word indicates a bodily oriented metaphor as it is used to characterize a person's character as being stubborn and tough by describing his/her bodily experience—making the neck stiff to show an unyielding attitude. This linguistic code gives rise to an image that is identifiable, understandable, empathetic, and easy to make sense of. Its accessibility and uniqueness meet the needs of language users for communication and for linguistic innovation. The study accounts for how, through frequent usage in media discourse, this metaphorical expression has become entrenched and conventionalized, by carrying a certain linguistic, pragmatic, and attitudinal meaning. News items with the key word were collected. In total, 2646 items of news with 2,045,263 characters were retrieved, and 3357 tokens were found for analysis. Using digital tools for computation and calculation, the data were analyzed regarding their longitudinal (Fig. 9.1) and theme distributions (Fig. 9.2).

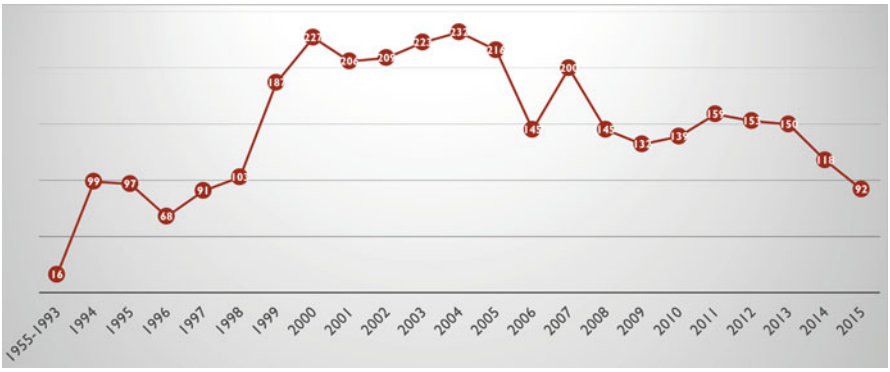



**Fig. 9.1** Yearly distribution

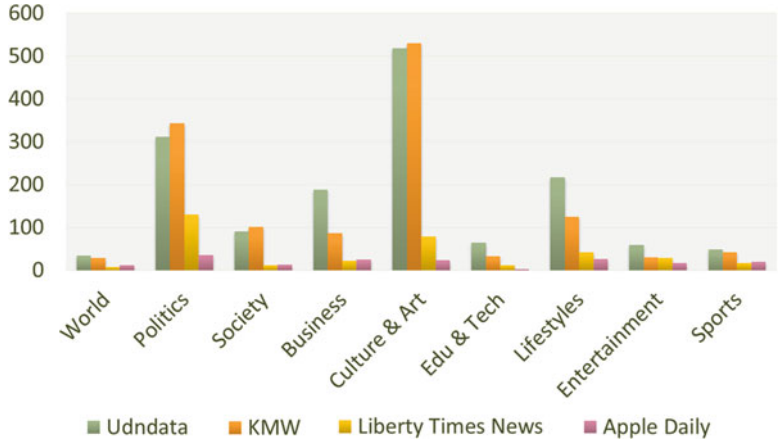



**Fig. 9.2** Topic distribution

The main findings are as follows. The high peaks in Fig. 9.1 indicate that the frequency of the usage closely correlates with the years when important political events happened in Taiwan society. While originally used to carry a negative connotation, it now carries a positive connotation often used to characterize Hakka-related matters, greatly showing an ideological effect toward Hakka ethnic groups in media discourse. In addition, through mechanisms of denotational extension, metonymy, and metaphor, its usage has repeatedly increased. It has become a handy attractor exploited by the media—portraying matters from prototypical Hakka ones to non-Hakka ones, and representing various themes from international affairs, publicity for entertainment or sports, to technology and finance matters. In short, it emerges as an iconographic reference in the news discourse serving as a simplistic image with judgment and values. The following headlines taken from Udndata illustrate how it is used to portray the protagonist's perseverance or indomitability: in (1) of Angelina Jolie, in (2) of the New York Knicks, and in (3) of objects such as stocks.

1. 裘莉硬頸挺難民.

   Qiúlì yìngjǐng tǐng nànmín.
   Angelina Jolie-stiff neck-back up-refugees.
   “Jolie strongly backs refugees up.”

2. 傷兵不怕戰尼克超硬頸.

   Shāngbīng búpà zhàn níkè chāoyìngjǐng.
   wounded players-no-fear-fighting-Knicks extremely stiff neck.
   “The wounded players do not fear; New York Knicks really tough!”

3. 台泥登50元水泥股硬頸.

   táiní dēngwǔshíyuán shuǐnígǔ yìngjǐng.
   TaiNi-climb up-NT50-dollars-cement stock-stiff neck.
   “TaiNi climbs up to 50 dollars. The cement stock stands steady!”

This case study demonstrates research integrating computational and corpus methods for the linguistic analysis of news discourse. In particular, it shows how the recontextualization of metaphor is contingent upon language use in context. The result draws attention to the complex intertwining of language use, communication, and society.

## 9.4 Taiwan Mandarin, Spoken Corpus, and Linguistic-Gestural Behaviors and Cognition

When people engage in a conversational talk, the spontaneous use of gestures along with speech is prevalent and indispensable (Goldin-Meadow 1999; Kendon 2004; McNeill 1992, 2000). Gestures are mainly performed in the central gesture space with noticeable and discernable configurations. Speakers make utterances and move their hands and arms simultaneously to convey meaning, whereas addressees comprehend the linguistic and gestural meaning for understanding the messages. People’s conception is, thus, claimed to include both linguistic and imagery content. A corpus of spoken data is indispensable for the research on the form, meaning, and use of language and gesture, and the linguistic-gestural behaviors and cognition of the language community of Taiwan Mandarin.

The NCCU Corpus of Spoken Mandarin has been collecting data of spontaneous face-to-face conversations since 2006. The participants were recruited to hold a conversation with acquaintances. They were free to initiate any topics of interest and develop the sequential turns in their own way. See the screenshots of some of the recordings in Fig. 9.3. At present, the corpus includes thirty-two conversational excerpts totaling about 800 min of talk. With the written consent from the participants, the data are available online.


**Fig. 9.3** Recordings of daily conversation

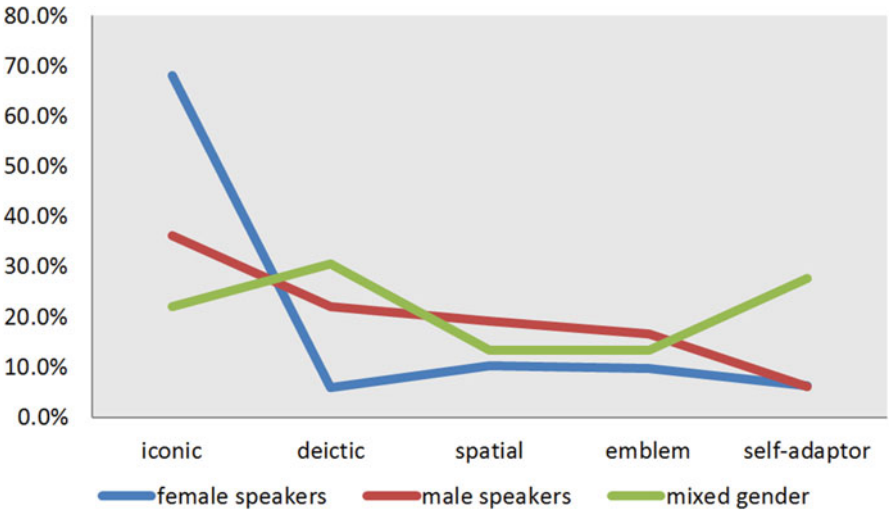



**Fig. 9.4** Occurrence of gesture across different types of interaction

We talk; we gesture. Various kinds of gesture occur: "iconic gesture" bears a direct semantic relation to speech, depicting the meaning of a concept; "deictic gesture" points to a referent in the physical speaking environment; "spatial gesture" locates a referent in the gesture space; "emblematic gesture," such as waving the hand for "goodbye," conveys a conventionalized meaning; "self-adaptor" is a self-touching gesture. In 15 conversational excerpts that totaled 300 min of talk, 2012 gestures were found across male-speaker conversations, female-speaker conversations, and mixed-gender conversations. Figure 9.4 presents the occurrence rates of the five types of gesture across the three types of interaction; gender appears to affect the production of gesture.

Face-to-face conversations are the most fundamental type of talk-in-interaction (Bavelas and Chovil 2006; Clark 1996; Stivers et al. 2009). The naturally occurring data including speech and gesture have been digitized for computing the real-time manifestation of linguistic-gestural behaviors and cross-modal cognition.

Manual configurations convey meaning. "Speech and gesture refer to the same event and are partially overlapping, but the pictures they present are different"

(McNeill 1992, p. 13). Both modalities create a more complete message than either modality can alone (Clark 1996). In Mandarin conversational discourse, complementary gestures, which do not have a direct syntactic and semantic relation with particular linguistic constituents in the utterances, may depict attitudinal, script-evoked, and topical information to enrich speech events or maintain the continuity of a topic under discussion. In sequential exchanges, the speaker also uses gesture to complete the meaning of the utterance, whereas the addressee acknowledges or negotiates with the speaker of the gestural information to accomplish conversational coherence. Moreover, gestures can be used to compensate for the lack of path information in speech. Since English and Turkish speakers do not gesture in the same way (Özyürek et al. 2005), the gestural behavior in Taiwan Mandarin demonstrates cross-linguistic variation and language specificity. Finally, for the joint construction of meaning across speakers (Clark 1996; Holler and Wilkin 2011), mimicking someone's gesture in interaction can achieve common ground and mutual understanding. A mimicked gesture also functions as a semantic basis upon which new information can be presented in discourse. All of these linguistic-gestural patterns arise naturally in real-time interactions.

The inevitable, indispensable occurrence of gesture in daily communication and the tight semantic relationship between the two modalities attest to the cognitive unity of language and gesture (McNeill 1992, 2000; Núñez and Cooperrider 2013). In Taiwan Mandarin, metaphoric gestures provide visible evidence for cross-domain cognitive mappings and the embodiment of conceptual metaphors in people's perceptual and bodily experiences in recurrent sociocultural activities and individual incidences. Metaphoric gestures also reveal the speaker's real-time focus of attention on a particular aspect of conceptualization in communication. Second, knowledge in conceptual frames is also readily manifested in gesture. Gestures not only reveal roles and role relations in a scene as distinct from those in speech, they also co-occur with speech to jointly reveal scriptal knowledge. In line with Langacker's (2008) view of prominence, gesture depicting the part of frame knowledge being selected for expression is also the speaker's focal attention on a certain aspect of the scene during speaking. "[G]esture is an inherent part of language – gestures work as signs communicating thought" (Lakoff 2008, p. 284). The use of language and gesture in daily conversation bears out embodied cognition, conceptual complexity, modal-specificity, and culture-specificity in conceptualization.

To conclude, a spoken corpus of Taiwan Mandarin provides rich speech and gesture data for the understanding of the nature of cross-modal communication, gesture in thought, and the cognitive unity of speech and gesture. As the neural integration of language and gesture has been widely attested by neuroimaging studies (such as Özyürek et al. 2007; Holle et al. 2008; Straube et al. 2012; Dick et al. 2014), the corpus-based studies of the form, meaning, and use of language and gesture constitute the empirical base for the further research on the neurocognition of linguistic-gestural universality and specificity in Taiwan Mandarin.

## 9.5 Linguistic and Cognitive Analyses of Narrative Ability: A Comparison of Typical Children and Children with Autism Spectrum Disorder

Digitized language data has also been employed to explore the recurrent patterns of linguistic form and use in atypical children's narrative discourse. Because narrative production involves an integration of social–emotional, cognitive, and linguistic knowledge, studies of oral narratives produced by individuals with autism spectrum disorder (ASD) have revealed rich information about the social–communicative impairments in this population (for a review, see Stirling et al. 2014). As Loveland and Tunali (1993) suggested, the difficulties that autistic individuals experience in narrative production reflect their impaired mindreading abilities that render them less sensitive to listeners' informational needs and perspectives. Among various indices of narrative abilities, referential choice is regarded as an important window to gauge the interaction between linguistic patterns and cognition, and, in particular, to show a speaker's sensitivity to listeners' needs. In the study by Colle et al. (2008), for instance, the subtle but significant impairment in the referential use of pronouns in English-speaking individuals with ASD is considered relevant to the insensitivity of the affected individuals as to what listeners need.

Though there have been few detailed investigations about the narrative abilities of Mandarin-speaking children with ASD (e.g., Sah and Torng 2015; Tsou and Cheung 2007), we still lack knowledge about how Mandarin-speaking autistic children use referential terms in narratives. This study, therefore, examined a Chinese corpus of oral narratives produced by Mandarin-speaking children with ASD and typically developing (TD) children. The narrative data was analyzed in terms of referential forms and pragmatic functions. The referential forms included: (1) nominal forms, (2) pronominal forms, and (3) null forms; the pragmatic functions were examined when a referential form is used to (1) introduce, (2) maintain, or (3) reintroduce a referent in a narrative.

The results reveal that TD and ASD children are comparable in using nominal forms, irrespective of the pragmatic functions involved. Further detailed analysis indicates: Both groups display a preference to introduce and to reintroduce referents with nominal forms, whereas null forms are the most commonly used way for maintaining reference. The predominance of nominal forms for introducing reference is consistent with the previous findings for English-speaking children with ASD (Colle et al. 2008). Inconsistent with the previous findings, however, null forms, rather than pronominal forms, appear to be the dominant device for Mandarin-speaking children, both ASD and TD, to maintain reference. According to Chafe (1994), a subsequently mentioned referent is the active referent and it is expected to be realized by means of pronominal forms. Our data, however, display a picture different from what was expected. One possible reason for this discrepancy is that null forms are the most preferred device for this kind of maintaining function for Mandarin-speaking children. As Hickmann and Hendriks (1999) indicated, null forms are used very frequently by Chinese speakers, though they might be

unacceptable in many languages. Another related finding about the preponderance of null forms is noted in the report on Japanese narratives, in which young children are found more likely to use null forms for maintaining reference (Clancy 1982).

To recapitulate, Mandarin-speaking children with ASD are sensitive to discourse-pragmatic functions and listeners' needs while making referential choices in narrative discourse. Both ASD and TD groups tend to use nominal forms for the purposes of introducing and reintroducing referents, which suggests that they would make efforts to reduce the potential uncertainty regarding the referents they are talking about. It is worth noting that null forms, rather than pronominal forms, appear to be the least effortful way for both ASD and TD children to maintain reference. This particular finding is relevant to a characteristic feature of Mandarin Chinese that the language allows arguments to be grammatically null, and thus, referents that can be understood from discourse contexts do not need to be overtly specified. Plausible as this interpretation may seem, further empirical inquiry into the use of referential choice in different narrative genres is needed to illuminate the nature of such choice by Mandarin-speaking children with ASD. Nevertheless, this study shows how corpus-based studies can further our understanding about the interaction between linguistic forms and functions in oral narratives by typical and atypical children.

## 9.6 Conclusion

Usage-based approaches to language analyses advocate that meaning drives grammar, grammaticality is an entrenchment, form-meaning pairings are analytic units, categorization is based on encyclopedic semantics, and that language should be accounted for through the interaction of all of the dimensions of meaning (Glynn 2010). These viewpoints underlie our corpus-based linguistic studies that demonstrate the intricate patterns of language as manifested in different social–cultural discourse in Taiwan. Moreover, the use of language and gesture in spoken Taiwan Mandarin manifests the linguistic-gestural behaviors and linguistic-gestural cognition embodied in people's perceptual and bodily experiences in recurrent individual, social–cultural practices. Finally, as narrative data is critical for the investigation of pragmatic and social–cognitive abilities, the comparison of the narrative data produced by typical and atypical children bears out the developmental integration of social–emotional, cognitive, and linguistic abilities across the two groups of young language users. The baseline findings from these studies lay the important empirical foundation for further research on speech communication in other language communities in Taiwan.

## References


Bavelas, J. B., & Chovil, N. (2006). Hand gestures and facial displays as part of language use in face-to-face dialogue. In V. Manusov & M. L. Patterson (Eds.), *The Sage handbook of nonverbal communication* (pp. 97–115). Thousand Oaks, CA: Sage Publications.

Beckner, C., Blythe, R., Bybee, J., Christiansen, M. H., Croft, W., Ellis, N. C., et al. (2009). Language is a complex adaptive system: Position paper. *Language Learning, 59*(s1), 1–26.

Biber, D. (2010). Corpus-based and corpus-driven analyses of language variation and use. In B. Heine & H. Narrog (Eds.), *The Oxford handbook of linguistic analysis* (pp. 195–223). Oxford: Oxford University Press.

Boers, F. (1999). When a bodily source domain becomes prominent: The joy of counting metaphors in the socio-economic domain. In R. W. Gibbs Jr. & G. J. Steen (Eds.), *Metaphor in cognitive linguistics* (pp. 47–56). Philadelphia, PA: John Benjamins.

Cameron, L. J., & Deignan, A. (2006). The emergence of metaphor in discourse. *Applied Linguistics, 27*(4), 671–690.

Chafe, W. (1994). *Discourse, consciousness, and time: The flow and displacement of conscious experience in speaking and writing*. Chicago: University of Chicago Press.

Clancy, P. M. (1982). Written and spoken style in Japanese narratives. In D. Tannen (Ed.), *Spoken and written language: Exploring orality and literacy* (pp. 55–76). Norwood, NJ: Ablex.

Clark, H. H. (1996). *Using language*. Cambridge: Cambridge University Press.

Colle, L., Baron-Cohen, S., Wheelwright, S., & van der Lely, H. K. (2008). Narrative discourse in adults with high-functioning autism or Asperger syndrome. *Journal of Autism and Developmental Disorders, 38*(1), 28–40.

Deignan, A. (2005). *Metaphor and corpus linguistics*. Philadelphia: John Benjamins.

Deignan, A. (2008). Corpus linguistics and metaphor. In R. W. Gibbs Jr. (Ed.), *The Cambridge handbook of metaphor and thought* (pp. 280–294). New York: Cambridge University Press.

Deignan, A. (2012). Figurative language in discourse. In H. J. Schmidt (Ed.), *Cognitive pragmatics* (pp. 437–462). Berlin: Walter de Gruyter.

Dick, A. S., Mok, E. H., Beharelle, A. R., Goldin-Meadow, S., & Small, S. L. (2014). Frontal and temporal contributions to understanding the iconic co-speech gestures that accompany speech. *Human Brain Mapping, 35*(3), 900–917.

Geeraerts, D. (2010). *Theories of lexical semantics*. New York: Oxford University Press.

Glynn, D. (2010). Corpus-driven Cognitive semantics: Introduction to the field. In D. Glynn & F. Kerstin (Eds.), *Quantitative methods in cognitive semantics: Corpus-driven approaches* (pp. 1–41). Berlin: Walter de Gruyter.

Goldin-Meadow, S. (1999). The role of gesture in communication and thinking. *Trends in Cognitive Science, 3*, 419–429.

Hickmann, M., & Hendriks, H. (1999). Cohesion and anaphora in children's narratives: A comparison of English, French, German, and mandarin Chinese. *Journal of Child Language, 26*, 419–452.

Holle, H., Gunter, T. C., Rüschemeyer, S. A., Hennenlotter, A., & Iacoboni, I. (2008). Neural correlates of the processing of co-speech gestures. *NeuroImage, 39*(4), 2010–2024.

Holler, J., & Wilkin, K. (2011). Co-speech gesture mimicry in the process of collaborative referring during face-to-face dialogue. *Journal of Nonverbal Behavior, 35*, 133–153.

Kendon, A. (2004). *Gesture: Visible action as utterance*. Cambridge: Cambridge University Press.

Kövecses, Z. (2002). *Metaphor: A practical introduction*. New York: Oxford University Press.

Lai, H. L. (2017). Understanding ethnic visibility through language use: The case of Taiwan Hakka. *Asian Ethnicity*, *18*, 406–423.

Lakoff, G. (1993). The contemporary theory of metaphor. In A. Ortony (Ed.), *Metaphor and thought* (pp. 202–251). Cambridge: Cambridge University Press.

Lakoff, G. (2008). The neuroscience of metaphoric gestures: Why they exist. In A. Cienki & C. Müller (Eds.), *Metaphor and gesture* (pp. 283–289). Amsterdam, Philadelphia: John Benjamins.

Lakoff, G., & Johnson, M. (1980). *Metaphors we live by*. Chicago: University of Chicago Press.
Langacker, R. W. (2008). *Cognitive grammar: A basic introduction*. New York: Oxford University Press.
Loveland, K., & Tunali, B. (1993). Narrative language in autism and the theory of mind hypothesis: A wider perspective. In S. Baron-Cohen, H. Tager-Flusberg, & D. J. Cohen (Eds.), *Understanding other minds: Perspectives from autism* (pp. 247–266). Oxford: Oxford University Press.
McNeill, D. (1992). *Hand and mind: What gestures reveal about thought*. Chicago: The University of Chicago Press.
McNeill, D. (Ed.). (2000). *Language and gesture*. Cambridge: Cambridge University Press.
Núñez, R., & Cooperrider, K. (2013). The tangle of space and time in human cognition. *Trends in Cognitive Sciences, 17*(5), 220–229.
Özyürek, A., Kita, S., Allen, S., Furman, R., & Brown, A. (2005). How does linguistic framing of events influence co-speech gestures? Insights from cross-linguistic variations and similarities. *Gesture, 5*, 215–237.
Özyürek, A., Willems, R. M., Kita, S., & Hagoort, P. (2007). On-line integration of semantic information from speech and gesture: Insights from event-related brain potentials. *Journal of Cognitive Neuroscience, 19*(4), 605–616.
Porto, M. D., & Romano, M. (2013). Newspaper metaphors: Reusing metaphors across media genres. *Metaphor and Symbol, 28*, 60–73.
Radden, G., & Kövecses, Z. (1999). Towards a theory of metonymy. In K. U. Panther & G. Radden (Eds.), *Metonymy in language and thought* (pp. 17–59). Philadelphia: John Benjamins.
Sah, W. H., & Torng, P. C. (2015). Narrative coherence of Mandarin-speaking children with high-functioning autism spectrum disorder: An investigation into causal relations. *First Language, 35*(3), 189–212.
Semino, E., Deignan, A., & Littlemore, J. (2013). Metaphor, genre, and recontextualization. *Metaphor and Symbol, 28*, 48–59.
Stirling, L., Douglas, S., Leekam, S., & Carey, L. (2014). The use of narrative in studying communication in autism Spectrum disorders: A review of methodologies and findings. In J. Arciuli & J. Brock (Eds.), *Communication in autism* (pp. 169–216). Amsterdam: John Benjamins.
Stivers, T., Enfield, N. J., Brown, P., Englert, C., Hayashi, M., Heinemann, T., Hoymann, G., Rossano, F., Ruiter, A. P. D., Yoon, K. E., & Levinson, S. C. (2009). Universals and cultural variation in turn-taking in conversation. *Proceedings of the National Academy of Sciences of the United States of America, 106*(26), 10587–10592.
Straube, B., Green, A., Weis, S., & Kircher, T. (2012). A supramodal neural network for speech and gesture semantics: An fMRI study. *PLoS One, 7*(11), e51207.
Tsou, C. Z., & Cheung, H. (2007). Narrative story telling of high-functioning children with autism spectrum disorders. *Bulletin of Special Education, 32*(3), 87–109.
Yu, N. (2008). Metaphor from body and culture. In R. W. Gibbs Jr. (Ed.), *The Cambridge handbook of metaphor and thought* (pp. 247–261). New York: Cambridge University Press.

# Chapter 10
# From Naive Expectation to Realistic Progress: Government Applications of Big Data on Public Opinions Mining


**Naiyi Hsiao, Zhoupeng Liao, and Don-Yun Chen**

## 10.1 Big Data Analytics and Public Opinions Mining

Identifying public policy agenda and relevant issues have served one of the crucial stages in public policy analysis. In addition to the long existing channels such as telephones and newspapers, the Internet has been emerging as the most challenging source of citizens' complaints and comments on public policy as the netizens, namely the citizens on the Internet, have been voicing since 1990s. In contrast with the traditional methods of public opinions survey such as face-to-face and telephone (including mobile phones) interviews, the Internet survey with the general public has been criticized with its sampling bias and representativeness due to uncontrolled demographics of the respondents. Besides, the existing survey methods may also raise reasonable suspect of methodological obtrusiveness where the contacted respondents are likely to provide ignorant and even twisted answers contingent upon their own interests and political preferences.

The technical and computational advances in recent decades of statistics and semantic analyses such as data mining and text mining have shed light on alternatives to the previous concerns. Particularly, machine-learning algorithms that can more reliably and validly parse texts provide promising solutions to collect and analyze public comments on the Internet. The rising tides of Big Data Analytics have appeared expected promising and in some cases demonstrated with actual

N. Hsiao (✉) · D.-Y. Chen
Department of Public Administration & Taiwan E-Governance Research Center, National Chengchi University, Taipei, Taiwan
e-mail: nhsiao@nccu.edu.tw; donc@nccu.edu.tw

Z. Liao
Department of Public Administration & Taiwan E-Governance Research Center, National Open University, New Taipei City, Taiwan
e-mail: zpliao@mail.nou.edu.tw



S.-H. Chen (ed.), *Big Data in Computational Social Science and Humanities*, Computational Social Sciences, https://doi.org/10.1007/978-3-319-95465-3_10

benefit. The existing practice and literature, however, remain insufficient to provide systematic investigation for conducting Internet public opinions analysis (IPOA). Moreover, there is virtually no in-depth observation regarding how governments in charge of public policies may implement and benefit from IPOA as well as the accompanying costs and challenges.

Partnering with a private IPOA service provider, the present study reflects upon planning and implementing IPOA in a public agency via a series of interviews and field observation. The prototyping approach was conducted during May and November in 2014 to collect and analyze unstructured public opinions regarding Free Economic Trade Zone (FETZ) charged by National Development Council (NDC), Taiwan. Weekly or bi-weekly reports were produced by the research team composed of experts and academicians in computing technology and public administration and policy. The public officials of NDC closely worked with the research team and revised the IPOA implementation processes contingent upon their onsite evaluation. Two rounds of focus group interviews containing experts in various domains such as law, technology, public relations, and political practitioners were hosted to explore and resolve challenging issues along the prototyping process. At the second half in the IPOA implementation process, nearly 40 career officials from various ministries in Taiwanese central government participated in three rounds of reviews to make sense of the previous prototyping results of IPOA.

## 10.2 Big Data and its Applications to Government and Public Policy

Due to substantial progress and popularity of information and communication technology, the digital data has exponentially accumulated in the recent decades. The term Big Data emerged in the 2000s as a rhetorical phrase to constantly increasing amounts of data. Actually, Big Data is an evolving concept that also refers to how data are collected, analyzed, and optimized for business process and citizen value. Big Data is often defined along three-Vs dimensions, namely volume, velocity, and variety (Desouza 2014; TechAmerica Foundation 2012). It meanwhile represents various challenges in employing advanced techniques and technologies to enable the capture, storage, distribution, management, and analysis of the data. Most of the current studies generally agree that Big Data, given proper management, has the potential to transform government by providing greater insight and by better serving the citizenry, society, and the world. Specifically, Big Data holds vast potential for improving the decision-making processes by domain-specific analytics needed in many critical and high-impact application areas involving the economy, health care, job creation, education, natural disasters, terrorism, etc. (Desouza and Jacob 2014; Kim et al. 2014). As Data Science has emerged and continued to progress, several Big Data initiatives have recently emerged in the public sector. For example, in March 2012, the Obama Administration put forward the Big Data

Research and Development Initiative which was used to understand the technologies needed to manipulate and mine massive amounts of information, in order to apply that knowledge to apply for many public issues and to engage citizens with public data (Desouza and Jacob 2014, p. 2).

However, as a large portion of the Big Data literature focuses on its expected benefit, there are obstacles needed to be addressed while applying Big Data in the public sector, including key governmental data analysis issues—governance and privacy (Desouza and Jacob 2014). Firstly, as Kim et al. (2014, p. 80) point out, government data can be categorized as silo, security, and variety, therefore, there is the lack of standardized solutions, software, and cross-agency solutions for extracting useful information from discrete datasets in multiple government agencies and insufficient funding due to government austerity measures to develop and implement these solutions. Therefore, overcoming IT and data governance difficulty within the government data nature remains a thorny issue as we deploy Big Data to enhance government capacity to more effectively address major public problems.

Secondly, Big Data can be especially powerful as it connects and explores correlations between previously disparate datasets in a wide array of public domains. However, this powerful dimension of Big Data has raised concerns about the citizens' privacy because these connections can reveal hidden behavioral patterns about individual citizens who do not consent to (Desouza and Jacob 2014, p. 9). For example, the cameras which were set in a city can be used to provide necessary information for crime prevention, but it can also be a useful tool to harm the citizens' rights by an illegal scrutiny. Besides, criminals may have the chance to use videos to target home owners who do not stay at home for stealing their private properties.

We now live in a digital world where the scale and scope of value derived from data exhibit an exciting progress. Hidden in immense volume, variety and velocity of data produced is new information—facts, relationships, indicators, and pointers—that either cannot be practically discovered in the past or simply do not exist before (TechAmerica Foundation 2012, p. 37). Many practitioners and researchers have recognized that there is a great opportunity for adopting Big Data in public domains (Chen et al. 2012). Especially when government and political processes become more transparent, participatory, online, and multimedia-rich, several useful Big Data techniques, including public opinions mining, social network analysis, text analysis, and sentiment analysis, can support online political participation and e-democracy, e-government service delivery, and accountability.

## 10.3 Conducting Internet Public Opinions Analysis (IPOA)

As the research design summarized above, a complete IPOA process was planned and implemented in NDC in charge of one of the most controversial policies then in Taiwan, Free Economic Pilot Zone[1] (FEPZs). NDC, in Taiwanese central

[1]Refer to the official website of Free Economic Pilot Zones (FEPZs): http://www.fepz.org.tw/en_index.aspx.

government Executive Yuan, serves as planning and evaluation of economic and social development policy. NDC has been invested in policy research regarding various policy issues including public governance and e-governance topics.

Since March 2014, the FEPZs initiative was submitted to Legislative Yuan for legislative review and approval; hence it drew substantial public attention especially in the same period a student movement was also marched between March and April. One of the major reasons for FEPZs controversy stems from its free trade agreement (FTA) with Mainland China. Most suspicion has been raised for possible loss of political independence due to increasing economic dependency on Mainland China. It was also noting that the core members and supporters of the student movement, coined as Sunflower Student Movement,[2] heavily utilized the Internet-related platforms and tools, such as PTT (one of the most influential BBS in Taiwan), YouTube, and Web seminars, to record and promote the movement. Working with the authors, NDC therefore recommended FEPZs as the policy issue that deserves exploration of how IPOA may help collect and analyze public opinions.

### *10.3.1 The Prototyping Process for IPOA Implementation*

NDC management in April 2014 decided to implement IPOA along with FEPZs legislative process by contracting with the authors and technical service provider, eLand Technologies. Based on several rounds of experimental reports and face-to-face discussion among the three parties (NDC, eLand, and the authors), an iterative process of IPOA implementation was developed to transform Internet-based public opinions into policy-relevant information meaningful to policy decision makers. The key action of IPOA witnessed in the process, as in Fig. 10.1, includes the following components.

1. Policy domain experts in NDC and other ministries work with the external policy consultants and IPOA technical provider to propose and decide FEPZs-related subjects and keywords, as well as frequency and time frame of the follow-up IPOA reports. Due to various wordings posted by the general public, trials and errors via the IPOA online query (Fig. 10.2) contributes to both NDC experts and the external consultants to develop a set of subjects and keywords that can effectively collect netizens' comments relevant to the targeted policy FEPZs from various sources of Internet media including news websites, forums, blogs, and social media such as Facebook, Twitter, and PTT.
2. In addition, NDC management has also to specify key events and policy-relevant stakeholders (either individuals or groups) related to FEPZs. The step may also expand the Internet media specified above as the events and stakeholders of policy authority's special interest may not be included by default.

[2]Refer to the Wiki website of Sunflower Student Movement: http://en.wikipedia.org/wiki/Sunflower_Student_Movement.

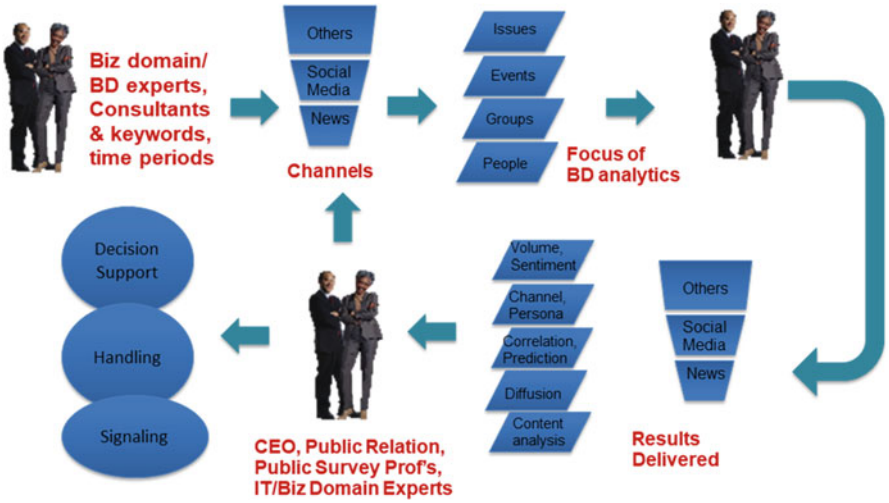



**Fig. 10.1** The iterative process of IPOA implementation


**Fig. 10.2** The online query of IPOA

3. Preliminary IPOA results, including volume, sentiment analyses, and so on (detailed below), are produced for further examination and discussion by policy domains officials, public relations managers, and public survey professionals. The preliminary results usually remain improved for completeness and readability.
4. The ultimate purposes of IPOA results and interpretation should be linked to strategy, decision-making, and follow-up actions. In the final stage, collective sense-making and brain-storming activities from high-ranked public officials and decision makers significantly determine the actual efficacy of the overall IPOA implementation.

### *10.3.2 Presentation and Interpretation of IPOA Results*

The IPOA implementation in the study produces a series of reports that attempt to capture the public opinions and their attributes on the Internet. Firstly, Fig. 10.3 exhibits the storm of FEPZs policy-relevant keywords that extract from all negative public comments on the Internet. The distance between the extracted keywords and the lower left corner, FEPZs, represents the degree of connection among the keywords. The colors and sizes of the circles symbol the volumes of the negative public comments with the corresponding keywords. In Fig. 10.3, the storm center stands for a series of events "Public Conference for FEPZs" hosted by NDC in June and July, 2014. The "storm" diagram shows that, between June 1 and July 7, some negative comments on the Internet related to the keywords have collected on the Internet. The policy makers would expect to benefit from further identification and analyses of the negative comments.

While Fig. 10.3 summarizes the keywords from public comments with negative sentiments, Fig. 10.4 below sketches the longitudinal volumes and sentiments by connecting them to the public events in the same time periods. All including negative and positive public comments are categorized by their channels of being posted, namely news and social media. The news channels include all mainstream newspapers and TV channels, and the social media consist of bulletin board systems (BBS), blogs, and all types of social networking sites such as Facebook, Twitter, and PTT, indicated above as one of the most popular and influential BBS and social networking sites especially among college students and young netizens. In Fig. 10.4, some volume peaks of public opinions concerning FEPZs may be identified by the NDC officials and further explored by their correlation with particular events and news.

In addition, the netizens' positive and negative sentiments are also depicted in Fig. 10.4, where significant positive/negative sentiments are attached to the public comments. The Internet media contents can be parsed and matched with sufficient meanings of positive and negative wordings. As public comments may contain both positive and negative wordings, positive and negative sentiments are analyzed independently. That is, a specific piece of public comments may contain

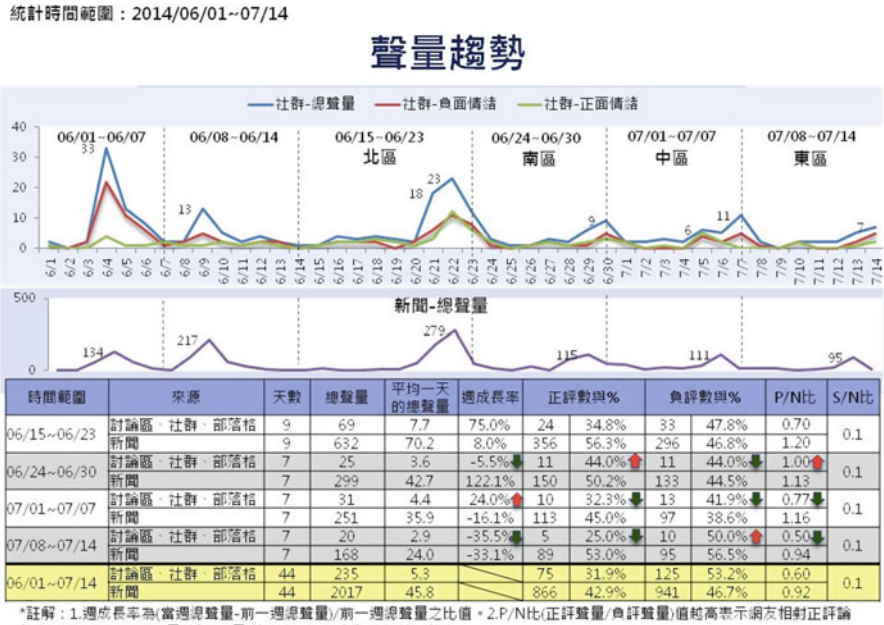

| 新增負評關鍵字 | 相關內容 | 曝光討論版 |
|---|---|---|
| 許柱森、農委會 | 許柱森自爆「本身就是受益者」，表示自己當青年農民貸款買了土地，「聽了政府的話，所以過得很好。」忽然他想起經貿圈是會議有網路直播，急忙指點直播人員：「這個不要轉進去。」讓網友質疑此舉掩守有問題。 | Ptt > Gossiping |
| | 農委會副會長許柱森，也在會議中對用青農低利貸款買地一事做出澄清，強調自己的農民身分沒問題 | Ptt > Gossiping |

**Fig. 10.3** Storm diagram of negative comments and relevant keywords

both significant positive and negative sentiments. The machine-learning algorithms for sentiments analysis have been widely adopted worldwide for text mining and analytics. Moreover, the research team has invested in quality measurement by content analyses from human readers. In average the accuracy reaches 80% of relevance and sentiments, which is regarded acceptable by the policy makers.

Over the time periods between June 1 and July 14, as shown in the last row of the summary table in Fig. 10.4, the percentages of negative comments, namely the public comments judged with significantly negative sentiments, appear higher than positive comments especially in those posted on social media channels. P/N Ratio, 0.60 and 0.92, respectively, calculated by the percentages of positive comments divided by negative comments, then stands for the overall evaluation of public sentiments concerning FEPZs over the time period. The results by P/N Ratio appear overall negative public sentiments towards FEPZs. S/N Ratio in the last column calculates the volumes on the news websites versus the social media websites. The numbers, 0.1 for all time periods, show that the public comments on FEPZs are mostly posted on the news websites rather than on the social media.

Aside from analyzing volumes and sentiments, IPOA also attempts to capture the correlation between specific events and news with public opinions. As shown in Fig. 10.5, inspecting the rise and fall of negative comments concerning FEPZs over 2 weeks, we can possibly infer the increasing volumes in July 13 to the public




| 時間範圍 | 來源 | 天數 | 總聲量 | 平均一天的總聲量 | 週成長率 | 正評數與% | | 負評數與% | | P/N比 | S/N比 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 06/15~06/23 | 討論區、社群、部落格 | 9 | 69 | 7.7 | 75.0% | 24 | 34.8% | 33 | 47.8% | 0.70 | 0.1 |
| | 新聞 | 9 | 632 | 70.2 | 8.0% | 356 | 56.3% | 296 | 46.8% | 1.20 | |
| 06/24~06/30 | 討論區、社群、部落格 | 7 | 25 | 3.6 | -5.5% | 11 | 44.0% | 11 | 44.0% | 1.00 | 0.1 |
| | 新聞 | 7 | 299 | 42.7 | 122.1% | 150 | 50.2% | 133 | 44.5% | 1.13 | |
| 07/01~07/07 | 討論區、社群、部落格 | 7 | 31 | 4.4 | 24.0% | 10 | 32.3% | 13 | 41.9% | 0.77 | 0.1 |
| | 新聞 | 7 | 251 | 35.9 | -16.1% | 113 | 45.0% | 97 | 38.6% | 1.16 | |
| 07/08~07/14 | 討論區、社群、部落格 | 7 | 20 | 2.9 | -35.5% | 5 | 25.0% | 10 | 50.0% | 0.50 | 0.1 |
| | 新聞 | 7 | 168 | 24.0 | -33.1% | 89 | 53.0% | 95 | 56.5% | 0.94 | |
| 06/01~07/14 | 討論區、社群、部落格 | 44 | 235 | 5.3 | | 75 | 31.9% | 125 | 53.2% | 0.60 | 0.1 |
| | 新聞 | 44 | 2017 | 45.8 | | 866 | 42.9% | 941 | 46.7% | 0.92 | |

*註解：1.週成長率為(當週總聲量-前一週總聲量)/前一週總聲量之比值。2.P/N比(正評聲量/負評聲量)值越高表示網友相對正評論較高。3.S/N比(社群聲量/新聞聲量)值越高表示網友相對討論度較高。

**Fig. 10.4** Longitudinal volumes and sentiments of internet public opinions

promotion event (world café). Based on the moving average (dotted line) towards the future, the same event in July 13 and 20 may also draw public attention.

While the previous Figs. 10.4 and 10.5 are interpreted as the volumes and sentiments of the general public opinions, Figs. 10.6 and 10.7 further look into where the public Internet-based comments come from. In Fig. 10.6, for the past 2 weeks (July 1–14) and the whole period with policy promotion activities (June 1–July 14), most (at least 64%) of the public opinions about FEPZs are found on social media websites (red bars) such as Facebook and PTT in Taiwan, followed by public comments on Internet forums (BBSs, blue bars) and a small portion on blogs (green bars). Figure 10.7 shows the top ten popular channels of Internet media carrying on the public comments on FEPZs, including PTT > Gossiping and so on.

Figure 10.8 then uncovers who are the leading voices by defining "leading" as the ratio of the responding comments due to the original comments. It is worth noting that on top of the opinion leaders list is the official Facebook ID "Go! Go! Taiwan!" The single message rated as positive (0.594 from the scale 0 to 1) attracts 377 responses from the netizens, though rated relatively negative with 62 significantly positive posts, 228 significantly negative posts, and leading to a low P/N Ratio 0.27. The second and third IDs, both PTT > britvic and PTT > 900183 users on PTT, on the list appear to be ordinary netizens both with negative sentiments and also drawing negative comments towards FEPZs.

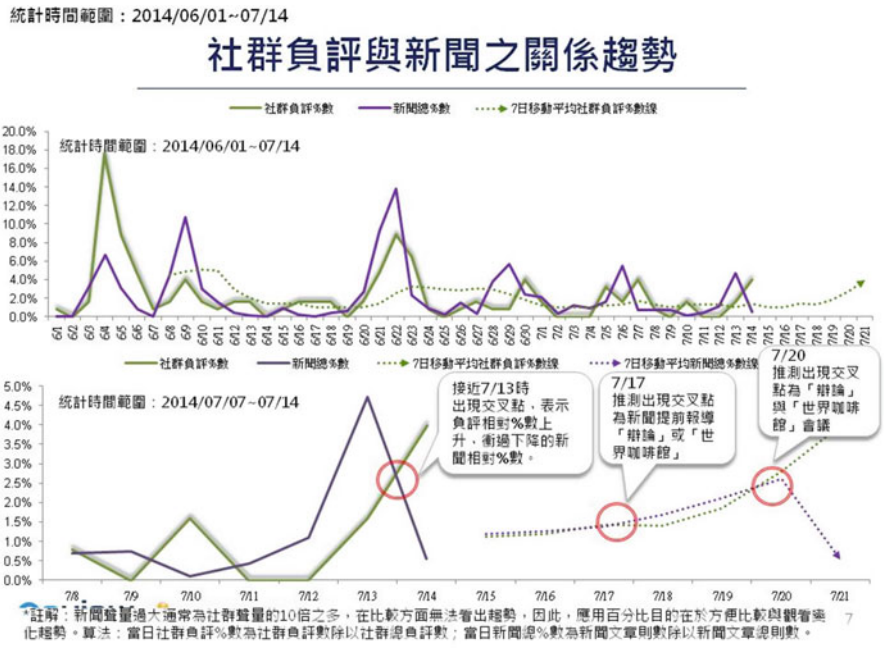



**Fig. 10.5** Correspondence between volumes and sentiments on social media

The most challenging results demonstrate how the NDC policy makers and the research team may inspect the contents of the public comments following specific events such as public promotion activities. It may involve more or less sampling bias to decide which pieces of public opinions should be particularly examined and interpreted. The content analyses, as well as the interpretations of IPOA statistics and charts presented above, can inspire the NDC policy makers working with the research team to figure out policy-relevant strategies, decisions, and actions.

## 10.4 Discussions and Concluding Remarks

The most straightforward understanding of IPOA, as in Fig. 10.9, involves the entities and their relationships attached to specific attributes, underlying the present study that transforms naïve expectation of Big Data applications in government into realistic progress. IPOA aspires to analyze the netizens' opinions posted on Internet media regarding policy issues of their interests. The netizens' anonymous IDs may provide individual and group demographics perceived by policy domain experts. While the types of Internet media are generally news websites, discussion lists (forums), blogs, and social networking sites, policy domain experts may recognize and recommend specific media to be included for public opinions collection and analysis. Lastly, the public opinions concerning a specific policy issue such as

FEPZs in the study can virtually produce all possible cross-tabulation for all attributes attached to netizens, Internet media, and policy issues concerned as long as numerical and text data are available on the Internet.

### *10.4.1 IPOA Contribution to Policy Analysis*

The field experience and observation reported above contribute to the development of a step-by-step process (Fig. 10.1) to facilitate how career public authority of policy domains, such as FEPZs in the study, interact with consulting professionals and the IPOA technical service provider. Unlike transaction-oriented information systems, developing IPOA is much similar to a decision support system that requires iterative communication and interpretation. The preceding empirical results show that implementing IPOA requires significant and effective cross-discipline and cross-agency collaboration. Particularly, business units, public relation units, research and development units, and information technology units have to work closely to achieve the expected benefit. The various units in a public organization have also to collaborate properly with the computing providers and the external consulting team.

Moreover, volumes and sentiments analyses (Fig. 10.4) across time frames and Internet media channels (Figs. 10.6, 10.7, and 10.8) have effectively provided

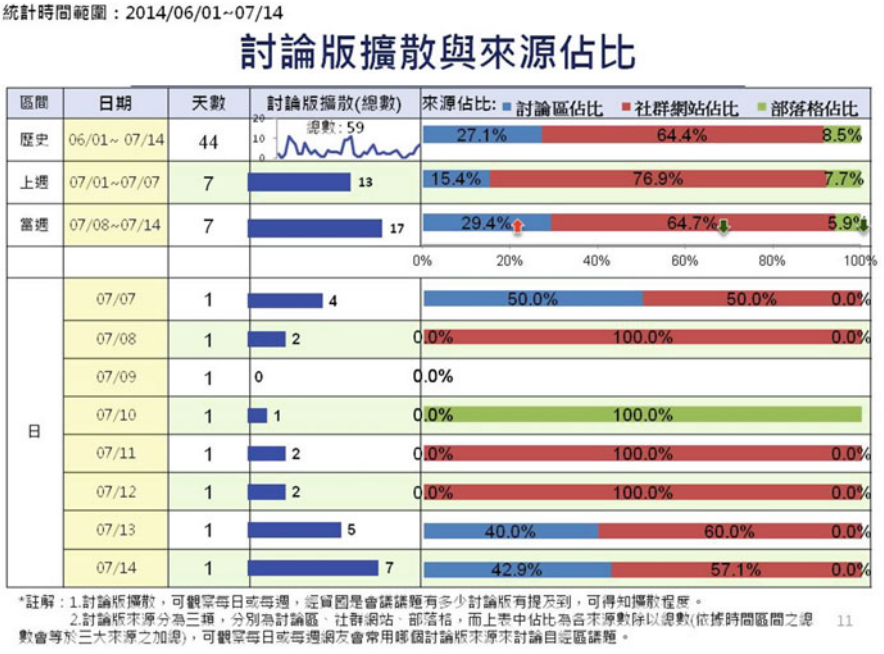
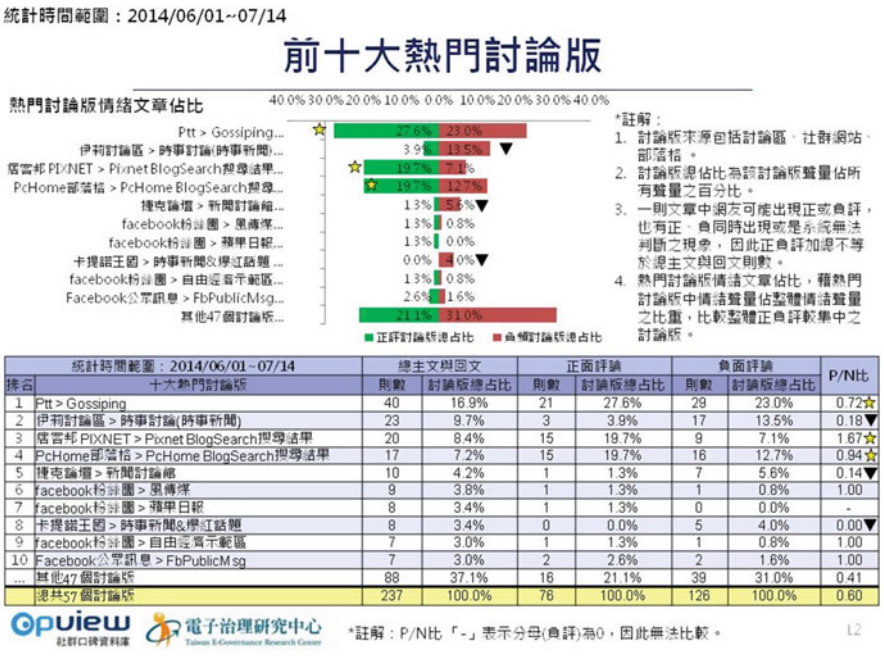



**Fig. 10.6** Popular types of internet media on FEPZs




| 統計時間範圍：2014/06/01~07/14 | | 總主文與回文 | | 正面評論 | | 負面評論 | | P/N比 |
|---|---|---|---|---|---|---|---|---|
| 排名 | 十大熱門討論版 | 則數 | 討論版總占比 | 則數 | 討論版總占比 | 則數 | 討論版總占比 | |
| 1 | Ptt > Gossiping | 40 | 16.9% | 21 | 27.6% | 29 | 23.0% | 0.72☆ |
| 2 | 伊莉討論區 > 時事討論(時事新聞) | 23 | 9.7% | 3 | 3.9% | 17 | 13.5% | 0.18▼ |
| 3 | 痞客邦 PIXNET > Pixnet BlogSearch搜尋結果 | 20 | 8.4% | 15 | 19.7% | 9 | 7.1% | 1.67☆ |
| 4 | PcHome部落格 > PcHome BlogSearch搜尋結果 | 17 | 7.2% | 15 | 19.7% | 16 | 12.7% | 0.94☆ |
| 5 | 捷克論壇 > 新聞討論區 | 10 | 4.2% | 1 | 1.3% | 7 | 5.6% | 0.14▼ |
| 6 | facebook粉絲團 > 黑傳媒 | 9 | 3.8% | 1 | 1.3% | 1 | 0.8% | 1.00 |
| 7 | facebook粉絲團 > 蘋果日報 | 8 | 3.4% | 1 | 1.3% | 0 | 0.0% | - |
| 8 | 卡提諾王國 > 時事新聞&爆紅話題 | 8 | 3.4% | 0 | 0.0% | 5 | 4.0% | 0.00▼ |
| 9 | facebook粉絲團 > 自由逗貢示範區 | 7 | 3.0% | 1 | 1.3% | 1 | 0.8% | 1.00 |
| 10 | Facebook公眾訊息 > FbPublicMsg | 7 | 3.0% | 2 | 2.6% | 2 | 1.6% | 1.00 |
| .... | 其他47個討論版 | 88 | 37.1% | 16 | 21.1% | 39 | 31.0% | 0.41 |
| | 總共57個討論版 | 237 | 100.0% | 76 | 100.0% | 126 | 100.0% | 0.60 |



**Fig. 10.7** Popular channels of internet media on FEPZs

fundamental insight pubic opinions exploration and agenda setting for FEPZs. Corresponding to public comments sensitive to accompanying events, the policy makers in NDC can identify the keywords and comments adopted by the netizens rapidly changing over time (Fig. 10.2). Although some critics argue that most netizens are not professionals and their opinions appear relatively useless in terms of policy relevancy and profession, the instant grasp of public concerns can at least earn more response time to public relations and crisis management. Complicated interaction of netizens' comments with policy-related events and news (Fig. 10.5) particularly initiated and released by governments also enables the policy makers to more efficiently examine the effect of agenda setting activities especially looking into the contents. The in-depth analyses regarding types and channels of Internet media help target the attentive public and policy stakeholders concerning FEPZs.

### *10.4.2 Challenges and Future Development*

Based on the previous experience and reflection, nevertheless, the IPOA results remain limited for the following aspects and therefore deserve further exploration

統計時間範圍：2014/06/01~07/14

## 前十大意見領袖

| 統計時間範圍：2014/06/01~07/14 | | 發文引發討論狀況 | | | 主文情緒強度 | | 回文情緒狀況 | | |
|---|---|---|---|---|---|---|---|---|---|
| 排名 | 討論版名稱與作者（議題） | 平均發布一則引發回文數 | 回文總數 | 發布則數 | 正評強度 | 負評強度 | 正評數 | 負評數 | P/N比 |
| 1 | facebook粉絲團_台灣加油讚<br>【DPP拒出席】 | 377 | 377 | 1 | 0.594 | 0.292 | 62 | 228 | 0.27☆ |
| 2 | Ptt_britvic<br>[新聞]徐旭東超時發言不審持國發會:抱歉持續 | 118 | 118 | 1 | 0.246 | 0.454 | 4 | 52 | 0.08▼ |
| 3 | Ptt_good900183<br>[新聞]學者：國富再談均富 | 101 | 101 | 1 | 0.226 | 0.497 | 1 | 31 | 0.03▼ |
| 4 | Ptt_prince101<br>[新聞]客運駕駛王應傑籲開放外勞 | 91 | 91 | 1 | 0.232 | 0.224 | 4 | 28 | 0.14 |
| 5 | Ptt_kwm<br>[新聞]政黑互鬥徐旭東：幼稚 | 81 | 81 | 1 | 0.268 | 0.548 | 7 | 40 | 0.18 |
| 6 | Ptt_Xhocer<br>[新聞]「保障過頭」企業不想用建教生 | 72 | 72 | 1 | 0.241 | 0.492 | 4 | 30 | 0.13 |
| 7 | facebook粉絲團_謝長廷<br>民進黨從反ECFA、反服貿、到反自經區 | 71 | 71 | 1 | 0.285 | 0.232 | 5 | 33 | 0.15 |
| 8 | Ptt_zzyyxx77<br>[新聞]徐旭東：經濟再不努力未來糾糾 | 70 | 70 | 1 | 0.415 | 0.262 | 5 | 52 | 0.10▼ |
| 9 | facebook粉絲團_朱學恒的阿宅萬事通事務所<br>訪問經貿圈是會議的藉問 | 66 | 131 | 2 | 0.464 | 0.475 | 18 | 48 | 0.38☆ |
| 10 | facebook粉絲團_中時電子報<br>總拒「經貿圈是會議」藍表遺憾 | 57 | 57 | 1 | 0 | 0.312 | 7 | 29 | 0.24☆ |

*註解：1. 意見領袖定義為每發布一則主文，引起最多回應之作者。
2.正、負評強度，大於0.3代表顯著，另外，假如發文數大於兩則以上取平均數
3. 回文情緒狀況為該作者發言引發之所有回應，依據一則主文之各篇回文合自經系統判斷之正評與負評數量；一則文章中網友回文可能出現正或負評也有正、負同時出現或是系統無法判斷之現象，因此正負評加總不等於回文總數。

**Fig. 10.8** Public opinions leaders on internet media

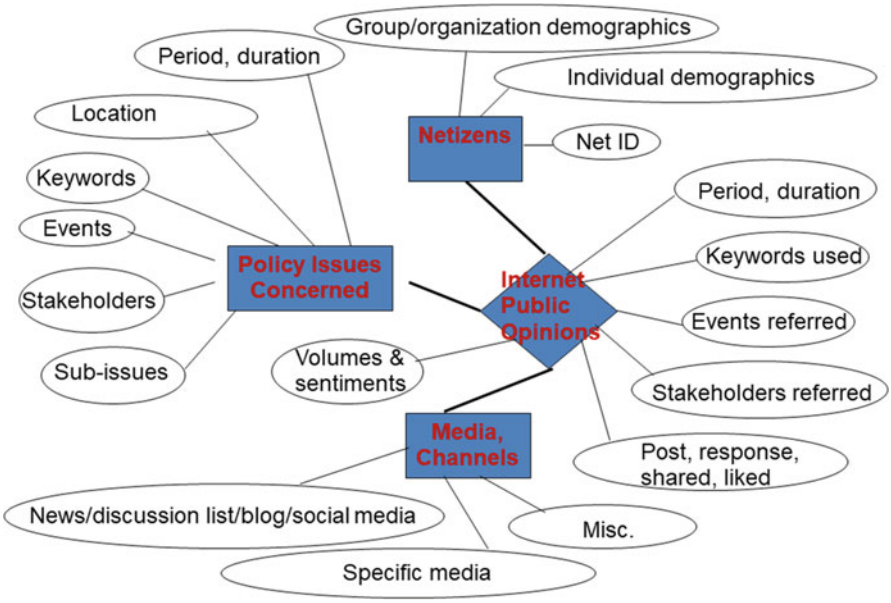



**Fig. 10.9** Entities and attributes of internet public opinions analysis

by both research and practice communities collaboratively. First of all, many practitioners participating in IPOA in our field study have raised suspicion for quality of IPOA statistics and charts in terms of reliability and validity. Some high-ranked public officials also questioned whether the public sentiments are equivalent to their preferences and positions towards the policy issue. By manually inspecting the IPOA online query results (Fig. 10.2) we actually found that around 20% of public comments are misjudged by computer programs either for their relevance or sentiments towards FEPZs. Also identified in our study were public comments with significantly negative sentiments that actually agreed, rather than disagreed, with FEPZs. The ambiguity and subtle meanings of texts and languages should be responsible for the inevitable misinterpretation by machine-learning algorithms. Although the technical providers in Big Data industry nowadays appear to accept the current level of text mining quality, they meanwhile keep endeavoring to develop more effective algorithms. Based on the field observation for the IPOA iterative process (Fig. 10.1), the present study suggests that policy makers should work with academics and IPOA service providers to improve its accuracy and precision.

Despite preliminary cross-examination of public comments and sentiments as well as the accompanying events, their causality in some cases should be more carefully inspected and confirmed especially by policy domain experts and policy makers. Concurrent correlations, strictly speaking, cannot always guarantee one-way and mutual causal relationships as confounding variables may act behind the scene without being exposed and diffused by Internet media. This also serves as a persuasive reason to support our previous arguments that policy domain experts in government should collaborate with external policy consultants and IPOA technical providers.

Lastly, the preceding IPOA implementation in government and its application to public policy analysis call for promising future research and practice to develop more comprehensive methods to collect and analyze public opinions. Telephone (including mobile phone) survey, panel survey by emails, and online survey by websites and mobile applications (apps) have been developed and conducted by various organizations. IPOA reported in the study can contribute to the arsenal of public opinions survey methods by its unobtrusiveness, updatedness, and possibility of longitudinal inspection. For example, policy officials in governments may conduct a telephone survey with more rigid sampling procedure based on the keywords and concerns extracted from the netizens' comments. Some distorted facts and biased arguments instantly found in the netizens' comments by in-depth IPOA content analyses may be included and traced by panel survey by emails over a period of time to test whether policy marketing and communication activities have achieved expected effect. While the collaborative and integrated adoption may be experimented in practice, the research community may endeavor to build up methodological and analytical foundation for IPOA playing as another contributor to public policy and governance.

## References

Chen, H., Chiang, R. H. L., & Storey, V. C. (2012). Business intelligence and analytics: From big data to big impact. *MIS Quarterly, 36*(4), 1165–1188.

Desouza, K. C. (2014). *Realizing the promise of big data: Implementing big data projects*. Washington, DC: IBM Center for the Business of Government.

Desouza, K.C. & B. Jacob (2014) Big data in the public sector: Lessons for practitioners and scholars. *Administration Society*. 1–22. doi: 10.1177/0095399714555751.

Kim, G. H., Trimi, S., & Chung, J. H. (2014). Big data applications in the government sector. *Communications of the ACM, 57*(3), 78–85.

TechAmerica Foundation (2012). Demystifying big data. Washington, DC: TechAmerica Foundation. Retrieved May 21, 2014, from http://www.whitehouse.gov/sites/default/files/microsites/ostp/big_data_press_release_final_2.pdf.

# Chapter 11
# Understanding "The User-Generated": The Construction of the "ABC Model" and the Imagination of "Digital Humanities"


**Hui-Wen Liu, I-Ying Lin, Ming-Te Chi, and Kuo-Wei Hsu**

## 11.1 Background

Facebook has become one of the most influential media for accessing news, information, and social network. In addition to fields of personal social activities, Facebook has become a critical platform for information sharing and exchange among social activists. The "Sunflower Movement," which sprouted in the March of 2014, was viewed as the best evidence to show how students spread information and get organized through Facebook (Lin 2016; Kung 2016). The academic society has made efforts to account for the role of social media in Sunflower Movement (e.g., some articles in Journal of Communication Research and Practice 6(1) and Mass Communication Research, 2015–16), showing that Facebook have had a great influence on Taiwan's political and social life.

Compared with studies concerning relationships between Facebook and political-social life, this study focuses on influence of a group of fan pages serving a social movement. Except for the fan page of the initiator "Black Island Youth Font, BIYF," a constellation of fan pages was established around the movement. What is the significance of this kind of constellation formation? Does each fan page take its own specific role? Can we penetrate deeper via footprints of participants on these fan pages? That is what we try to answer in this study.

In this study we propose a data-driven approach based on the analysis of digital footprints of users who have actedon a fan page centering on the Sunflower

H.-W. Liu (✉) · I.-Y. Lin
College of Communication, National Chengchi University, Taipei, Taiwan
e-mail: huiwen@nccu.edu.tw; iyinglin@mail2000.com.tw

M.-T. Chi · K.-W. Hsu
Department of Computer Science, National Chengchi University, Taipei, Taiwan
e-mail: mtchi@cs.nccu.edu.tw



S.-H. Chen (ed.), *Big Data in Computational Social Science and Humanities*,
Computational Social Sciences, https://doi.org/10.1007/978-3-319-95465-3_11

Movement or the issue of "anti-CSSTA." Visualization tools applied in this study are constructed based on the "ABC model" and evolved along with the need of analysis. Data-driven approach is a good complementary of traditional methods based on self-report or self-interpretation. In addition, it can help researchers to discover new puzzles or tackle new questions that are beyond the perspectives of established theories.

## 11.2 Literature Review

### *11.2.1 User Description: Traditional Social Science Methods vs. Data-Driven Approach*

Traditional social science methods, for example, online survey, in-depth interviews, and field observations, are the most common methods used for depicting Facebook users.

Wilson et al. (2012) found that Facebook's influence on social interaction is the most popular theme for researchers. Surveys and in-depth interviews are frequently adopted.

Many investigations on relationships between individuals or groups (Ellison et al. 2006; Joinson 2008; Saleh et al. 2011; Bazarova & Choi 2014; Gosling et al. 2007), including student–faculty (Mazer et al. 2009), business–employee (Binder et al. 2009), and customer–company interaction (Cvijikj et al. 2011), are conducted through in-depth interviews.

Based on in-depth interviews, Preece (2001) made a field observation of exam sociability and usability in online communities. Köhl and Götzenbrucker (2014) combined qualitative interviews and online surveys to account for networked technologies-brewed emotional cultures' potential for young users to express themselves with less restriction.

Surveys are also widely applied. Ellison et al. (2007) conducted a survey of 286 undergraduate students to consolidate a strong relation between Facebook use and social capitals. As for political participation, Lee et al. (2014) conducted a national survey in USA to test relationships between social media, social network service (SNS) network heterogeneity, and opinion polarization.

Investigations on posts put emphasis on messages and author–reader interactions. Ross et al. (2015) analyzed 1148 posts of New Zealand Members of Parliament (MPs) leading up to the 2011 general election. They found that politicians seldom open dialogues with their fans, making Facebook a platform for broadcasting.

The computational turn (Berry 2011) has fertilized data-driven approaches. In addition to data crawling, Graph API has enabled researchers to gather users' digital footprints on Facebook through developers' applications. Applications and algorithms helped researchers to deal with huge amount of data, broadening our visions and understandings of digitalized life.

From 2007 to 2012, Kosinski et al. (2015) have gathered more than 6,000,000 consenters' psychological test results and more than 4,000,000 Facebook profiles via a popular Facebook application called myPersonality. They provide these anonymous datasets to registered collaborators for research works.

Backstrom (2011) suggested that the "six degree of separation" (Milgram 1967; Granovetter 1974, 1983) has fallen to 5.28 in early 2008 and to 4.74 in early 2011. Edunov et al. (2016) reported a new low, 3.57 in 2016.

Nazir et al. (2008) invented three Facebook applications, which have collected a dataset of activity records over eight million users. This may be the first research analyzing user activities on SNS. They found a small group of users account for the majority of activity within their applications. User Locality makes no difference to user response time.

### *11.2.2 Investigating Online Activism: Keyword Mining vs. "Fan Page-Centered"*

There are several approaches to grasp the interaction between social movements and social media. Lim (2012) explored relationship between social media and Egypt political change in a wider context of media use and online activism history.

Cheng and Chen (2016) collected tweets containing at least one of the related "keywords" (24 in this study) via Twitter search API. They identified influential Twitter accounts and hyperlinks in communities speaking Chinese, Japanese, and English.

Mining Facebook posts with "keywords" can help researchers to gain a better understanding of how people think of an issue or event, which words co-occur frequently (Cheng and Chen 2016, 2014; Burgess & Bruns 2012), or to detect social events (Sayyadi et al. 2009).

Besides users' digital footprints scattering around the Facebook sea, fan pages of activist groups or campaigns have garnered public attention as well as accumulated subscription and mobilization power.

Waters et al. (2009) analyzed 275 NPOs' strategies of fan page management and evaluation of effects. Briones et al. (2011) investigated how American Red Cross builds relations via social media. This kind of studies surge in business and marketing research fields and may shed lights on activists' fan page management. For instance, Swani et al. (2013) used content analysis and HLM to analyze 1143 posts of 193 Fortune 500 Facebook accounts, exploring the relationship between message strategies and "likes." They found that including emotional sentiments is particularly effective for B2B and service marketers. Coursaris et al. (2013) have extracted seven categories and 23 subcategories from posts of Delta Airlines, Wal-Mart, and McDonald's, aiming to assist further marketing strategy designs.

The Sunflower Movement started on the night of March 18, 2014, when students and activists occupied the Taiwan parliament. They protested the inadequate pass of

CSSTA. Besides the fan page of the leading organization, Black Island Youth Front, BIYF, and two fan pages concentrating on this movement during the occupation (News e-forum and Anti-Media Monsters Youths, AMMY), at least 13 fan pages were created to support this movement.

We see the need to analyze these pages together since they interact with each other more or less, for instance, messages travelled from center (the leading organization, BIYF, or the medium, News e-forum) to second-grade information center (AMMY, for example) to satellites, Alliance of anti-CSSTA, etc.

Fan pages of "Occupy Movement" in different cities serve for its own location but share the same idea. Del Vicario et al. (2015) inspected 179 public pages and found that activities are driven by pages linked to major cities instead of geographically close pages. Unlike "Occupy Wall Street," Sunflower Movement participants gathered around one site, the parliament, to support activists who locked themselves inside. Fan pages are not differentiated according to areas (except one page in the name of a city). After the end of the occupation, legal and political struggles around CCSTA remained. Do users keep track of the process and continue paying attention to CCSTA? Adding a time dimension to analysis can help us to investigate the same.

## 11.3 Data Collection

We chose 16 fan pages into analysis. Most influential fan pages, the leading group of this "occupy parliament event," Black Island Youth Front, BIYF, a follow-up campaign called "Appendectomy Project," pro-movement medium run by a coalition of students, "News e-forum," were selected. 11 fan pages containing "CSSTA" in the titles and one in the name of "Sunflower Movement" were found through the Facebook search engine. In addition, a page named "Anti-Media Monsters Youths, AMMY" updated intensively on this on-site protest during the occupy period was also included.

Researchers applied a software program called "Pagedata" (Xiong et al. 2014) to collect records of users' activity through Graph API provided by Facebook. User records on the 16 fan pages dating from March 18th to April 11th in 2014 were collected. Facebook sets all interaction data in a fan page as "open to public," and thus we do not have to deal with the privacy issue.

4780 posts, 119,693 shares, 185,619 comments, and more than 11,900,000 likes created by 680,512 users were collected from 16 fan pages.

Among all the fan pages, most "shares" were made on BIYF (18,066 shares, 44%) and News e-Forum (17,432, 42%). These two fan pages amount to more than 80% of the shared messages. 31,046 users (11%) share posts from both fan pages.

BIYF gained 97,926 comments, accounting for 53%. News e-Forum totals 43,673 comments, accounting for 24%. Again, these two "stars" breed almost 80% of the comments.

Although a dozen of fan pages were established during the Sunflower Movement, more than 80% of the "shared" messages originated from these two information czars. In addition, most users chose these two fan pages as channels for self-expression or deliberation with others (leaving comments). It shows that information dispatching and opinion exchanging about the Sunflower Movement is quite centralized. What roles do those "satellite" fan pages play? We constructed a model and visualization tools to scrutinize this.

### *11.3.1 The ABC Model*

Constrained by the design of user interface, a Facebook user can produce four kinds of "digital footprints" in a Facebook fan page, including becoming a fan ("Like" the fan page), liking a specific post, commenting on a specific post, and sharing a specific post. In this study we highlight relationships between user participation and posts, and thus solely becoming a fan without any further action is excluded from our observation data.

To construct an index of user participation, we aggregate three indexes to compose the "ABC model." A represents "activity," summing up activities of each user within a fan page, including likes, comments, and shares. B represents "broadness," totaling the number of posts each user has participated. Likes, comments, or shares of a single post all count. C means "continuity," calculating the duration time of each user, from the first time they participated to the last time they left their digital footprints in one fan page. The ABC model is good at depicting a fan page via each fan's digital footprints. Moreover, it can be used to address the diversified functions contributing by different fan pages in a given social movement.

## 11.4 Discussion: Investigating the Constellation of Fan Pages with ABC Visualization

Collected user footprints were poured into ABC model showing that the majority of the data is (1, 1, 0), which means most users whoever left their digital footprints only visited one of these fan pages one time, with only one action (Activity = 1, Broadness = 1, Continuity = 0).

It echoes the power law found in previous studies (Nazir et al. 2008; Rheingold and Weeks 2014; Del Vicario et al. 2015; Cheng and Chen 2016). Since the powerful 20s have greater influence than the average 80s, we try to reveal their action patterns in this study.

To gain the characteristics of the "powerful 20s," we apply a 3D visualization tool based on ABC model to display their activity trend. In this primitive visualization, one dot represents a fan, who has acted at least once on that fan page. X axis exhibits

the amount of "activity," while Y axis displays the number of posts for which one has pressed "like," commented, or shared. Z axis shows the time dimension, marking the duration between one's first and last action on that fan page.

Thirteen of the 16 fan pages show a pattern of "wings-shape," displaying long and thick clusters of dots located in both low and high z axis (continuity) but sparse dots around the middle area. That means most of the "powerful 20" can be categorized into two groups, one group acts intensively within a short period while the other sticks longer and accumulates their fruitful participation slowly. Void along the middle area of the z axis indicates this polarization while the portions vary according to fan page popularity. Some fan pages have large vacancy, showing no or few fans exerting a middle time-span activity. The most popular two fan pages, BIYF and News e-forum have narrower void than the others, showing more diversities of fans' "revisiting rate." However, polarization in low and high continuity still prevails (Fig. 11.1).

### *11.4.1 The Powerful 20s: One Act for one Post in Average*

To inspect the dual relationships among Activity, Broadness, and Continuity, slices of A–B, B–C, and C–A axes were produced for comparison. We discovered that most 2D slices of A-B axes display a 1:1 ratio, indicating that most fans take one action with a single post, such as liking, commenting, or sharing. It implies most Sunflower Movement supporters on the social media are not prone to exhaust all tools of Facebook interface to interact with a post they feel interested.

Besides the majority in near-perfect correlation, there are dots scattering on the right side of the fitted line. They are the most active 20s of the powerful 20s. Power law applies again. Popular fan pages like BIYF and News e-Forum have a dense plotting along the 1:1 ratio line. Dots away from the line go more sparsely as the distance increases. Compared with News e-Forum, BIYF has more dots carrying higher value of A (activity). It shows that BIYF has more engaging fans who participate more intensively via higher action frequency or/and diverse action types in a single post.

We see that messages from the initiator of the Sunflower Movement (BIYF) have greater appeal for fans' interaction than the medium serving the movement.

Slice of the fan page of Anti-CSSTA Student Organization, ACSO displayed a digression, showing more than a dozen of dots carrying high-A and low-B values. It represents that there are some fans participating extremely deep (more than 70 times) in very few posts. Whether there are interaction-soliciting factors embedded in the posts, proposed events, or dimensions of the issue in that fan page should be asked and go further to find out. This illustrates the exploration potential of "research questions finding" of a data-driven approach (Fig. 11.2).

Three fan pages do not display the trend of polarization. "Understandable CSSTA" has too few fans to display a pattern of digital footprints (183 fans). "Appendectomy Project" and "Anti-Media Monsters Youths, AMMY" do not have

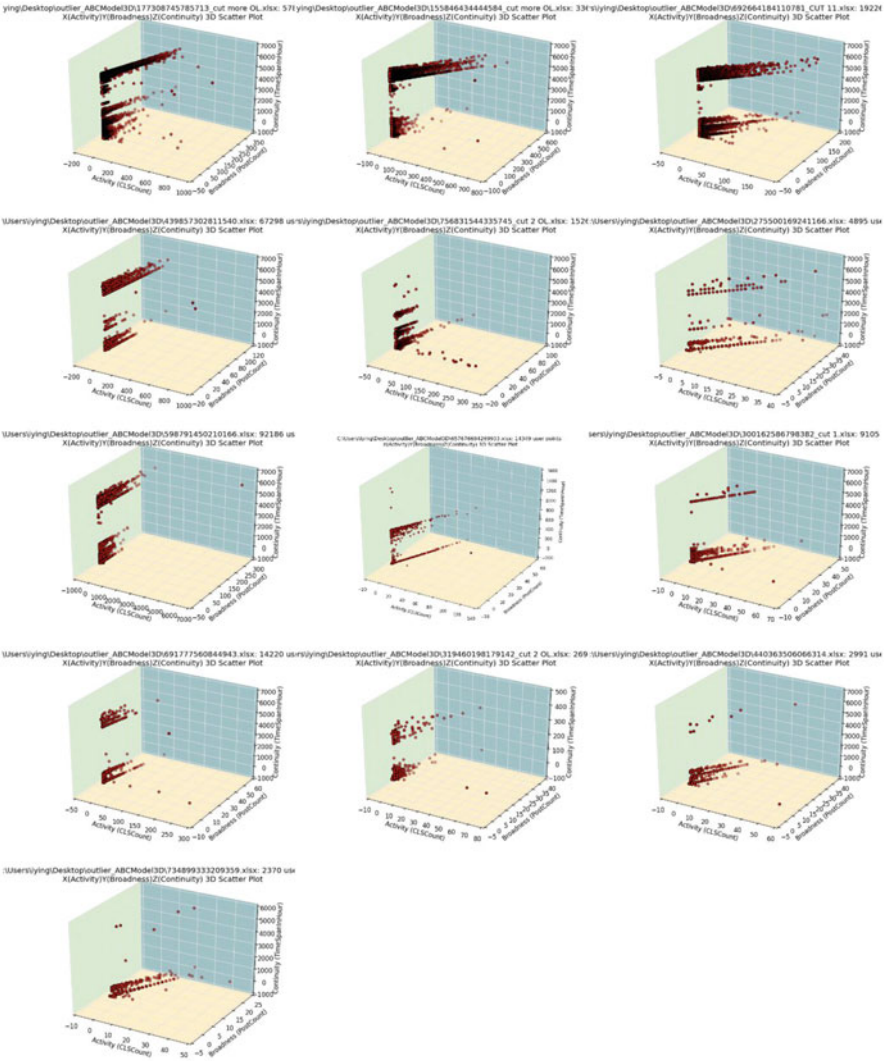


**Fig. 11.1** 13 fan pages show a pattern of “wings-shape,” displaying a polarization in continuity

vacancy in the z axis. Does it imply that these two fan pages have more loyal fans who keep participating all the time? AMMY has low values in both A and B but a uniform distribution in C. Why do not users with longer participation produce more A and B as time goes by? Are there factors discouraging participation in this fan page? We checked the A-B slice to know that the ratio is 1; thus the reasonable explanation is that the time span is too short (no more than 600 h) to have the effect of polarization. Appendectomy Project has a similar trend in graph and the time duration is less than 10 days. If we check the participation time of BIYF, duration

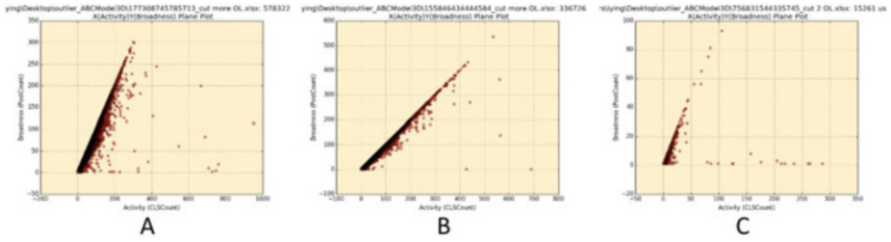



**Fig. 11.2** A–B slice of three fan pages (**a**) BIYF, (**b**) News e-forum, and (**c**) ACSO

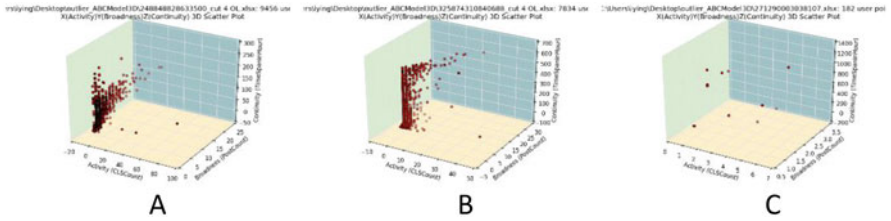



**Fig. 11.3** ABC model of three fan pages (**a**) Appendectomy project, (**b**) AMMY, and (**c**) Understandable CSSTA

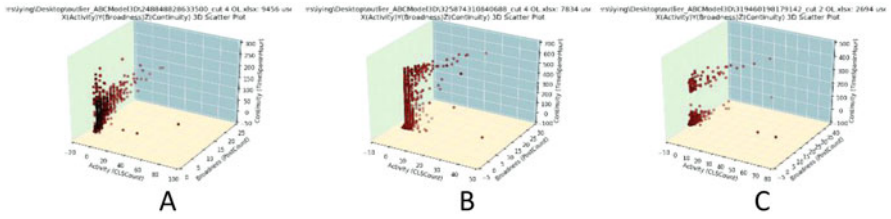



**Fig. 11.4** ABC model of three fan pages (**a**) Appendectomy project, (**b**) AMMY, and (**c**) Sunflower Movement

more than 1000 belongs to the first exploration of participation. That means AMMY and Appendectomy Project do not introduce another pattern (Fig. 11.3).

"Sunflower Movement" displays a polarization in duration so that it is grouped with other 12 fan pages. If we compare it with AMMY and Appendectomy Project, however, we will find that it is the only page with polarization in a short time span (under 700 h) (Fig. 11.4).

If we fix the scale of every fan page, we can see this easily; however unfitted scale will compress the distribution trajectory and may erase important trends. A dynamic visualization tool is needed to put various fan pages into the same scale as well as remain flexible to shift according to exploration and situated comparison conditions.

### 11.4.2 A Dynamic 3D Visualization Tool: The Calculated Display

To compare a constellation of fan pages, we developed a 3D visualization tool.

Each symbol representing a user was marked in different shape according to what fan page one acted on. Researchers can decide which fan pages to be displayed together or to be hidden (Fig. 11.5).

We applied a two-step preprocess to reduce the visual cutter cause by dense overlapped points. First, points carrying the same value are merged into a single point. For example, 100 points carrying the value of (1, 1, 0) are merged into a point (1, 1, 0) with a weight of 100. Next, we converge the neighboring point pair iteratively to overview the global distribution. We select a pair of points containing the shortest distance each iteration, and then merge them into a single point with the summed weight and the interpolated position. The convergence will continue until no more than 100 points are left.

Integrating data of 16 fan pages together with the data reduction helps us to differentiate roles played by each fan page.

Two stars, the leading organization BIYF and the supporting medium, NEF attracted most users' activities. NEF activated nearly four times of posts than BIYF but gained only 40% of its number of shares and 45% comments. It indicates that official proclamations earned more discussion and spread over than relatively neutral "news reports."

The Appendectomy Project produced relatively less posts but earned the third place in terms of sharing and the fifth place in terms of commenting. Two youth organizations, AMMY and OSSAC both ranked sixth in terms of total influence. It reveals the discussion power and spreading potentials of cohesive subgroups.

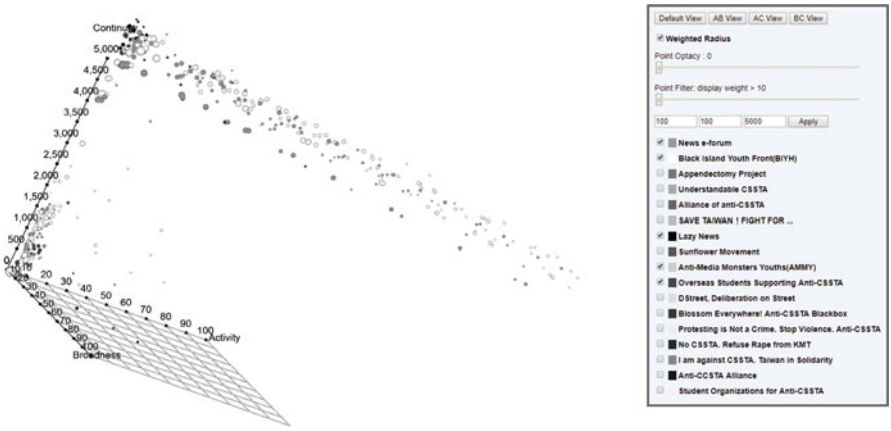



**Fig. 11.5** A visualization of anti-CSSTA fan pages via the 3D tool (http://cglabnccu.github.io/ABC-model/view/html/sunflower.html?weighted_radius=true)

## 11.5 Conclusion

In this study we have showed that dynamic visualization with calculating functions can help researchers to scrutinize the activity trends of the "powerful" minorities. Contrary to previous studies investigating the "Occupy Movement," we add a time dimension instead of examining locality significance.

Actions of liking, commenting, and sharing as well as participation time and engaged post numbers upon 16 Facebook fan pages for Taiwan Sunflower Movement were collect and recorded into the ABC model. Most fans only visit once and leave a sole footprint. The active 20s of users either compact their engagement in a short period or construct their involvement during a rather long period. A polarization of participation time is found in every fan page. Most "20s" allocate one action for one post while the 20s of the 20s exert dense participation in each post.

Through fans' action characteristics we can construct the role played by each fan page. BIYF and News e-Forum maintained heat of discussion and sharing while the second-grade, ISAC, ABC, and Appendectomy Project contribute about one-tenth of BIYF's achievement (comments and shares). Two of four young subgroups, AMMY and OSSAC rank high in participation results, showing more energy in sharing than leaving comments. Youngsters in different communities built their own pages and cultivate on them. The cohesive power deserves attention. How come some pages trigger more shares than comments and some run vice versa needs more in-depth investigation. Some small fan pages allowed fans' posts, which did attract a small group of extremely active fans to stick to them.

## References

Backstrom, L. (2011, November 22). *Anatomy of Facebook*. Palo Alto, CA: Facebook. Retrieved August 8, 2016, from https://www.Facebook.com/notes/Facebook-data-team/anatomy-of-Facebook/10150388519243859

Bazarova, N. N., & Choi, Y. H. (2014). Self-disclosure in social media: Extending the functional approach to disclosure motivations and characteristics on social network sites. *Journal of Communication, 64*, 635–657.

Berry, D. M. (2011). Thecomputational turn: Thinking about the digital humanities. *Cultural Machine*, *12*. Retrieved August 8, 2016, from https://www.culturemachine.net/index.php/cm/article/download/440/470

Binder, J., Howes, A., & Sutcliffe, A. (2009). The problem of conflicting social spheres: Effects of network structure on experience tension in social network sites. In *Proceedings of the 27th International Conference on human factors in computing systems* (pp. 965–974). New York, NY: ACM.

Briones, R. L., Kuch, B., Liu, B. F., & Jin, Y. (2011). Keeping up with the digital age: How American Red Cross uses social media to build relationships. *Public Relations Review, 37*, 37–43.

Burgess, J., & Bruns, A. (2012). (not) the twitter election: The dynamics of the #ausvotesConversation in relation to the Australian media ecology. *Journalism Practice, 6*(3), 384–402.

Cheng, Y., & Chen, P. (2014). Global social media, local context: A case study of Chinese-language tweets about the 2012 presidential election in Taiwan. *AJIM, 66*(3), 342–356.

Cheng, Y.-C., & Chen, P.-L. (2016). Online real-time civic engagement in a networked movement: A case study of Taiwan's 318 movement. *Journal of Communication Research and Practice, 6*(1), 117–150.

Coursaris, C. K., Van Osch, W., & Balogh, B. A. (2013). A social media marketing typology: Classifying brand Facebook page messages for strategic consumer engagement. *ECIS 2013 Proceedings*, AIS Electronic Library.

Cvijikj, I., Spiegler, E., & Michahelles, F. (2011). The effect of post type, category and posting day on user interaction level on Facebook. In *2011 IEEE Third International Conference on Privacy, Security, Risk and Trust Abd 2011 IEEE Third International Conference on Social Computing* (pp. 810–813). Los Alamitos, CA: CPS.

Del Vicario , M., Zhang, Q., Bessi, A., Zollo, F., Scala, A., Caldarelli, G., & Quattrociocchi, W. (2015). *Structural patterns of the occupy movement on Facebook*. arXiv Preprint arXiv, 1501.07203.

Edunov, S., Diuk, C., Filiz, I. O., Bhagat, S., & Burke, M. (2016, February 04). Three and a half degrees of seperation. *Research at Facebook*. Retrieved August 8, 2016, from https://research.Facebook.com/blog/three-and-a-half-degrees-of-separation/

Ellison, N., Heino, R., & Gibbs, J. (2006). Managing impressions online: Self-presentation processes in the online dating environment. *Journal of Computer Mediated Communication, 11*(2), 415–441.

Ellison, N., Heino, R., and J. Gibbs. (2006). Managing Impressions Online: Self-Presentation Processes in the Online Dating Environment. In *Journal of Computer Mediated Communication, 11*(2), 415–441.

Gosling, S. D., Gaddis, S., & Vazire, S. (2007). *Personality impressions based on Facebook profiles*. ICWSM'2007. Retrieved June 27, 2016, from http://icwsm.org/papers/3--Gosling-Gaddis-Vazire.pdf

Granovetter, M. S. (1974). *Getting a job: A study of contacts and careers*. Cambridge, MS: Havard University Press.

Granovetter, M. S. (1983). The strength of weak ties: A network theory revisited. *Sociological Theory, 1*, 201–233.

Joinson, A. N. (2008). Looking at, looking up or keeping up with people? Motives and use of Facebook. In *Proceedings of the 26th Annual SIGCHI Conference on Human Factors in Computing Systems* (pp. 1027–1036). New York, NY: ACM.

Kosinski, J., von Appen, A., Ori, A., Karius, K., MÜller, C. W., & Beck, M. (2015). Xlink Analyzer: Software for analysis and visualization of cross-linking data in the context of threedimensional structures. *Journal of Structural Biology, 189*(3), 177–183.

Köhl, M. M., & Götzenbrucker, G. (2014). Networked technologies as emotional resources? Exploring emerging emotional cultures on social network sites such as Facebook and Hi5: A transcultural study. *Media, Culture & Society, 36*(4), 508–525.

Kung, L.-S. (2016). A media revolution driven by 318 student movement: Reflecting on the live broadcasting mode established by students with flip-flops and Ipad. *Journal of Communication Research and Practice, 6*(1), 229–250.

Lee, J. K., Choi, J., Kim, C., & Kim, Y. (2014). Social media, network heterogeneity, and opinion polarization. *Journal of Communication, 64*, 702–722.

Lim, M. (2012). Clicks, cabs, and coffee houses: Social media and oppositional movements in Egypt, 2004–2011. *Journal of Communication, 62*(2), 231–248.

Lin, L. (2016). The practice of students are 'NTU E news forum' in the sunflower movement. *Journal of Communication Research and Practice, 6*(1), 251–269.

Mazer, J. P., Murphy, R. E., & Simond, C. J. (2009). The effects of teacher self-disclosure via Facebook on teacher credibility. *Learning, Media and Technology, 34*, 175–183.

Milgram, S. (1967) *The small-world problem*. Psychology Today, *1*, 61–67.

Nazir, A., Raza, S., & Chuah, C. (2008). Unveiling Facebook: A measurement study of social applications. In *Proceedings of the 8th ACM SIGCOMM Conference on Internet Measurement* (pp. 43–56). New York: ACM.

Preece, J. (2001). Sociability and usability: Twenty years of chatting online. *Behavior and Information Technology Journal, 20*(5), 347–356.

Rheingold, H., & Weeks, A. (2014). *Net smart: How to thrive online*. Cambridge: Mit Press.

Ross, K., Fountaine, S., & Comrie, M. (2015). Facing up to Facebook: Politicians, publics and the social media(ted) turn in New Zealand. *Media, Cuture & Society, 37*(2), 251–269.

Saleh, F., Jani, H., Marzouqi, M., Khajeh, N., & Rajan, A. (2011, October). *Social networking by the youth in the UAE: A privacy paradox*. Paper presented at the 2011 International Confernce and Workshop on Current Trends in Information Technology, Dubai, UAE.

Sayyadi, H., Hurst, M., & Maykov, A. (2009). *Event detection and tracking in social streams*. Paper presented at International Conference on Weblogs and Social Media. Retrieved August 9, 2016, from https://www.aaai.org/ocs/index.php/ICWSM/09/paper/viewFile/170/493

Swani, K., Milane, G., & Brown, B. P. (2013). Spreading the word through likes on Facebook: Evaluating the message strategy effectiveness of *Fourtune* 500 companies. *Journal of Research in Interactive Marketing, 7*(4), 269–294.

Waters, R. D., Burnett, E., Lamm, A., & Lucas, J. (2009). Engaging stakeholders through social networking: How nonprofit organizations are using Facebook. *Public Relations Review, 35*, 102–106.

Wilson, R. E., Gosling, S. D., & Graham, L. T. (2012). A review of Facebook research in the social science. *Perspectives on Psychological Science, 7*(3), 203–220.

Xiong, K.-W., Wei, H.-S., & Chi, M.-T. (2014, July). *Visualization of infoemation diffusion on social media: A case study on Facebook shareposts*. Paper presented at Computer Graphics Workshop 2014. Taipei City: National Taipei University of Technology.

# Part II
# Survey and Challenges

# Chapter 12
# Big Data Finance and Financial Markets


**Dehua Shen and Shu-Heng Chen**

## 12.1 Introduction

Financial markets have always been the most aggressive adopters of new technologies for most of their history, and thus technology has had a huge impact on financial markets. In the early nineteenth century, the communication of financial information was mainly conveyed by messengers riding on horseback. Financiers who owned private networks of horses could generate profits from exploiting the information before it spread to others. The networks of messengers on horseback were subsequently replaced by the faster "technology" of carrier pigeons. Later, pigeons were rendered obsolete by the telegraph, and the telegraph was in turn replaced by telephones. Telephone-based communications dominated the financial markets for the first 70 years of the twentieth century. After that, the shift to PC-based trading systems meant that automated trading could start to perform functions previously carried out only by traders, and computers could monitor stock prices and issue orders to buy or sell if a stock's price rose above or below a specified threshold or "trigger price." For example, prior to the introduction of the "Big Bang," daily trades on the London Stock Exchange numbered about 20,000, but within a few months of the "Big Bang" the number of daily trades had risen to an average of 59,000. These figures would have been impossible to reach without technology that could have reduced the time taken to complete a deal and handle massive volumes. Over the past 10 years, the emergence of social media has become a mainstream platform for information gathering, processing and interaction. In financial markets,

D. Shen (✉)
College of Management and Economics, Tianjin University, Tianjin, China
e-mail: dhs@tju.edu.cn

S.-H. Chen
AI-ECON Research Center, Department of Economics, National Chengchi University, Taipei, Taiwan



S.-H. Chen (ed.), *Big Data in Computational Social Science and Humanities*,
Computational Social Sciences, https://doi.org/10.1007/978-3-319-95465-3_12

social media has not only enhanced existing processes but has also created new interaction channels, e.g., a stock message board, Twitter, Facebook, and Wechat. These geometric increases in digital information have made it possible to address the predictability and dynamics of financial markets from a completely different perspective, i.e., financial big data. In that sense, the technology has already had an impact and is expected to have a truly transforming influence on financial markets.

In spite of different forms of technologies, the essential issue in financial markets is the diffusion of information to investors. From a retrospective point of view, two theories stand out prior to the efficient market hypothesis (EMH, Fama 1970) becoming prevalent. On the one hand, in response to the question of what is the problem when we try to construct a rational economic order, Hayek (1945) argued that the problem concerned how to make the best use of the knowledge in society, which was based on the assumption that individuals had perfect information and computational capacities. Therefore, the challenge regarding making the best use of knowledge is that the knowledge itself never exists in an integrated form, but is incompletely separated in individuals. Thus, the "data" used in the economic calculus is only a part of the data provided by the society. In addition, Simon (1955) further argued that the "economic man" postulated by traditional economic theory is still problematic and should take into account their limited computational and predictive ability, i.e., there exists bounded rationality.

In considering these two aspects mentioned above, it is natural to ask whether the financial big data provides us with a way to aggregate the scattered data, thereby revealing the magnitude of the investors' rationality and eventually altering the predictability of the EMH as well as the dynamics of the financial markets. This chapter aims to provide an overview of the current state of the art related to the utilization of technology, i.e., financial big data. The remainder of this chapter is organized as follows. Section 12.2 provides the reader with an understanding of the changing landscape from conventional media to big data. Section 12.3 summarizes the main empirical findings from the perspective of the medium effect on the EMH and market dynamics. Section 12.4 explains the reasons why big data might work, and Sect. 12.5 proposes potential avenues for further research.

## 12.2 Understanding Big Data

### *12.2.1 Big Data in Financial Markets*

The increasing amount of big data that reflects various aspects of human activity provides a crucial new opportunity for scholars to address fundamental research questions. Financial markets are the main battlegrounds for this quantitative investigation. Each day social media generate millions of pieces of firm-specific and market-wide information in financial markets worldwide. Einav and Levin (2014) claim that "There was five exabytes of information created between the dawn of

civilization through 2003, but that much information is now created every two days, and the pace is increasing." The efficiency of financial markets may be largely attributed to the amount of information and its diffusion process. In that sense, social media is undoubtedly playing a crucial role in financial markets. In this section, we define the concept of financial big data as consisting of both firm-specific and market-wide news, rumors, and information that appear in social media (e.g., stock message boards, search engines, microblogging, and spam emails) as well as information provided by newly emerging online information interaction channels (e.g., online investor education and online investor relationship management), which can potentially alter the formation and updating of investors' expectations, influence their decision-making and eventually have a material impact on stock prices. Compared with conventional media, financial big data have distinct characteristics whereby "former information sources" are no longer able to control the information and "former audiences" are no longer receivers of passive information. This definition is mainly derived from the perspective of complementing the "data" for understanding the dynamics of financial markets. As shown in Fig. 12.1, financial big data give rise to an additional information source that can be employed to verify financial theories and market dynamics. Other sources of information include the information gleaned from conventional media (e.g., newspapers, advertising, and television) and stock market data (e.g., stock prices, trading volume, and short-term

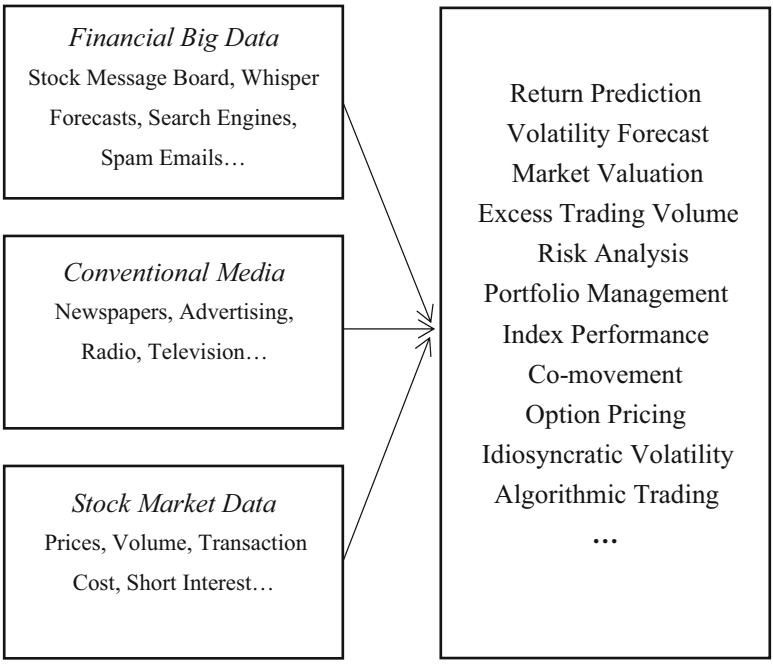



**Fig. 12.1** Definition and role of financial big data

interest rates). Only when we manage to fully integrate the financial big data, as well as information on conventional media and stock market data into the analytics and computing, can we gain a broad understanding of financial markets.

This figure illustrates the definition and role of financial big data. The definition of financial big data is obtained from the perspective of complementing the "data" for understanding the predictability and dynamics of financial markets.

### *12.2.2 The Changing Landscape: From Conventional Media to Big Data*

The crucial issue in financial markets is the way in which information is generated, diffused and utilized. A large number of empirical studies investigate the role of conventional media (e.g., newspapers, advertising, and television) in financial markets. Table 12.1 summarizes the main findings with the utilization of conventional media. We find that the majority of these studies focus on the frequency of the headlines appearing in the mainstream national newspapers (Cutler et al. 1989; Klibanoff et al. 1998; Chan 2003; Griffin et al. 2011), the expenditures on advertising (Grullon et al. 2004) or even the number of interviews given by the CEO (Meschke 2003), and only a few employ the information content (Tetlock 2007; Tetlock et al. 2008). Besides, they mainly focus on the predictability of media coverage on stock returns or on the explanation of the return premium (Fang and Peress 2009; Tetlock 2010; Tetlock 2011; Gurun and Butler 2012; Solomon 2012; Ahern and Sosyura 2014; Solomon et al. 2014; Dang et al. 2015). Table 12.2 summarizes the main findings with the utilization of financial big data. As for the methodology aspect, this has not only aroused research interest from financial economists but has also called for some interdisciplinary efforts from scholars in computer science, mathematics, complex systems, and econophysics. Besides, given the increasingly microcosmic data, scholars can now construct intraday or hourly proxies for different research purposes (Shen et al. 2016; Sun et al. 2016). More importantly, the research focus has shifted away from solely investigating the predictability of financial big data to paying more attention to the market dynamics and recognizing the crucial role in financial markets.

To sum up, investors are no longer passive information receivers, but also make explicit attempts to search for information through information outlets. This dramatically changing information environment for investors and the various types of newly emerging information are providing new insights into the dynamics of financial markets.

**Table 12.1** Main empirical findings with the utilization of conventional media

| Author | Media type | Usage form | Data frequency | Methodology | Main findings |
|---|---|---|---|---|---|
| Cutler et al. (1989) | New York Times | Frequency | Monthly | VAR | News can explain small proportion of variance in aggregate prices |
| Mitchell and Mulherin (1994) | Dow Jones | Frequency | Daily | Multivariate analysis | Number of news items is related to market activity |
| Klibanoff et al. (1998) | New York Times | Frequency | Weekly | OLS | News affects the response of closed-end fund prices to asset values |
| Liang (1999) | Wall Street Journal | Frequency | Daily | Event study | Recommendations have impact on stock prices and trading volume |
| Chan (2003) | Dow Jones | Frequency | Monthly | Portfolio construction | Drift and reversal exist after headlines |
| Meschke (2003) | CNBC | Frequency | Daily | Event study | Price and volume react to CEO interviews on CNBC |
| Grullon et al. (2004) | Advertising | Frequency | Monthly | Regression | Advertising is positively related to liquidity and ownership |
| Tetlock (2007) | Wall Street Journal | Content | Daily | VAR | Media sentiment can predict prices and trading volume |
| Tetlock et al. (2008) | DJNS and WSJ | Content | Daily | OLS | Linguistic media content captures fundamentals |
| Fang and Peress (2009) | LexisNexis[a] | Frequency | Daily | Multivariate analysis | Return premium exists on stocks with no media coverage |
| Engelberg and Parsons (2011) | ProQuest's[b] | Frequency | Daily | Multivariate analysis | Local media coverage strongly predicts local trading |

[a]LexisNexis database mainly includes the New York Times, USA Today, Wall Street Journal and Washington Post in the final sample of Fang and Peress (2009)
[b]ProQuest's newspaper database includes the Boston (Globe), Denver (Post), Detroit (News), Houston (Chronicle), Las Vegas (Review Journal), New York (Times), Pittsburgh (Post Gazette), San Antonio (Express News), San Diego (Union Tribune), San Francisco (Chronicle), Seattle (Post Intelligencer), St. Louis (Post Dispatch), St. Petersburg (Times), Minneapolis (Star Tribune), Atlanta (Journal Constitution), Sacramento (Bee), Washington (Post), and New Orleans (Times Picayune) in the final sample of Engelberg and Parsons (2011)

**Table 12.2** Main empirical findings with the utilization of financial big data

| Author(s) | Media type | Usage form | Data frequency | Methodology | Main findings |
|---|---|---|---|---|---|
| Wysocki (1998) | Yahoo! Finance | Frequency | Daily | OLS | Message-posting volume is related to firms' fundamentals |
| Bagnoli et al. (1999) | Fool and Techstocks | Content | Quarterly | Construct strategy | Whispers contain information not contained in first call forecasts |
| Tumarkin and Whitelaw (2001) | RagingBull | Frequency | Daily | VAR | Message board activity cannot predict returns or trading volume |
| Antweiler and Frank (2004) | Yahoo!/RagingBull | Content | Daily | Naive Bayes coding | Stock message board helps predict market volatility |
| Das and Chen (2007) | Yahoo! Finance | Content | Daily | Classifier | Message board sentiment index is closely related to the MSH index |
| Hanke and Hauser (2008) | Crummy | Content | Daily | Event study | Spam has significant impact on return, turnover and price range |
| Bollen et al. (2011) | Twitter | Content | Daily | NLP[a] | Twitter mood can predict the changes in closing value of the DJIA |
| Da et al. (2011) | Google Trends | Frequency | Weekly | Multivariate analysis | Search volume index (SVI) is related to returns and IPO anomaly |
| Drake et al. (2012) | Google Trends | Frequency | Daily | Multivariate analysis | SVI represents information demand and spikes at the announcement |
| Zhang et al. (2013) | Baidu Index | Frequency | Daily | OLS | Internet enhances the speed of information diffusion |
| Vozlyublennaia (2014) | Google Trends | Frequency | Weekly | VAR | Increased SVI diminishes return predictability |
| Chen et al. (2014) | Seeking Alpha | Content | Daily | OLS | Extracted opinions predict future stock returns and earnings surprises |

| | | | | | |
|---|---|---|---|---|---|
| Blankespoor et al. (2014) | Twitter | Content | Daily | Multivariate analysis | Twitter reduces information asymmetry and increases liquidity |
| Siganos et al. (2014) | Facebook | Content | Daily | VAR | Happiness index has a contemporaneous relation to stock returns |
| Zhang et al. (2014) | Baidu News | Frequency | Daily | GARCH | Internet information has a positive impact on return persistence |
| Da et al. (2015) | Google Trends | Frequency | Daily | OLS | FEARS can predict returns, volatility and mutual fund flows |
| Dimpfl and Jank (2016) | Google Trends | Frequency | Daily | VAR | SVI can Granger-cause volatility |
| Shen et al. (2016) | Baidu News | Frequency | Intraday | GARCH | Internet information can reduce return persistence at stock-level |
| Zhang, Li, et al. (2016); Zhang, Song, et al. (2016) | NetEase | Content | Daily | Event study | Market reaction to internet information with both PPH and IDH[b] |
| Sun et al. (2016) | TRMI[c] | Content | Half hour | OLS | Half-hour sentiment can predict S&P index return |

[a]NLP denotes natural language processing. Bollen et al. (2011) use the OpinionFinder and Google-Profile of Mood States (GPOMS) to extract sentiment from Tweets

[b]PPH denotes the price pressure hypothesis and IDH denotes the information diffusion hypothesis

[c]TRMI provides comprehensive sentiment data by distilling a massive collection of news and social media content. See: https://www.marketpsych.com/

## 12.3 The Medium Effect of Big Data

The aim of this section is to illustrate the main empirical findings regarding the association between financial big data and financial markets. This section focuses exclusively on the medium effect of the financial big data and is naturally divided into two subsections based on the different research focuses.

### *12.3.1 Medium Effect on the Efficient Market Hypothesis*

Stock returns are not predictable according to the efficient market hypothesis, which models random walks in stock returns. Therefore, the common starting point focuses on the utilization of the financial big data in prediction. The earliest discussion on the use of financial big data in stock market predictions may be traced back to the reports and comments that appeared in mainstream newspapers in the USA, e.g., the *Seattle Times* (Batsell 1998), *Dow Jones News Service* (Bennett 1998), *Dallas Morning News* (Goldstein 1998), *New York Times* (Harmon 1998), *Wall Street Journal* (Maremount 1998) and *Chicago Daily Herald* (Medill 1998). All of these articles raise a very intriguing topic: does financial big data have some predictive power in relation to the stock market? Even though the Computer Business Review reported that "comments posted to a Yahoo! Finance message board have so devalued a Las Vegas, Nevada-based biotech stock that the company's chief executive officer held a conference call to reassure investors and analysts that fundamentals are sound," little is known about the relationship between financial big data and the predictability of stock returns.

The simplest measurements of financial big data are the number of times certain stock names are mentioned, the search frequency of certain keywords as well as the sentiment extracted from the content. To the best of our knowledge, Wysocki (1998) was the first to document the relationship between the message-posting volume of more than 3000 stocks listed on the Yahoo! Finance Message Board and stock market performance. Wysocki found that the overnight message-posting volume could be used to predict changes in the next day's stock returns, leading to the conclusion that the message-posting volume is related to the fundamentals of underlying companies. Bagnoli et al. (1999) collected the whisper forecasts of quarterly earnings from the websites of the Wall Street Journal, fool.com and techstocks.com and found that the whisper forecasts disseminated on the Internet contained information that was not conveyed in the official forecasts (e.g., the First Call analyst forecasts). Bollen et al. (2011) extracted data on the collective mood states, including "alert," "sure," "vital," "kind," "calm," and "happy", from Twitter and found that these moods could predict the changes in the closing value of the Dow Jones Industrial Average (DJIA). Da et al. (2011) constructed a direct proxy for investor attention with the search frequency of stock names of the Russell 3000 in Google Trends from 2004 to 2008. They found that an increase

in investor attention could predict higher stock returns in the following 2 weeks and that investor attention could also explain the large debut-day return and long-term underperformance. Sabherwal et al. (2011) constructed the "online trader's credit-weighted sentiment index" from Lion.com and found that this index could predict the subsequent two trading days' returns. Zhang et al. (2012) downloaded message board postings from The Lion Wall Street Pit and showed that the sentiment generated by text classifiers was a negative indicator of the next day's stock return. Zhang et al. (2013) employed the search frequency of stock names in the Baidu Index as a proxy for investor attention in China and concluded that investor attention increased the predictive power of abnormal returns. Vozlyublennaia (2014) investigated the relationship between index performance and Google search volume, and found that there existed a significant short-term impact of increased attention on index returns, but a long-term impact of return shocks on the change in attention, and further demonstrated that the increased investor attention reduced index return predictability and simultaneously improved the market's efficiency. Other similar studies include Clarkson et al. (2006), Hanke and Hauser (2008), Zhang and Swanson (2010), Bank et al. (2011), Saxton (2012), and Vlastakis and Markellos (2012).

### *12.3.2 Medium Effect on the Market Dynamics*

The medium effect of the financial big data on the market dynamics usually relies on some financial theories. In general, there are mainly three strands of the literature that focus on this issue. The first category refers to studies that elaborate on the search data from investors as a reflection of the investors' state of mind (De Long et al. 1990; Lakonishok et al. 1994; Barberis et al. 1998; Daniel et al. 1998). Joseph et al. (2011) elaborated on the search frequency from Google Insights as "a set of beliefs about cash flow and investment risks," i.e., investor sentiment, and found that this sentiment could predict trading volume and abnormal stock returns. Following the intuition that investors react to uncertainty by intensifying their search behavior, Dzielinski (2012) expounded on the search frequency from Google Trends as reflecting economic uncertainty and found that this indicator was significantly correlated with stock returns and volatility. Irresberger et al. (2015) constructed a market-level crisis sentiment with the search volume from Google Trends and found that this sentiment could lead investors to devalue the bank stocks. Da et al. (2015) also constructed the investor sentiment using Google Trends and found that this FEARS (Financial and Economic Attitudes Revealed by Search) index could predict temporary increases in volatility, mutual fund flows, and short-term price reversals.

As for the second category, empiricists have used the adoption of a newly emerged information channel as a natural experiment to investigate the impact of an additional information diffusion channel on the diffusion of firm-specific information. Using a large sample of firms from the S&P 100, Jones (2006)

empirically found that there existed significant increases in daily trading volume and firm-specific volatility after firms established their message boards on Yahoo! Finance. Blankespoor et al. (2014) focused on firms using Twitter to send links to investors and found that this additional diffusion channel was associated with greater abnormal depths and lower abnormal bid–ask spreads, which was consistent with the theory that firm disclosures could reduce information asymmetry. As for emerging markets, Jin et al. (2016) showed that Microblogging (Sina Weibo) in China increased trading volume and decreased firm-level volatility with a sample of stocks from the Shanghai and Shenzhen CSI 300 Index. All these findings suggest that the newly emerging information diffusion channels do not just convey preexisting information to investors, but that they are also evidence of a new information dynamic that has altered the behavior of investors in financial markets.

The third category treats the analysis of return volatility as a function of the information flow proxied by financial big data. Zhang et al. (2014) employed the number of news items appearing in Baidu News as a direct proxy for information arrival, and incorporated this proxy into the conditional variance equation of the GARCH model. They found that Baidu News could explain volatility clustering. This study was further extended by Shen et al. (2016) who separated the news into trading and non-trading periods. They showed that both the lead information and the aggregate (the sum of the lead and contemporaneous) information could explain the volatility clustering of individual stocks and naturally provided alternative evidence for the Mixture of Distribution Hypothesis.

## 12.4 Underlying Mechanisms of Big Data

As we all know, there are some stylized facts in the prices of stocks (Cont 2001). A few studies have begun to observe the statistical properties of financial big data and have incorporated them into the analysis of the return volatility (Shen et al. 2016). Therefore, the most straightforward reason why financial big data have superior predictive and explanatory ability compared to conventional media is that the financial big data present similar stylized facts to those observed in relation to stock returns.

Besides, according to psychology, humans can express their own thoughts through introspection via social interaction and thus a human's mental state can influence the behavior of others. In that sense, financial big data that create new forms of information diffusion channels can serve as a way of observing the behavior of other investors. Therefore, if investors share their opinions on the same platform, their expectations may be reconciled and converge in one direction, i.e., the herding effect. Conversely, the big data platform may also parallel investors' expectations, i.e., the disagreement effect. The outcomes of both effects can make the aggregate decision-making of investors more consistent and therefore the association between financial big data and market performance is even closer.

Last but not least, we all know that the ultimate aim of financial big data is to shorten the diffusion period from the time the information is generated to when it is conveyed to investors. Unlike the single-direction information diffusion mode, financial big data has the advantage of enabling investors to promptly receive the information. Besides, the transaction could be finished with the trading system apps being directly linked to notifications from social networking software. It is very likely that investors may change from reflective thinking to reflexive thinking, i.e., investors make snap decisions without taking the time to get all the information needed. In that sense, investors receiving the same information may react in a consistent manner.

## 12.5 Conclusions

It can be reasonably inferred from the abovementioned empirical findings that big data are consistently playing a more crucial role in crystalizing our understanding of financial markets. There are already signs that the literature is focusing on constructing proxies to investigate financial theories. The big data can serve as a unique source of data for empiricists to examine untested theories. Examples include the relationships between various forms of soft information (e.g., voice, gestures, and facial expressions) based on managers' and investors' perceptions, the dynamic asset pricing model based on the propagation path of the information, as well as the impact of biased information on expectations formation.

Another promising avenue for future research would be to combine multiple sources of information. Most existing studies only adopt a single source of information. The information sharing mechanism is a means of coping with the low value density of big data. The successful implementation of the sharing mechanism requires basic research from computer scientists, mathematicians and econometricians. Meanwhile, the privacy problem should be highlighted.

It would be worth investigating the function of big data as an early warning sign of contagion or interdependence. If the distributions and dynamics of big data are correlated with the state of mind or the range of knowledge of hundreds of millions of online users, it will become possible to predict the macrobehavior emerging from investors' micromotives. To some extent, this point has somehow been ignored as scholars focus on building more sophisticated models or just notice it without delving into the underlying issue more deeply.

We also believe that a most challenging avenue of future research would be to combine the financial big data with the data generated by agent-based economic models (Chen et al. 2012; Chen and Venkatachalam 2017). These two sources of data definitely go hand in hand, and supplement and complement each other in helping to understand the dynamics of financial markets as well as engage in data-driven decision-making.

Last but not least, we firmly believe that another most challenging and promising avenue of future research would be to further investigate the dark side of the big

data in financial markets. Because the convenience of communications technology has changed the constitution of the investor structure, some "ineligible" investors without financial savvy may be lured into the financial markets and their irrational behavior may deeply hurt the investing public as a whole.

**Acknowledgement** The first author is grateful for the research support in the form of National Natural Science Foundation of China (Grant number: 71701150 and 71320107003), whereas the second author is grateful for the research support in the form of Ministry of Science and Technology (MOST) Grants, Taiwan, MOST 106-2410-H-004-006-MY2.

## References

Ahern, K. R., & Sosyura, D. (2014). Who writes the news? Corporate press releases during merger negotiations. *Journal of Finance, 69*(1), 241–291.

Antweiler, W., & Frank, M. Z. (2004). Is all that talk just noise? The information content of internet stock message boards. *Journal of Finance, 59*(3), 1259–1294.

Bagnoli, M., Beneish, M. D., & Watts, S. G. (1999). Whisper forecasts of quarterly earnings per share. *Journal of Accounting and Economics, 28*(1), 27–50.

Bank, M., Larch, M., & Peter, G. (2011). Google search volume and its influence on liquidity and returns of German stocks. *Financial Markets and Portfolio Management, 25*(3), 239–264.

Barberis, N., Shleifer, A., & Vishny, R. (1998). A model of investor sentiment. *Journal of Financial Economics, 49*(3), 307–343.

Batsell, J. (1998). Gossip central—internet message boards can leave some stocks hanging by a thread. *Seattle Times*, September 14.

Bennett, J. (1998). Traffic on financial web pages rises when the market falls. *Dow Jones News Service*, September 1.

Blankespoor, E., Miller, G. S., & White, H. D. (2014). The role of dissemination in market liquidity: Evidence from firms' use of Twitter™. *Accounting Review, 89*(1), 79–112.

Bollen, J., Mao, H., & Zeng, X. (2011). Twitter mood predicts the stock market. *Journal of Computational Science, 2*(1), 1–8.

Chan, W. S. (2003). Stock price reaction to news and no-news: Drift and reversal after headlines. *Journal of Financial Economics, 70*(2), 223–260.

Chen, H., De, P., Hu, Y., & Hwang, B.-H. (2014). Wisdom of crowds: The value of stock opinions transmitted through social media. *Review of Financial Studies, 27*(5), 1367–1403.

Chen, S.-H., Chang, C.-L., & Du, Y.-R. (2012). Agent-based economic models and econometrics. *Knowledge Engineering Review, 27*(Special Issue 02), 187–219.

Chen, S. H., & Venkatachalam, R. (2017). Agent-based modelling as a foundation for big data. *Journal of Economic Methodology, 24*(4), 362–383.

Clarkson, P. M., Joyce, D., & Tutticci, I. (2006). Market reaction to takeover rumour in internet discussion sites. *Accounting & Finance, 46*(1), 31–52.

Cont, R. (2001). Empirical properties of asset returns: Stylized facts and statistical issues. *Quantitative Finance, 1*(2), 223–236.

Cutler, D. M., Poterba, J. M., & Summers, L. H. (1989). What moves stock prices? *Journal of Portfolio Management, 15*(3), 4–12.

Da, Z., Engelberg, J., & Gao, P. (2011). In search of attention. *Journal of Finance, 66*(5), 1461–1499.

Da, Z., Engelberg, J., & Gao, P. (2015). The sum of all FEARS investor sentiment and asset prices. *Review of Financial Studies, 28*(1), 1–32.

Dang, T. L., Moshirian, F., & Zhang, B. (2015). Commonality in news around the world. *Journal of Financial Economics, 116*(1), 82–110.

Daniel, K., Hirshleifer, D., & Subrahmanyam, A. (1998). Investor psychology and security market under- and overreactions. *Journal of Finance, 53*(6), 1839–1885.
Das, S. R., & Chen, M. Y. (2007). Yahoo! for Amazon: Sentiment extraction from small talk on the web. *Management Science, 53*(9), 1375–1388.
De Long, J. B., Shleifer, A., Summers, L. H., & Waldmann, R. J. (1990). Noise trader risk in financial markets. *Journal of Political Economy, 98*(4), 703–738.
Dimpfl, T., & Jank, S. (2016). Can internet search queries help to predict stock market volatility? *European Financial Management, 22*(2), 171–192.
Drake, M. S., Roulstone, D. T., & Thornock, J. R. (2012). Investor information demand: Evidence from Google searches around earnings announcements. *Journal of Accounting Research, 50*(4), 1001–1040.
Dzielinski, M. (2012). Measuring economic uncertainty and its impact on the stock market. *Finance Research Letters, 9*(3), 167–175.
Einav, L., & Levin, J. (2014). The data revolution and economic analysis. *Innovation Policy and the Economy, 14*(1), 1–24.
Engelberg, J. E., & Parsons, C. A. (2011). The causal impact of media in financial markets. *Journal of Finance, 66*(1), 67–97.
Fama, E. F. (1970). Efficient capital market: A review of theory and empirical work. *Journal of Finance, 25*(2), 383–417.
Fang, L., & Peress, J. (2009). Media coverage and the cross-section of stock returns. *Journal of Finance, 64*(5), 2023–2052.
Goldstein, A. (1998). Money messages: Electronic message boards are a good way to get investing facts and fiction. *Dallas Morning News*, August 3.
Griffin, J. M., Hirschey, N. H., & Kelly, P. J. (2011). How important is the financial media in global markets? *Review of Financial Studies, 24*(12), 3941–3992.
Grullon, G., Kanatas, G., & Weston, J. P. (2004). Advertising, breadth of ownership, and liquidity. *Review of Financial Studies, 17*(2), 439–461.
Gurun, U. G., & Butler, A. W. (2012). Don't believe the hype: Local media slant, local advertising, and firm value. *Journal of Finance, 67*(2), 561–598.
Hanke, M., & Hauser, F. (2008). On the effects of stock spam e-mails. *Journal of Financial Markets, 11*(1), 57–83.
Harmon, A. (1998). The market turmoil: Investors on line. *New York Times*, September 1.
Hayek, F. A. (1945). The use of knowledge in society. *American Economic Review, 35*(4), 519–530.
Irresberger, F., Mühlnickel, J., & Weiß, G. N. F. (2015). Explaining bank stock performance with crisis sentiment. *Journal of Banking & Finance, 59*, 311–329.
Jin, X., Shen, D., & Zhang, W. (2016). Has microblogging changed stock market behavior? Evidence from China. *Physica A: Statistical Mechanics and its Applications, 452*, 151–156.
Jones, A. L. (2006). Have internet message boards changed market behavior? *Info, 8*(5), 67–76.
Joseph, K., Babajide Wintoki, M., & Zhang, Z. (2011). Forecasting abnormal stock returns and trading volume using investor sentiment: Evidence from online search. *International Journal of Forecasting, 27*(4), 1116–1127.
Klibanoff, P., Lamont, O., & Wizman, T. A. (1998). Investor reaction to salient news in closed-end country funds. *Journal of Finance, 53*(2), 673–699.
Lakonishok, J., Shleifer, A., & Vishny, R. W. (1994). Contrarian investment, extrapolation, and risk. *Journal of Finance, 49*(5), 1541–1578.
Liang, B. (1999). Price pressure: Evidence from the "dartboard" column. *Journal of Business, 72*(1), 119–134.
Maremount, M. (1998). Predeal trading in U.S. Surgical puts spotlight on cyberinvestors. *Wall Street Journal*, May 28.
Medill, G. (1998). Chicago firm wants to know what Yahoo! left messages. *Chicago Daily Herald*, October 12.
Meschke, F. (2003). CEO interviews on CNBC, AFA 2003 Washington, DC Meetings.
Mitchell, M. L., & Mulherin, J. H. (1994). The impact of public information on the stock market. *Journal of Finance, 49*(3), 923–950.

Sabherwal, S., Sarkar, S. K., & Zhang, Y. (2011). Do internet stock message boards influence trading? Evidence from heavily discussed stocks with no fundamental news. *Journal of Business Finance & Accounting, 38*(9–10), 1209–1237.

Saxton, G. D. (2012). New media and external accounting information: A critical review. *Australian Accounting Review, 22*(3), 286–302.

Shen, D., Zhang, W., Xiong, X., Li, X., & Zhang, Y. (2016). Trading and non-trading period Internet information flow and intraday return volatility. *Physica A: Statistical Mechanics and its Applications, 451*, 519–524.

Siganos, A., Vagenas-Nanos, E., & Verwijmeren, P. (2014). Facebook's daily sentiment and international stock markets. *Journal of Economic Behavior & Organization, 107*(Part B), 730–743.

Simon, H. A. (1955). A behavioral model of rational choice. *Quarterly Journal of Economics, 69*(1), 99–118.

Solomon, D. H. (2012). Selective publicity and stock prices. *Journal of Finance, 67*(2), 599–638.

Solomon, D. H., Soltes, E., & Sosyura, D. (2014). Winners in the spotlight: Media coverage of fund holdings as a driver of flows. *Journal of Financial Economics, 113*(1), 53–72.

Sun, L., Najand, M., & Shen, J. (2016). Stock return predictability and investor sentiment: A high-frequency perspective. *Journal of Banking & Finance, 73*, 147–164.

Tetlock, P. C. (2007). Giving content to investor sentiment: The role of media in the stock market. *Journal of Finance, 62*(3), 1139–1168.

Tetlock, P. C. (2010). Does public financial news resolve asymmetric information? *Review of Financial Studies, 23*(9), 3520–3557.

Tetlock, P. C. (2011). All the news that's fit to reprint: Do investors react to stale information? *Review of Financial Studies, 24*(5), 1481–1512.

Tetlock, P. C., Saar-Tsechansky, M., & Macskassy, S. (2008). More than words: Quantifying language to measure firms' fundamentals. *Journal of Finance, 63*(3), 1437–1467.

Tumarkin, R., & Whitelaw, R. F. (2001). News or noise? Internet postings and stock prices. *Financial Analysts Journal, 57*(3), 41–51.

Vlastakis, N., & Markellos, R. N. (2012). Information demand and stock market volatility. *Journal of Banking & Finance, 36*(6), 1808–1821.

Vozlyublennaia, N. (2014). Investor attention, index performance, and return predictability. *Journal of Banking & Finance, 41*, 17–35.

Wysocki, P. (1998). Cheap talk on the web: The determinants of postings on stock message boards. University of Michigan Business School Working Paper (98025).

Zhang, W., Li, X., Shen, D., & Teglio, A. (2016). Daily happiness and stock returns: Some international evidence. *Physica A: Statistical Mechanics and its Applications, 460*, 201–209.

Zhang, W., Shen, D., Zhang, Y., & Xiong, X. (2013). Open source information, investor attention, and asset pricing. *Economic Modelling, 33*(0), 613–619.

Zhang, Y., Feng, L., Jin, X., Shen, D., Xiong, X., & Zhang, W. (2014). Internet information arrival and volatility of SME PRICE INDEX. *Physica A: Statistical Mechanics and its Applications, 399*(0), 70–74.

Zhang, Y., Song, W., Shen, D., & Zhang, W. (2016). Market reaction to internet news: Information diffusion and price pressure. *Economic Modelling, 56*, 43–49.

Zhang, Y., & Swanson, P. E. (2010). Are day traders bias free?—Evidence from internet stock message boards. *Journal of Economics and Finance, 34*(1), 96–112.

Zhang, Y., Swanson, P. E., & Prombutr, W. (2012). Measuring effects on stock returns of sentiment indexes created from stock message boards. *Journal of Financial Research, 35*(1), 79–114.

# Chapter 13
# Applications of Internet Methods in Psychology


**Lee-Xieng Yang**

## 13.1 Introduction

The birth of the Internet has greatly changed our lives. The search engine on the Internet (e.g., *Google* and *Wikipedia*) is replacing traditional libraries; the digital shopping platform (e.g., Amazon) is replacing traditional department stores; the social media (e.g., Facebook, Twitter, and LINE) is replacing traditional paper media. Most of our needs, which in the past could be fulfilled only via specific channels, now can be done via the Internet. Thus, no one would disagree that the Internet is not only a digital device to us, but also a part of our real live. Psychologists have noticed the potential benefit of the Internet to psychological research in at least three respects: as a data-collecting platform (see Harlow and Oswald 2016), as an online database, and as the research field to uncover the mystery of human mind and behavior. The first is specifically referred to as the Web-based psychological studies. Whether or not, or to what extent, the psychological research is suitably conducted online is the first concern for transferring the laboratory-based study to Web-based study. In Sect. 13.2, I review the studies addressing this issue. In addition to behavioral studies (e.g., experiments and surveys), the database approach study is not rare in psychology. However, in the past, the kinds and numbers of databases were limited to the institutes for maintaining them. Now, with the search engine on the Internet, researchers can gain access to all Web pages of their interest around the world, which provide a bigger than ever database for research. In Sect. 13.3, the studies with the Web search engines are introduced. Since the emergence of blogs, microblogs, and social network sites, people are more and more used to sharing their live events with friends on social media. Thus, the

L.-X. Yang (✉)
Department of Psychology, National Chengchi University, Taipei, Taiwan



S.-H. Chen (ed.), *Big Data in Computational Social Science and Humanities*,
Computational Social Sciences, https://doi.org/10.1007/978-3-319-95465-3_13

social media themselves can become a research field for psychologists to observe and understand people. In Sect. 13.4, some pioneer studies in this approach are introduced. Following these three sections is the conclusion of this chapter.

## 13.2 Online Psychological Studies

As a science, one important principle of psychological research methods is that a study should be replicable. Due to the sampling error and the bigger variance of psychological phenomenon, this is a goal not that easy to achieve for psychology studies. Nonetheless, at least the psychological studies are conventionally asked to be conducted with standardized instruction, stimuli, and procedure and in well-controlled environment. However, the limited size and representativeness of sample (e.g., most participants are college students) are always the barrier to the establishment of ecological validity. When the idea of crowdsourcing emerges, on the Internet, one single task (or experiment) can be assigned to more than hundreds of workers (or participants) all over the world. This seems to be a solution to the low ecological validity of psychological studies. However, the control for the testing environment might not be as standardized as usual. The past studies suggest that the online survey indeed provides data of higher representativeness. For the online experiments, the effect size may be lower than the experiment in laboratory.

### *13.2.1 Crowdsourcing and Psychological Study*

The emergence of the Internet has brought innovations in many aspects of our life. One of them is so called the concept of crowdsourcing first proposed by Jeff Howe (2006).

> *Crowdsourcing is the act of taking a task traditionally performed by a designated agent (such as an employee or a contractor) and outsourcing it by making an open call to an undefined but large group of people.* (Howe 2008, p. 1)

It is not hard to see in this definition that the characteristic of crowdsourcing is an undefined large group of people who work independently to complete a task. The high speed and the security of exchanging information, ideas, and data files on the Internet render the plausibility of crowdsourcing. Although it may look strange to traditional business to have unrelated people independently contribute to a same project, it may be simply nature in psychological studies. Suppose we would like to verify the hypothesis that a new teaching program can effectively enhance students' performance in mathematics. To this end, we might have two groups of students, one as the control group taught with the old way and the other as the experimental

group taught with the new way. After being taught for a certain time period, both groups of students will get a test and the comparison between the average scores of these two groups can help us verify our hypothesis.

In order to prevent any confounding effect from the characteristics of participants (e.g., the control group are all males and the experimental group are all females), we will randomly sample the students from the student pool and randomly assign them to one group or the other. In the view point of crowdsourcing, the experiment is just the project or the task to accomplish. As the students are ideally to be independent to each other[1], the participants in an experiment can be viewed as the undefined group of people in crowdsourcing. Also, it is acknowledged, in order to get the power of statistical analysis large enough, the more participants the better. According to the definition of crowdsourcing, the group size must not be small. Thus, running a psychological experiment is quite similar to doing a project in the way of crowdsourcing, except that the former is traditionally conducted in the laboratories and the latter on the Internet.

In fact, psychologists have begun to collect data online. For instance, on this website (the URL is https://www.socialpsychology.org/expts.htm) maintained by Scott Plous, Wesleyan University, you can find more than 500 online experiments, surveys, and other social psychology studies. By clicking on the title of the study in which you are interested, you will be directed to the instruction Web page of that study and the test will begin after you sign up the informed consent by clicking the button of "I agree." In addition to the website maintained personally, some enterprise also provides similar online platform for the need of crowdsourcing. Amazon Mechanical Turk (MTurk) is a good example, which can be accessed on https://www.mturk.com/mturk/welcome. The slogan on this Web page is as follows.

> *Mechanical Turk is a marketplace to work. We give businesses and developers access to an on-demand, scalable workforce. Workers select from thousands of tasks and work whenever it's convenient.*

It is clear that MTurk is a marketplace where some people as the *requesters* can post their signup sheet/job advertisement to recruit workers to temporarily work for them, while some other people can come and join in the project as the *workers*. On MTurk, *requesters* can list tasks (called "human intelligence task," or HITs) along with a specified compensation. HITs range widely in size and nature, requiring from seconds to hours to complete, and compensation varies accordingly ($0.01 $0.1 per HIT). After seeing the preview of the task, *workers* can choose to accept this HIT and complete the task. Of course, *workers* will get paid according to their performance. This way of recruiting *workers* resembles quite much the way to recruit participants in a psychological study. Apparently, the spirit of crowdsourcing is embodied on MTurk.

[1] Random sampling of participants.

### 13.2.2 Internet Methods as a Medium for Collecting Human Data?

Since conducting a psychological study is a kind of crowdsourcing, it seems to be proper to do psychological studies on MTurk or similar websites. On the one hand, this idea might be workable in terms of the need to quickly get a large random sample of subjects. On the other hand, this might be a dangerous idea in terms of the concerns about the quality of collected data (e.g., Mezzacappa 2000). After all, psychological surveys and experiments both demand the administration of testing and experimental procedure to be standardized and well controlled, that presumably challenges the Web-based testing. Indeed, there is a debate on whether MTurk or similar websites can be a participant pool in the community of psychology. Let us first see some positive evidence.

Gosling et al. (2004) compared the data on personality measures of Internet samples ($N = 361{,}703$) who are the visitors to the noncommercial advertisement-free website, outofservice.com, and 510 traditional samples published in the *Journal of Personality and Social Psychology* with respect to several domains, such as gender, race, socioeconomic status, geographic region, age, and so on. Their findings suggested that Internet samples are more representative than traditional samples with respect to gender, socioeconomic status, geographic location, and age, and are about as representative as traditional samples with respect to race.

Also, Buhrmester et al. (2011) administered personality questionnaires via MTurk to evaluate the quality of data collected in this way. The main findings include that (1) MTurk participants are significantly more diverse than typical American college samples; (2) although the compensation rate and task length will affect participation, participants can still be recruited rapidly and inexpensively; (3) the compensation rates do not affect data quality; and (4) the data obtained are at least as reliable as those obtained via traditional methods. In addition to the administration of questionnaires, economic experiments can be run on MTurk. Amir et al. (2012) conducted economic game experiments on MTurk (dictator game, ultimatum game, trust game, and public goods game) and found that the results were consistent with previous research conducted in the physical laboratory. Similarly, Mason and Watts (2009) examined the relationship between financial incentives and performance by an MTurk experiment, in which participants were asked to sort the given images in chronological order in one of the conditions crossed by four levels of task difficulty and four levels of compensation. How many sets of images the participants sorted is the measure of quantity and their sorting accuracy is the measure of quality. The results showed that increased financial incentives increase the quantity, but not the quality of participants' performance. Another economic game experiment, public goods game, was found to yield a decreasing trend on subjects' contribution over rounds, which was consistent with the findings of the traditional lab-based experiments (Suri and Watts 2011). Even the Cloze sentence completion task on MTurk showed a high correlation (rho = .75) on word predictability to the lab experiment (Schnoebelen and Kuperman 2010).

### 13.2.3 MTurk as a Platform for Conducting Psychological Experiments

It might still not be sufficient with the above instances to resolve the worries about running psychological experiments online, as most of the psychological experiments demand a high precision (up to milliseconds) in reaction time recording and stimulus presentation, which is not covered by the above instances. However, the current Web browser technology (such as HTML with Java script) does afford millisecond timing functions, which theoretically can fit the need of psychologists. In fact, Crump et al. (2013) endorsed the validity of the cognitive experiments on MTurk via replicating a wide variety of classic cognitive experiments. These authors run on MTurk the Stroop task (MacLeod 1991; Stroop 1935) and found the same response pattern reported in the traditional experiments, that the reaction time of naming a word's color is faster when the word's name is congruent with its color than when the word name and word color are incongruent. Similarly, these authors also replicated on MTurk the cost on switching tasks (see Jersild 1927; Monsell 2003), that is, that people spend a longer reaction time when the to-be-done task is changed to another one. Again, these authors even showed that the Flanker task[2] (Eriksen 1995; Eriksen and Eriksen 1974) and the Simon task[3] (Craft and Simon 1970; Lu and Proctor 1995) were of no problem to run on MTurk.

Crump et al. (2013) further demonstrated that some effects found with rapid stimulus presentation can be replicated on MTurk, such as the visual cuing effect that target is identified more quickly with a correct cue only when the cue-to-target time interval is a small, attentional blink that the detection of the second target will be impaired when the second target appears within 100 500 ms of the first one (Raymond et al. 1992; Shapiro and Raymond 1997), and the marked priming effect that the detection of the probe will be enhanced if the prime is the same as it and vice versa. However, when the task becomes complex, more caution is needed for using MTurk. Crump et al. (2013) tried to replicate the difficulty gradient of learning the six category structures first proposed by Shepard et al. (1961). The stimulus consisted of three dimensions: shape, color, and size as shown in Fig. 13.1.

The Type I problem is the easiest to learn, as the two categories (circles and squares) can be perfectly separated by attending to shape only. The Type II is the second easiest, followed by Type III, Type IV, Type V, and Type VI. The Type VI problem is the hardest, as the categorization rule nonlinear and consists of all three dimensions. This difficulty gradient had been replicated by many studies (e.g., Kruschke 1992; Lewandowsky 2011; Love 2002; Nosofsky et al. 1994). Although in the data of Crump et al. (2013) online experiment, the Type I and Type VI were the easiest and the hardies problems, the error curves for other types (specifically Type

[2]The reaction time of identifying the target alphabet in a line of alphabets becomes longer when the target is different from the others and vice versa.

[3]The reaction time is shorter when the spatial compatibility between stimulus and response key is held and vice versa.

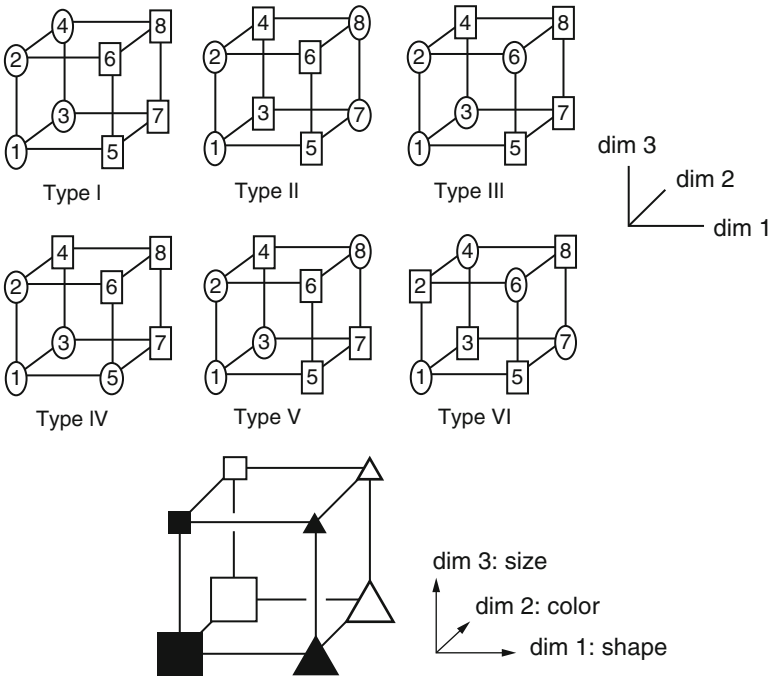



**Fig. 13.1** Six types of category learning problems and the stimulus dimensions used in Shepard et al. (1961), quoted from Nosofsk et al. (1994)

II and Type IV) did not match the pattern reported in the precedent studies. These authors initially suspected that the inconsistency came from weak motivation, so they manipulated the incentive level (i.e., raising the payment up to $2.00 and an extra bonus up to $2.50 based on task performance), but the learning patterns for Type II and Type IV were still inconsistent with the precedent studies. As argued by Kurtz et al. (2012) that the Type II advantage can be explained by the extent to which instructions emphasize verbal rules, the observed inconsistency might result from the understanding direction of participants about the instruction. To sum up, when the task is simple (i.e., short in time or simple in experimental design), such as those reaction time or attention and perception tasks, there is no problem to use MTurk; when the task is complex (i.e., complex in experimental design or sensitive to the understanding of instruction), it needs caution to conduct the experiment on MTurk. Nonetheless, the payment does influence the rate of joining in the experiment, but not the quality of performance.

### *13.2.4 Second Thoughts for MTurk as a Participant Pool*

Although the above reviewed studies point up the positive aspect of Web-based studies, some researchers otherwise have warned us of the potential crises in this approach. Among those warnings, sample quality and data quality are the main concerns. For the sample quality, the participants on MTurk often engaged in other affairs (e.g., reading MTurk blogs, listening to music, watching TV, or even chatting online) while doing a HIT (see Chandler et al. 2014). Thus, we do have reasons to doubt that the MTurk experiment is equivalent to those in laboratories. Also, cross talk happens among the participants on MTurk as *workers* can read the MTurk blogs or relevant forums to exchange information about HITs. These discussions are more about the pay rates or the reputation of *requesters*, not too much about the content of a particular HIT. Perhaps the duplication of participants across tasks is the most serious part of sample quality on MTurk. Normally, the researchers assume that the participants are naïve to the research materials because they are from a large participant pool or they have limited exposure to the research. Chandler et al. (2014) showed that some people have been the *workers* on MTurk for several years and are more likely than others to be sampled. Past research has noted that response to psychological measures correlate with proxies of prior participation in similar experiments, such as memory of prior participation (Greenwald and Nosek 2001) and memory of chronological order of studies themselves (Rand et al. 2014). Worse, the non-naïveté of participants can reduce the effect size of experimental treatment (Chandler et al. 2015).

The major concern about data quality resides on the prevalence of data exclusion. Chandler et al. (2014) did a meta-analysis for hundreds of papers published prior to December, 31, 2011, which conducted MTurk experiments and estimated that about one-third of these papers dropped *workers* post hoc for one reason or another. This high ratio is indeed a worry to those who want to conduct an MTurk experiment.

Although the issue of duplication of participants is annoying, it is not impossible to deal with. Chandler et al. (2014) suggested that Amazon Qualifications can be used as a tool for prescreening *workers*. *Requesters* can set up criteria in Qualifications for filtering *workers*, such as female only or the maximum number of HITs before the current task. Accordingly, we can select participants prior to the execution of task. Also, in virtue of a better control for the source of sample provided by Qualifications, it is likely to lower the exclusion rate of data. Thus, whether or not MTurk can be used for conducting psychological studies is not a simple yes/no question.

The Web-based (or MTurk) research indeed has some advantages that are better than the traditional studies, such as that it makes inexpensive and fast the recruitment of participants as well as it provides a relatively diverse sample. However, there is no such thing as a free lunch. It is worth noting that the validity of Web-based experiments might be reduced, due to the duplication of participants or/and inappropriate trimming of data post hoc. Nonetheless, as long as we consider

the nature of our task and do necessary prevention in advance (e.g., setting up Qualifications), this kind of research is still a viable choice for psychologists.

## 13.3 Psychological Studies with Online Search Engine

In addition to collecting data from real human beings, psychological studies can also be done with the established databases. For instance, the psycholinguistic study often relies on the corpus, which provides information about vocabulary, such as word frequency and word types. However, maintaining a database costs a lot of time and money. Also, it is not guaranteed to have a database matching our research interest. Thus, the feasibility of database approach research is constrained. Since the Internet is composed of billions of Web pages and each of which can be treated as a source of data, a straightforward idea is why not treat the whole Internet as the gigantic database for research. There are two types of studies embodying this idea. First, the search engine is used to provide data from Web pages. Second, the database of the search engine is the target for research. Both types of studies unveil a new landmark of database approach research in psychology.

### *13.3.1 Search Engine on the Internet as Research Tool*

In social sciences, database research has long been a typical way to provide a large-scale description for the topic of interest. Different databases contain data for different topics, such that the database of the World Bank contains data about economic indices of nations all over the world or that WordNet contains English nouns, verbs, adjectives, and adverbs which are structurally grouped as a useful tool for computational linguistics and natural language processing. A database is normally established and maintained by governments, academic institutions, or enterprises for public use. Most of the databases are established for a specific topic. For instance, WordNet is specifically established as a corpus of English. If we want to get the statistics of the number of vocabularies an English speaker would have, WordNet is a suitable choice. However, if we want to know GDPs of all nations in the world, we definitely will not go with WordNet. In addition to the specificity, sometimes a database can be accessed only by those who have been approved. It is noted that for the consideration of information security, some databases can be accessed only by the approved users.

In addition to the databases established for specific purposes, the search engine on the Internet, such as *Google* and Wikipedia, might also be a research tool, just like the database. However, with a number of essential differences, the search engine is not simply a database. First, the data contained in a regular database are structuralized, but the data in the Web pages, which can be accessed by the search engine, are not. For instance, the Web pages found by *Google* for our needs are often

metadata which cannot be used directly until being sorted out by suitable processes. Although the Wikipedia may be more structuralized with clear catalogs and indices, what the users can get on it basically are the Web pages and the contents of those Web pages are the source of data, not just the data.

Second, the search engine, especially *Google*, can provide a far wider variety of data than the regular database. Unlike the regular database which is established by a group of specialists following the coding instructions, the Web pages can be created by anyone all over the world. Also, there is no constraint on the format (e.g., spread sheet or a plain text), type (e.g., text or non-text), or length of the content of a Web page. Thus, the data transferred from the Web pages can be far more divergent than those structuralized data in a normal database. Third, the search engines can provide us not only the Web pages meeting our criteria but also the information about other people's search for those Web pages. For instance, the Web pages on the first page of the search results with *Google* are more likely searched by other people also. The normal database cannot provide such information. Some researchers in psychology and social sciences have started to apply *Google* and Wikipedia to research and gained quite inspiring results.

### *13.3.2 Research with Google and Wikipedia*

Stewart et al. (2005) proposed a simple model, the decision-by-sampling (DbS) model, to account for why the descriptive psychoeconomic functions take the form that they do. The idea of the DbS model is that people's decision for an attribute value depends on the relative rank of it in a random sample from their memory. These authors assumed that the contents of memory reflect the structure of the world. They also showed that the distribution of time in our memory can be gained by googling different strings which represent different time periods (e.g., 1 day, 2 days, ..., 1 year) and accumulating the counts of articles searched by *Google*. This distribution as other utility functions in Economics follows a power function. Similarly, Olivola and Sagara (2009) asked participants to estimate the frequencies of natural and industrial disasters given different number of deaths. The aggregated data show a function following a power function. These authors also searched with *Google News Archives* for the given numbers of death tolls (e.g., "10 people died"). The results showed that the distribution of fatalities from natural and industrial disasters made from *Google* search is quite similar to the one of real records and the participants' estimation followed a similar power function. These studies supported the idea that the result of *Google* search can reflect the structure of the real world and represent the contents of our memory.

Psychologists have long acknowledged that human beings are actually not as rational as assumed in economic theories on making their decision and instead we rely on many types of heuristic (Tversky and Kahneman 1974) to help us make decisions. These heuristics quite often lead us to fallacy in making our decisions. One of the famous fallacies is called the conjunction fallacy (Tversky and

Kahneman 1983). The typical problem is: "Linda is 31 years old, single, outspoken and very bright. She majored in philosophy. As a student, she was deeply concerned with issues of discrimination and social justice, and also participated in anti-nuclear demonstrations." The decision-makers are then asked whether she is more likely to be a bank teller or a feminist bank teller. Since the event of "bank teller" covers the event of "bank teller *AND* feminist," the probability for Linda being a bank teller should be higher than being a banker teller and feminist. However, people tend to choose the latter as their answer. Bhatia (2015) developed a heuristic judgment model which is a neural network model learning approximately 3.2 million articles on Wikipedia to recover the co-occurrence structure of 300,000 words. If two words appear in one article, then the frequency of co-occurrence for them is 1. Based on the co-occurrence structure implicitly retained in the articles on Wikipedia, this model predicts 67% of this conjunction fallacy and this result is positively correlated with another study with human participants ($r = .29$, Shafir et al. 1990).

More surprisingly, the search engine data can be used as a predictor of human action. Preis et al. (2013) found that words were classified as more financially relevant were more strongly related to subsequent stock market moves. Moat et al. (2013) even showed that increases in views of Wikipedia pages relating to companies listed in the Dow Jones Industrial Average or to more general economic concepts, tended to be followed by stock market drops.

## 13.4 Psychological Studies with Social Media

The emergence of social media extends our real (or offline) life to the virtual (or online) networks. Although there is not too much research so far, researchers have started to investigate the influence on people brought from social media. One interesting issue is whether the algorithm of social media to order the posts would manipulate people's attitude toward particular issue. The current evidence seems to support that the polarization of people's attitude does not result from the algorithm but people themselves. However, people's mood is indeed easily influenced by their friends even via their posts on social media. In addition to social influence, how people's footprints on social media can reflect their psychological construct is another interesting topic. Some research has started this quest and gained interesting findings. Although these cases are relevant to social and personality psychology, it can be reasonably expected that more studies with social media will be proposed in many other areas of psychology.

### *13.4.1 Social Influence on Individuals in Social Media*

It is acknowledged that we are influenced by important others in our social circles or the media and it becomes more salient on social media, such as Facebook. Some

researchers suspected that Facebook's social endorsement algorithm has something to do with the polarization of opinions and selective exposure on the media (Messing and Westwood 2012). However, some others defensed for the Facebook algorithm and showed that individuals' choices play a bigger role than the algorithm on selecting news (Bakshy et al. 2015). Nonetheless, it is of no doubt that we are living in a world without any influence from others, no matter whether the social endorsement algorithm exacerbates the fragmentation of the citizenry.

Another interesting case about the influence through the social media is the study of Coviello et al. (2014) on emotional contagion. These authors identified the emotional expression of millions of posts of Facebook users in terms of the words used in the posts. The regression results showed that the individual-specific factor (e.g., some people are always happier than others), the exogenous factor (e.g., rainfall), and the influence from the user's Facebook friends are valid predictors to the emotional expression of post. Generally speaking, when it rains, we feel sad. What these authors found gives us a surprising story, that even though there is no rain in my town, I could express negative emotion if my friend feels sad due to the rain in his/her town. The emotion contagion through social media is one instance of how psychologists can investigate individuals in a social context via social media. Of course, other topics in social psychology (e.g., group thinking) should be able to study via social media.

### *13.4.2 Private Traits are Predictable from Digital Records*

Psychologists always seek the valid criteria of the psychological traits or constructs. These criteria can be the items in a personality test, which can be a sentence or an adjective for people to judge to what extent the description matches their situation. For instance, a sentence that I like to go out with friend might be a criterion for extraversion.

Kosinski et al. (2013) investigated whether the digital records of human behavior can be used to estimate the personal attributes. These authors used the participants' Facebook likes as digital record of their behavior on Facebook. Also, these participants' scores on the famous Big-5 personality test were collected, using the questionnaire in the International Personality Item Pool (IPIP). These authors first constructed a matrix to represent the users and their likes. In order reduce the dimensionality of this matrix, the linear algebra technique SVD (Singular Value Decomposition) was applied to transfer the user-like matrix to user-component matrix. Subsequently, with this user-component matrix as a predictive variable, logistic regression model or linear multiple regression was established to predict the users' psychodemographic profiles. It was shown that their model can correctly discriminate between homosexual and heterosexual men, African Americans and Caucasian Americans, and between Democrats and Republicans. For the personality trait "Openness," the prediction accuracy is close to the test–retest accuracy of a

standard personality test. Therefore, it is supported that the digital records on social media can predict some personal traits that people would typically assume to be private.

## 13.5 Conclusion

There are basically three approaches psychologist would take to apply the Internet methods to their study. The first is using the Web environment as a new platform for recruiting participants and conducting studies. The advantage of this approach is that we can normally get the participants quickly and inexpensively. However, duplication of participants and high exclusion rate of data post hoc would be the cost. Whether or not a study is suitable for the Web environment depends on the nature of it (the type of dependent variable, prior knowledge interference, etc.). Instead of behavioral studies, the second approach goes for the search engine in the Internet (e.g., *Google* and Wikipedia) and can be regarded as a database study. With the gigantic amount of materials on the Internet as the representation of our knowledge about the world, it becomes possible to examine the mental representation and processing (memory, categorization, decision-making, etc.) at the population level. Different from the traditional experimental design, this approach requires some Internet techniques for dealing with the texts on Web pages (html coding, Web crawler, text mining, latent semantic analysis, etc.). Finally, the third approach provides a possibility to examine individual actions in social network. Comparing with the first two, this approach is more ambitious and requires more considerations on with respect to the procedure of data collection and the analysis and interpretation of data. Specifically, the related ethical issue about collecting the data of users on social media needs thorough considerations. Nonetheless, the stand of this chapter is that the Internet methods, just like any other research techniques (e.g., EEG or fMRI), can contribute to our studies of human beings, as long as the research topic, target behavior, and testing conditions are appropriate.

## References

Amir, O., Rand, D. G., & Gal, Y. (2012). Economic games on the Internet: The effect of $1 stakes. *PLoS ONE, 7*(2), e31461. https://doi.org/10.1371/journal.pone.0031461.

Bakshy, E., Messing, S., & Adamic, L. (2015). Exposure to ideologically diverse news and opinion on Facebook. *Science*, 1–5.

Bhatia, S. (2015). The power of the representativeness heuristic. In D. C. Noelle, R. Dale, A. S. Warlaumont, J. Yoshimi, T. Matlock, C. D. Jennings, & P. P. Maglio (Eds.), *Proceedings of the 37th Annual Meeting of the Cognitive Science Society* (pp. 232–237). Austin, TX: Cognitive Science Society.

Buhrmester, M., Kwang, T., & Gosling, S. D. (2011). Amazon's Mechanical Turk: A new source of inexpensive, yet high-quality, data? *Perspectives on Psychological Science, 6*, 3–5.

Chandler, J., Mueller, P., & Paolacci, G. (2014). Nonnaïveté among Amazon Mechanical Turk workers: Consequences and solutions for behavioral researchers. *Behavior Research Methods, 46*, 112–130. https://doi.org/10.3758/s13428-013-0365-7.

Chandler, J., Paolacci, G., Peer, E., Mueller, P., & Ratliff, K. (2015). Non-naïve participants can reduce effect sizes. In K. Diehl & C. Yoon (Eds.), *NA—advances in consumer research* (Vol. 43, pp. 18–22). Duluth, MN: Association for Consumer Research.

Coviello, L., Sohn, Y., Kramer, A. D. I., Marlow, C., Franceschetti, M., Christakis, N. A., et al. (2014). Detecting emotional contagion in massive social networks. *PLOS ONE, 9*, e90315. https://doi.org/10.1371/journal.pone.0090315.

Craft, J. L., & Simon, J. R. (1970). Processing symbolic information from a visual display: Interference from an irrelevant directional cue. *Journal of Experimental Psychology, 83*, 415–420.

Crump, M. J. C., McDonnell, J. V., & Gureckis, T. M. (2013). Evaluating Amazon's Mechanical Turk as a tool for experimental behavioral research. *PLoS ONE, 8*(3), e57410. https://doi.org/10.1371/journal.pone.0057410.

Eriksen, B. A., & Eriksen, C. W. (1974). Effects of noise letters upon the identification of a target letter in a nonsearch task. *Perception & Psychoophysics, 16*, 143–149.

Eriksen, C. W. (1995). The flankers task and response competition: A useful tool for investigating a variety of cognitive problems. *Visual Cognition, 2*, 101–118.

Gosling, S. D., Vazire, S., Srivastava, S., & John, O. P. (2004). Should we trust web-based studies? A comparative analysis of six preconceptions about Internet questionnaires. *American Psychologist, 59*, 93–104.

Greenwald, A. G., & Nosek, B. A. (2001). Health of the implicit association test at age 3. *Zeitschrift für Experimentelle Psychologie, 48*(2), 85–93.

Harlow, L. L., & Oswald, F. L. (2016). Big data in psychology: Introduction to the special issue. *Psychological Methods, 21*, 447–457.

Howe, J. (2006). The rise of crowdsourcing. *Wired Magazine, 14*(06), 1–5.

Howe, J. (2008). *Crowdsourcing: Why the power of crowd is driving the future of business*. New York: Crown Publishing Group.

Jersild, A. T. (1927). Mental set and shift. *Archives of Psychology*, *14*, (Whole No. 89).

Kosinski, M., Stillwell, D., & Graepel, T. (2013). Private traits and attributes are predictable from digital records of human behavior. *PNAS, 110*, 5802–5805.

Kruschke, J. K. (1992). ALCOVE: An exemplar-based connectionist model of category learning. *Psychological Review, 99*, 22–44.

Kurtz, K. J., Levering, K. R., Stanton, R. D., Romero, J., & Morris, S. N. (2012). Human learning of elemental category structures: Revising the classic result of Shepard, Hovland, and Jenkins (1961). *Journal of Experimental Psychology: Learning, Memory, and Cognition*, Online first publication. doi:https://doi.org/10.1037/a0029178

Lewandowsky, S. (2011). Working memory capacity and categorization: Individual differences and modeling. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 37*, 720–738.

Love, B. C. (2002). Comparing supervised and unsupervised category learning. *Psychonomic Bulletin & Review, 9*, 829–835.

Lu, C., & Proctor, R. W. (1995). The influence of irrelevant location information on performance: A review of the Simon and spatial Stroop effects. *Psychological Bulletin & Review, 2*, 174–207.

MacLeod, C. M. (1991). Half a century of research on the Stroop effect: An integrative review. *Psychological Bulletin, 109*, 163–203.

Mason, W., & Watts, D. J. (2009). Financial incentives and the performance of crowds. In *Proceedings of the ACM SIGKDD Workshop on Human Computation* (pp. 77–85). New York: ACM.

Messing, S., & Westwood, S. J. (2012). Selective exposure in the age of social media: Endorsements trump partisan source affiliation when selecting news online. *Communication Research, 41*, 1042–1063. https://doi.org/10.1177/0093650212466406.

Mezzacappa, E. (2000). Letter to the Editor. *APS Observer, 13*, 10.

Moat, H. S., Curme, C., Avakian, A., Kenett, D. Y., Stanley, H. E., & Preis, T. (2013). Quantifying Wikipedia usage patterns before stock market moves. *Scientific Reports, 3*, 1801.

Monsell, S. (2003). Task switching. *Cognitive Science, 7*, 134–140.

Nosofsk, Y. R. M., Gluck, M. A., Palmeri, T. J., McKinley, S. C., & Glauthier, P. (1994). Comparing models of rule-based classification learning: A replication and extension of Shepard, Hovland, and Jenkins (1961). *Memory & Cognition, 22*, 352–369.

Nosofsky, R. M., Palmeri, T. J., & McKinley, S. C. (1994). Rule-plus-exception model of classification learning. *Psychological Review, 101*, 53–79.

Olivola, C. Y., & Sagara, N. (2009). Distributions of observed death tolls govern sensitivity to human fatalities. *Proceedings of the National Academy of Sciences, 106*, 22151–22156.

Preis, T., Moat, H. S., & Stanley, H. E. (2013). Quantifying trading behavior in financial markets using Google Trends. *Scientific Reports, 3*, 1684.

Rand, D. G., Peysakhovich, A., Kraft-Todd, G. T., Newman, G. E., Wurzbacher, O., Nowak, M. A., et al. (2014). Social Heuristics shape intuitive cooperation. *Nature Communications*.

Raymond, J. E., Shapiro, K. L., & Arnell, K. M. (1992). Temporary suppression of visual processing in an RSVP task: An attentional blink. *Journal of Experimental Psychology: Human Perception & Performance, 18*, 849–860.

Schnoebelen, T., & Kuperman, V. (2010). Using Amazon Mechanical Turk for linguistic research. *Psihologija, 43*, 441–464.

Shafir, E. B., Smith, E. E., & Osherson, D. N. (1990). Typicality and reasoning fallacies. *Memory & Cognition, 18*, 229–239.

Shapiro, K. L., & Raymond, J. E. (1997). The attentional blink. *Trends in Cognitive Science, 1*, 291–296.

Shepard, R. N., Hovland, C. I., & Jenkins, H. M. (1961). Learning and memorization of classifications. *Psychological Monographs, 75*(13), 1–42 (Whole No. 517).

Stewart, N., Chater, N., & Brown, G. D. A. (2005). Decision by sampling. *Cognitive Psychology, 53*, 1–26.

Stroop, J. R. (1935). Studies of interference in serial verbal reactions. *Journal of Experimental Psychology, 18*, 643–662.

Suri, S., & Watts, D. J. (2011). Cooperation and contagion in web-based, networked public goods experiments. *PLoS ONE, 6*(3), e16836. https://doi.org/10.1371/journal.pone.0016836.

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science, 185*, 141–162.

Tversky, A., & Kahneman, D. (1983). Extensional versus intuitive reasoning: The conjunction fallacy in probability judgment. *Psychological Review, 90*, 293–315.

# Chapter 14
# Spatial Humanities: An Integrated Approach to Spatiotemporal Research


**David Blundell, Ching-Chih Lin, and James X. Morris**

## 14.1 Introduction

In recent decades we have entered an age where digital tools are ever increasing in capacity to help us with daily life. In the academic realms of text mining, network analysis, public history, heritage studies, and mapping we are coming of age in digital humanities and related disciplines (see Blundell and Hsiang 1999). Among these areas, many specialties focus on analyzing digital space through time. We call this area spatiotemporal research—mapping across time with digital computational methods providing a large array of information. This enhances our ability to observe data beyond an individual's abilities to perceive all the possible components. The possible data stems from aerial mapping, remote sensing, photometric imagery, random sampling archaeology, statistical programming with languages such as R, and contemporary software development for innovative methods to see beyond what we can see. When conducting fieldwork, you may find there are occasions when digitizing data becomes necessary. Whether this is due to limitations such as time or access, mobility issues requiring light travel, or due to chance, such as the occasional lucky find, digitization is an excellent method to collect spatiotemporal data. This chapter outlines several varying projects and methodologies in the

D. Blundell (✉)
Asia-Pacific SpatioTemporal Institute (ApSTi), National Chengchi University, Taipei, Taiwan

Electronic Cultural Atlas Initiative (ECAI), University of California, Berkeley, USA
e-mail: pacific@berkeley.edu

C.-C. Lin
Graduate Institute of Religious Studies, National Chengchi University, Taipei, Taiwan

J. X. Morris
International Doctoral Program in Asia Pacific Studies, National Chengchi University, Taipei, Taiwan



S.-H. Chen (ed.), *Big Data in Computational Social Science and Humanities*,
Computational Social Sciences, https://doi.org/10.1007/978-3-319-95465-3_14

digital humanities incorporating integrated approaches to spatial humanities and spatiotemporal research. We invite you to participate in this field of spatiotemporal methods to enhance your research. With this chapter we hope to inform, instruct, and inspire more research in this new and exciting area.

History is elusive and often written from specific perspectives (see Buckland 2004, p. 39). Regardless of our so-called objective research, there is a bias in data selection. Digital humanities as a scientific method gives us a view across big data providing unexpected ways of looking at and configuring information. Mapping is one of the most commonly used techniques in reviewing our 'sense of being' in space (see Cosgrove 2004; Blundell 2011, 2012).

As the authors of this chapter have experience in visualizing layers of ethno-archaeology from the perspective of sociocultural networks based on patterns of religious networks and self-represented environments, this chapter begins with an anthropological background of the individual and acculturation in space at a given time.

We are describing applications and methodological strategies for new ways to approach data from the grassroots. Our illustrations utilize geographic information systems (GIS) comprehending spatial happenings through time. The application of GIS to issues in history, archaeology, and anthropology is exciting in both digital and spatial humanities (see Gregory and Geddes 2014). Now historiography has fresh and innovative tools (Robertson 2012), and is no longer limited to literary text mining. Spatial humanities provide us with current advances in GIS computing and information infrastructures offering researchers the possibility of reconsidering the entire strategy of analysis and dissemination of information. It features "deep mapping," acknowledging multiple meanings in a place that "enables humanities scholars to discover relationships of memory, artifact, and experience that exist in a particular place and across time" (see Bodenhamer et al. 2010). 'Deep' cartographic explorations feature instructive approaches to multivalent ways for developing spatiotemporal voices in spatial anthropology (Roberts 2016).

Our time maps research contributes to important academic discourse in many ways. Time maps are utilized to trace stories by the way people move through time. Other digital humanities researchers map data to understand spatial patterns from original sources. These visualized spatiotemporal displays contribute to discovering knowledge, answering questions, and seeking other questions. Spatial humanities produce a cycle of questions creating layers of maps portrayed in different ways.

The experiential projects introduced here further promote such innovative media, both analog and digital methodologies that are important components to explore the aesthetics and ontology of perspectives of history, social and physical sciences, and humanities. We explore heritage *vis-à-vis* present-day society and the determinants of people in this context.

Our challenge is to imagine new methods for doing research and making those results available to broader user communities. Can we find meaning and innovation in the digital humanities beyond what scholarly efforts have already accomplished? Digital humanities have matured, crossing boundaries of geography and history to include a wider range of disciplines through our mutual transdisciplinary synergies. We comprehensively search for spatiotemporal points where networks existed and continue to do so.

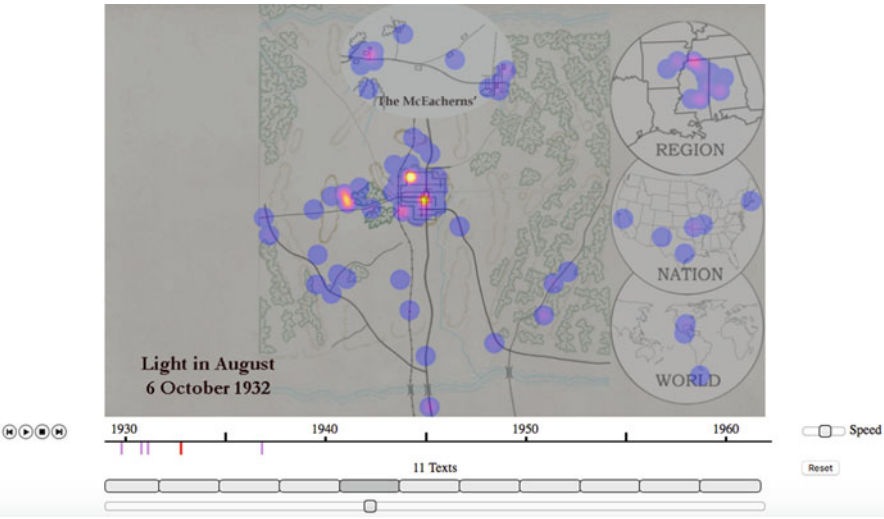



**Fig. 14.1** Spatial use of William Faulkner's map of Yoknapatwapha. Based on Faulkner's early hand-drawn map of Jefferson, Mississippi, a GIS heatmap from *Light in August* (1936) identifies story locations where events occur in the county, region, nation, and world (images retrieved from Digital Yoknapatwapha, University of Virginia, http://faulkner.iath.virginia.edu)

The models we share with you in this chapter are only several examples of the many ways GIS is used in this generation of digital humanities. Our purpose in sharing these projects is to visualize spatial concepts and systems that may not necessarily immediately spring to mind as being typical GIS models. The ability to utilize GIS for other areas of study has the benefit of allowing different realms of academia to begin thinking outside the box, and challenging our previous ways of understanding an area of study by providing new views of utilizing research methods.

An example of dynamic geographic visualizations from historical literature is mapping Yoknaphatwapha an imaginary town described in the award winning novels of William Faulkner based on his hometown in Lafayette County, as an interactive GIS time map produced by local and international scholarly technologists at the University of Virginia.[1] Faulkner drew a simple a map of his celebrated fictional county and devised an entire history for the space, from its origins as Chickasaw territory through the twentieth century, including locations, relationships of individuals, and other demographic information which intersected events experienced by denizens of Lafayette County (see Fig. 14.1). The history of these literary locations were examined in detail by Richard Reed and again its interwoven history with that of Lafayette County has been analyzed by Charles S.

[1]For more information, visit Digital Yoknapatawpha (http://faulkner.iath.virginia.edu).

Aiken (see Reed 1974; Aiken 1977, 1979). Ann Heylen featured in Chap. 4 of this volume utilizes this research model.

In many ways, the richness of events in Yoknapatwapha County draws parallels to that of Middle Earth of J. R. R. Tolkein. Defined geographic spaces and features and events of cultural significance for the characters have defined a 'real' space with as much significance to the reader of fiction as the streets of ancient Rome, hidden beneath a modern Italian metropolis, has for the archaeologist. Where GIS and remote sensing have been used to explore the buried and inaccessible sites of the Romans now buried beneath 2000 years of history, these spatiotemporal tools have been applied to the area of literature to help better interactively visualize and understand the social contexts of the novelist's geographic space (see Digital Yoknapatwapha).

Our spatiotemporal interfaces provide new methods of integrating primary source materials into crosswalks of interactive visualizations. Utilizing GIS we are able to chart the extent and dynamics of specific traits of cultural information to create layered maps. These elements are transmitted through time based on spatial points. The research outcome is a Web-based, interactive, and cultural platform acting as an atlas for a local community bulletin board designed for scholarly exchange.

The aim is to recount human transformations from cartography, historical records, aesthetic determinants, and community research partnerships. That implicit conceptual underpinning of advanced hermeneutics research in our qualitative tradition is critical and able to potentially enrich and deepen perspectives *based on elements seemingly unrelated, yet connected* (see Blundell 2016).

A far-reaching goal is to further standards in cartographic strategies through the utility of digitalization and animation of map content giving new possibilities in the hands of local and international collaborators. We provide examples for developing best practice standards applied to databases giving interactive multimedia utility aspects. This allows uniting the context of environmental landscapes with cultural data for making enhanced possibilities in spatial humanities with scales of data, large and small—with humanistic and scientific results.

For an entry into spatial humanities we consult Jo Guldi's introduction of the spatial turn for eight academic disciplines, "What is the Spatial Turn?" (see Guldi 2011) and Richard White's essay "What is Spatial History?" (White 2010). And how are we utilizing 'geolocative technologies' that are reshaping spatio-cultural narratives and experiences (see Stadler 2015)? Digital mapping today gives resource affordability and availability of resources to novice or advanced researchers who are not cartographers, abilities to chart information.

The interactivity of digital mapping allows one to filter data to the desired scale and includes a multitude of sources that, through their abundance, allow for data gathering and transdisciplinary comparisons to be made, such as analysis for arguing geographical claims and historical legal issues (Meccarelli and Sastre 2016). Sometimes, these techniques take confidence to master. We hope to provide case examples conducted by anthropologists and historians, not cartographers. We aim to open spatial humanities to anyone.

## 14.2 Our Experiential Projects

In 2015 we initiated our Asia-Pacific SpatioTemporal Institute (ApSTi), Top University Project in Digital Humanities, National Chengchi University, Taipei, Taiwan (see http://apsti.nccu.edu.tw). Here we have created an environment for synergies to occur between researchers serving to facilitate studies as a home for innovative geographic information systems (GIS)-based research and sharing advanced technologies in the digital humanities. Our institute offers a range of project services to facilitate new ways of configuring data based on geospatial tools. Through developing interfaces for spatiotemporal systems we are able to generate dynamic maps of unique information possibilities. Researchers from various disciplines contribute to dialogues about techniques, challenges, and results of digital humanities research. In brief, ApSTi facilitates capacity building and innovative ways of sharing information using digital methods for visualizing spatiotemporal aspects of the human experience (see Blundell and Jan 2016).

Here we will mention some of our projects currently underway. Let's open with a Web-based platform our affiliate the Electronic Cultural Atlas Initiative (ECAI) Austronesia Team guided by David Blundell assisting with projects of the ECAI *Atlas of Maritime Buddhism* conceived by Lewis Lancaster.

The project is developed to reach world audiences and features a high level of participatory interactive 3D visualizations to be more accessible with mobile phone applications (APPs) and multimedia museum displays (Figs. 14.2 and 14.3) (see Zerneke 2014). Featuring historic timelines, ships, trade routes and trade winds, traveling monks, life accounts at ports, and diaries integrate content and technology to enable our understanding of Monsoon Asia, its diffusion of culture, and oceanic navigation to become alive and accessible (Blundell and Zerneke 2014).

Michael Buckland states:

> It is useful to distinguish between the past—what happened, thus history, accounts of the past—and heritage, which are those parts of the past that affect us in the present. To be more precise as a student of history, it depends on the documentation of the past, as in writing. That is to say that the events that have transpired are no longer directly knowable. The past is knowable only indirectly through histories, such as descriptions and narratives of what happened. For any aspect of the past, there may be many narratives or none. Histories are always multiple and incomplete. Many factors influence what histories are, or can be written. As heritage is legacy from the past that we compose life with in the present and give to the future as reference for local identity, it is also a marker for universal human appreciation. (Buckland 2004, p. 39)

The challenge accepted by this project is to break new ground, developing new knowledge using digital tools to produce results that could not be achieved through traditional research in any single discipline. It is leveraging data from disparate databases to create integrated systems and customizable visualizations. Within spatiotemporal religious historical studies this opens new perspectives on the early Asia-Pacific transport systems of navigation in the region of the Indian Ocean. It will enable critical discussion on the extent Austronesian navigation


**Fig. 14.2** Google Earth visualization of the custom gazetteer developed to initiate research for the ECAI Atlas of Maritime Buddhism (courtesy of Jeanette Zerneke, Google Earth)

played in the transmission of religious beliefs. The *dharma* has not been necessarily associated with Austronesian cultures, yet the largest and Buddhist monument, Borobudur, Java, Indonesia, is in Austronesia. This will beg the question as to what is the early historical Austronesian contribution to world history. This contributes to our scholarly attention of indigenous cultures, trans-ocean navigation, migration, symbolism, international belief systems, and narratives of new dimensions through the innovative methodology of spatial humanities (Blundell 2014). It is a compelling demonstration of how GIS can contribute to our complex historical understanding.

Our knowledge derives from various research fields, and integrates many different types of data and analytical styles developing new research methodologies, creating paradigm shifts and multi-vocal views in the humanities. We are able to chart the extent and dynamics of specific elements of cultural data via maps using GIS.

By utilizing modern communication technologies, typically innovative mobile handheld tools, acquiring high-quality spatiotemporal information has become much more efficient and cost-effective than before. As a result, many researchers deem that data collected by volunteers with minimum training can be used for scientific researches if carefully designed quality assurance processes are performed. Our projects explore the concept of Monsoon Asia and globalization—positioning historical roots and contemporary languages and cultures as valuable peaceful and


**Fig. 14.3** Google Earth interactive spatiotemporal interface for visualization of Buddhist sites, trade routes, kingdoms, and empires from 500 to 1500 CE, with dynamic map overlays. Notice time bar scrolled to the year 1496 in upper left corner of map (courtesy of Jeanette Zerneke, Google Earth)

sustainable development tools for interconnectivity across Asia Pacific that can be used to seek collaboration and partnership due to their association with heritage. We try to establish a clear relationship between all aspects of our projects, particularly based on languages and cultural elements as logistical resources.

Today we have vast interconnectivity via handheld electronic mobile devices and APPs for people-to-people sharing from grassroots communities to world systems. Among our GIS tools we employ volunteered geographic information (VGI) methodologies to collect field data about various and selected aspects of cultural and natural resources within local communities. The local community residents use mobile devices such as smartphones, global positioning system (GPS) logger, and digital cameras for recording locational coordinates and images of important features. We develop software applications to be installed on smartphones and tablets, and handy tools for uploading and mapping data contributed by participants. This in turn becomes our Web-based knowledge bulletin board.

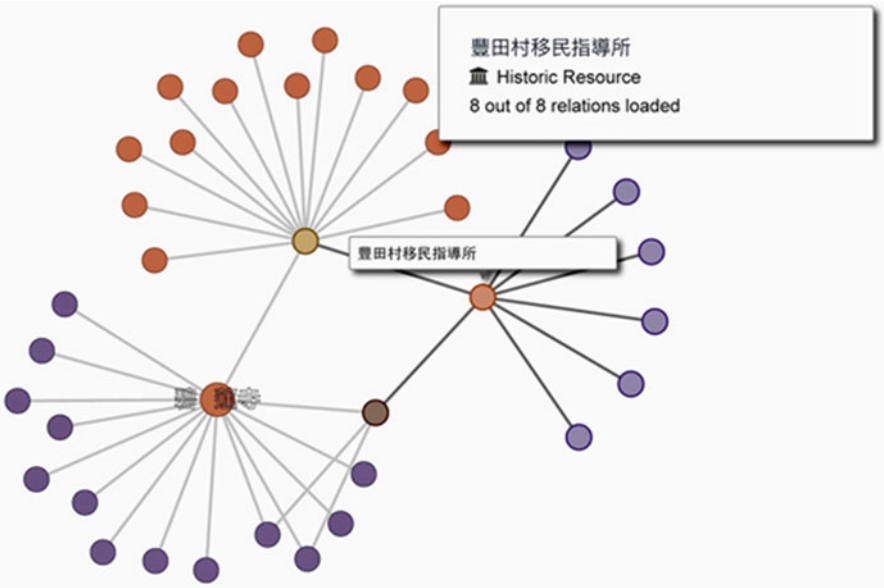



**Fig. 14.4** A screenshot of one instance of a dynamic interactive graph showing the relations network among cultural heritage resources. By selecting a node on the graph, additional interrelated data points will be populated, enabling the viewer to browse the connections within a network (produced by Jihn-fa Jan using Arches)

Jihn-fa Jan[2] in the Department of Land Economics at National Chengchi University, Taipei, is developing user-friendly applications for local community-volunteered geographic information, an innovative methodology in spatial humanities. His research demonstrates the procedures for obtaining high-quality natural and cultural information of various resource kinds from data collected by volunteers equipped with smart devices or global positioning system (GPS) taggers. Selected basic elements are specific to a local community, yet through mapping techniques of digital and spatial humanities he is able to chart network systems that could not be done previously.

His contribution utilizes the Arches (http://archesproject.org) system developed by the Getty Conservation Institute (http://www.getty.edu/conservation), an innovative open-source geospatial software system for heritage inventory and management for developing a spatial database through Web applications for community resources (Fig. 14.4). In order to satisfy the requirement of visualization, his research also uses the LizardQ system to take panoramic photographs of local community resources. After the database and Web application is established, they could be references for

[2]See Jihn-fa Jan, Chap. 2 in this volume.

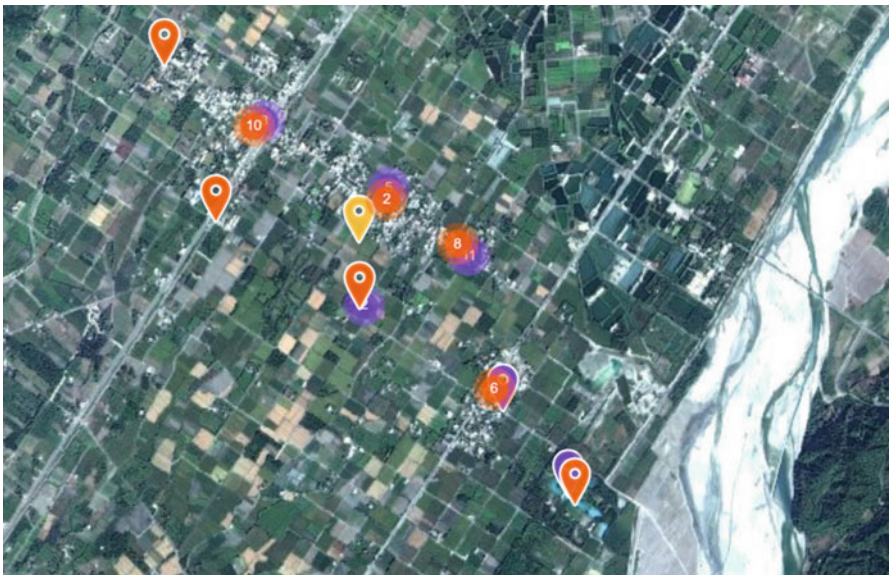



**Fig. 14.5** Locations of cultural heritage resources overlaid on aerial orthoimages identified by local people as significant sites of importance within their community (image by Jihn-fa Jan using Arches)

planning and managing community resources for long-term benefits and sustainable development (Jan and Mao 2016).

These mapping network systems designate data points to specific items which in turn can be connected with lines. People in a community contributing to the map will be able to see their basic elements have interconnectivity with other places distant and close to them. Meanings and shapes of the elements could then be observed, discussed, shared, and graphically displayed in a dynamic interactive map (Figs. 14.4 and 14.5).

Oliver Streiter[3] at National University of Kaohsiung, Taiwan, documents and researches burial grounds of Taiwanese and Chinese diaspora in connection with the East Asian region. Cemeteries are modeled as internally complex structures that are themselves composed of other more complex structures, such as tombs, grave markers, and temples. Streiter's work includes new technologies for digitalizing data including 360-degree photography of tomb sites, placing the viewer in the position of a visiting worshipper, and aerial photography of cemeteries through the use of remote control drones to gain a greater perspective of the orientation of tombs in addition to identifying and bringing inaccessible sites into range (Fig. 14.6).

Within the archive all structures are computationally represented through the use of tables (data frames), with columns that contain descriptions of the properties of

[3] See Oliver Streiter, Chap. 3 in this volume.

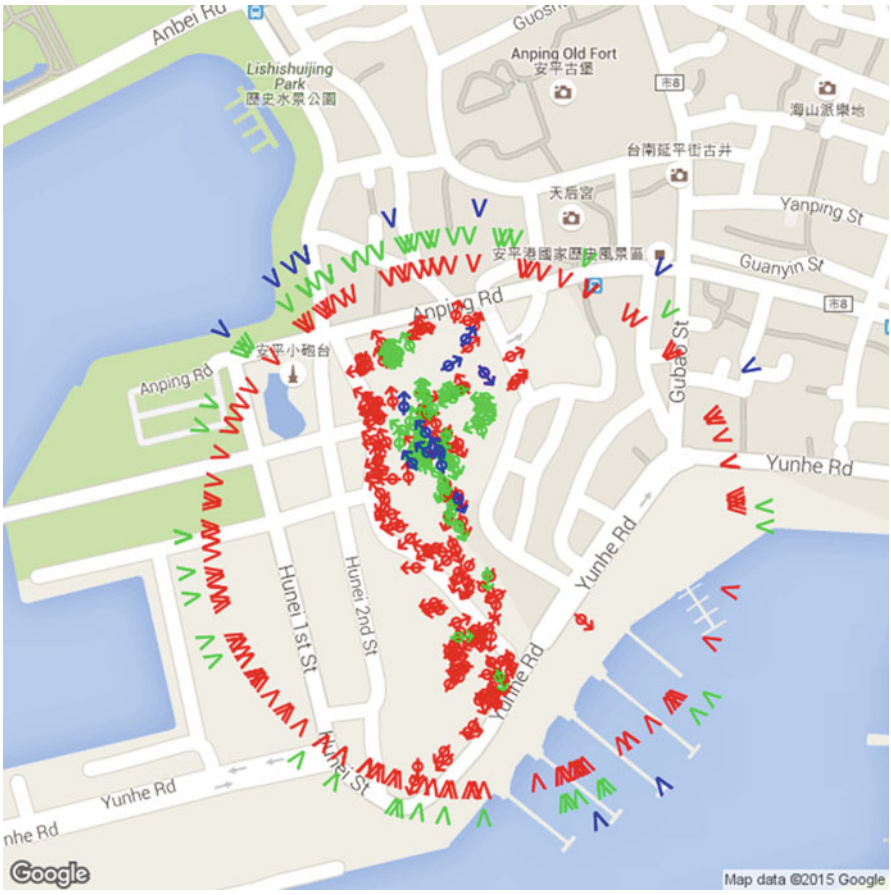



**Fig. 14.6** GIS map depicting orientation of Anping tombstones in Tainan, Taiwan. Legend in colors: blue, Qing dynasty; green, Japanese colonial (1895–1945); red, Republic era (1945–present) (graphic display compiled by Oliver Streiter, Google Maps)

each instance, and indices related to instances that can be found in other tables, e.g., common identifiers that can link a tomb and a tombstone (Streiter et al. 2011; Streiter and Goudin 2013; Streiter 2015).

The primary source of information used for filling the tables with content are georeferenced digital photos, typically taken in a way so that one photo represents one component of a tomb, i.e., one instance in a table. Every photo is also an instance in media-file tables that links photos and modeled structures through identities, e.g., a tomb identifier.

The data tables, a computational model of Taiwan burial grounds, and all media files are made available to researchers in order to promote digital approaches to the study of Taiwan in terms of history and culture. These resources are archived

**Fig. 14.7** Temple exteriors, altar icons and imagery, and donor placards and stones are among the digitized shrine elements in the corpus (courtesy of James X. Morris)

through the Data Archiving and Network Services (DANS)[4] and can be accessed as a set of 230,000 media files and, linked to it, as a set of data files describing 813 burial grounds and 62,361 tombs.[5] Within the data set, separated CSV files each store one table of tombs, tombstones, etc. Also made available is a bundle of helper functions written in R, tailored to easily access the data set and to prepare and elaborate the data for spatial and temporal analyses (Streiter and Morris 2015).

James X. Morris is documenting and mapping land god shrines to establish patterns of settlement, communal organization, and the historical trade networks of communities in Taiwan, southern maritime China, Hong Kong, and Macau. His research is ongoing with a primary focus on northern Taiwan and the greater Taipei region. The unique nature of land god shrines, their quick establishment within a newly settled location, relative permanence, and utilization as markers of neighborhood and community activity create a foundational social construct in traditional Chinese culture from which spatiotemporal GIS projects can be built upon (Fig. 14.7). Through documentation and mapping of these land god shrines his research is able to establish patterns of settlement and expansion, economic development, communal organization, historical trade networks, and intercommunal linkages.

[4]The Data Archiving and Networked Services (DANS) is an institute of the Royal Netherlands Academy of Arts and Sciences (KNAW) and the Netherlands Organization for Scientific Research (NWO) dedicated to the archiving of digital data for the humanities and social sciences.

[5]To access the set of media files (see https://doi.org/10.17026/dans-zmh-2jjs). To access the set of data files (please visit https://doi.org/10.17026/dans-zvy-rtju).

The database is primarily digital photos with geo-referenced metadata. Elements of each land god shrine are systemically photographed and stored within shrine-specific directories. Utilizing methods devised by Streiter et al., Morris' digital corpus now includes more than 500 land god shrines consisting of over 20,000 digitized elements found throughout maritime East Asia.

Through the use of historical maps, records, and data collected through extensive fieldwork Morris is reconstructing the original settlements and trade networks of the Taipei basin that have since been lost due to war, colonial suppression, and rapid development. Within historical Taipei his research suggests that the oldest networks of land god shrines were situated along waterways that have since been covered, filled in, redirected, or otherwise lost during the twentieth century. He is also exploring the origin and distribution of land god steles, enshrined carved stones bearing the land god's name, found in clusters throughout the Tamshui River watershed of northern Taiwan, focusing on character variations cross-referenced with tombstone inscription variations. The spread of these two technologies, stone tablets and variant character types, combined with records of known periods of settlement and historical maps, will provide clues of how far stone carving trade networks in northern Taiwan extended, thus giving researchers a new tool in their exploration of early internal commerce infrastructure. This type of research provides another example of how digital humanities based on spatiotemporal GIS platforms is changing historical investigations (Fig. 14.8).

The next step for this database is to build a Web-based platform for uploading and exploring data and metadata by users. By establishing a collaborative database-and-tool platform with GIS functionality the collection and dissemination of digital land god shrine data will enable researchers to not only explore communal shrines across an historical context, but they will also be able to sort, tag, collaborate, and share one another's data to create a more enriched experience. As with Streiter's tombs research, integrating new multimedia such as 360-degree photography to allow the users to explore the interior of shrines, video of festivals, ceremonies, and processions, and aerial drone photography to orient the shrine's place within a community will serve to deepen the understanding of land god shrines and social constructs within spatiotemporal humanities.

Ching-Chih Lin, teaching at the Institute of Religious Studies, National Chengchi University, is working with the Center for GIS at Academia Sinica in Taipei, Taiwan, to develop methods and resources in spatial humanities that can contribute to the study of Chinese religious systems. He provides spatiotemporal platforms for analysis and visualization, which will be explored in greater detail later in this chapter. Lin's work on methods and resources of spatial humanities find that there exists an abundance of online GIS (WebGIS) platforms that can contribute to the study of Chinese religions by providing spatiotemporal platforms for analysis and visualization (Fig. 14.9).

With their graphic religious mapping contributions, scholars could utilize a number of WebGIS platforms associated with religious texts, sites, and figures and use GPS devices and GIS applications to create their own digital spatiotemporal

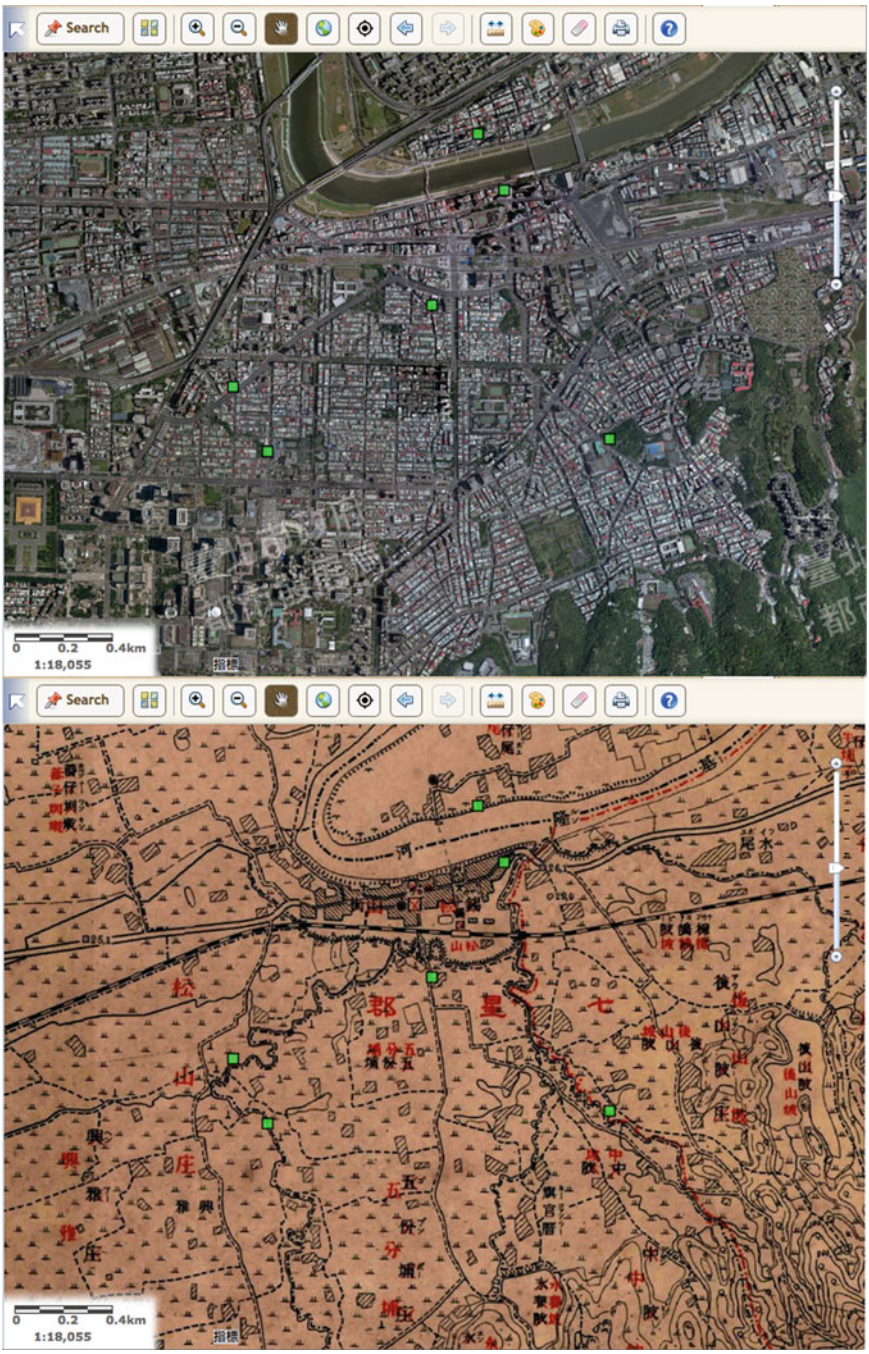



**Fig. 14.8** Utilizing GIS spatiotemporal tools available online, points can be added to a modern map of Taipei (above) to identify the sites of land god shrines near present-day Songshan Station and Raohe Night Market in Taipei. Below, utilizing historical map overlays (early twentieth century) the original terrain features and riverside placement of shrines containing steles can be seen clearly (map imagery provided by Department of Urban Development, Taipei, Taiwan)

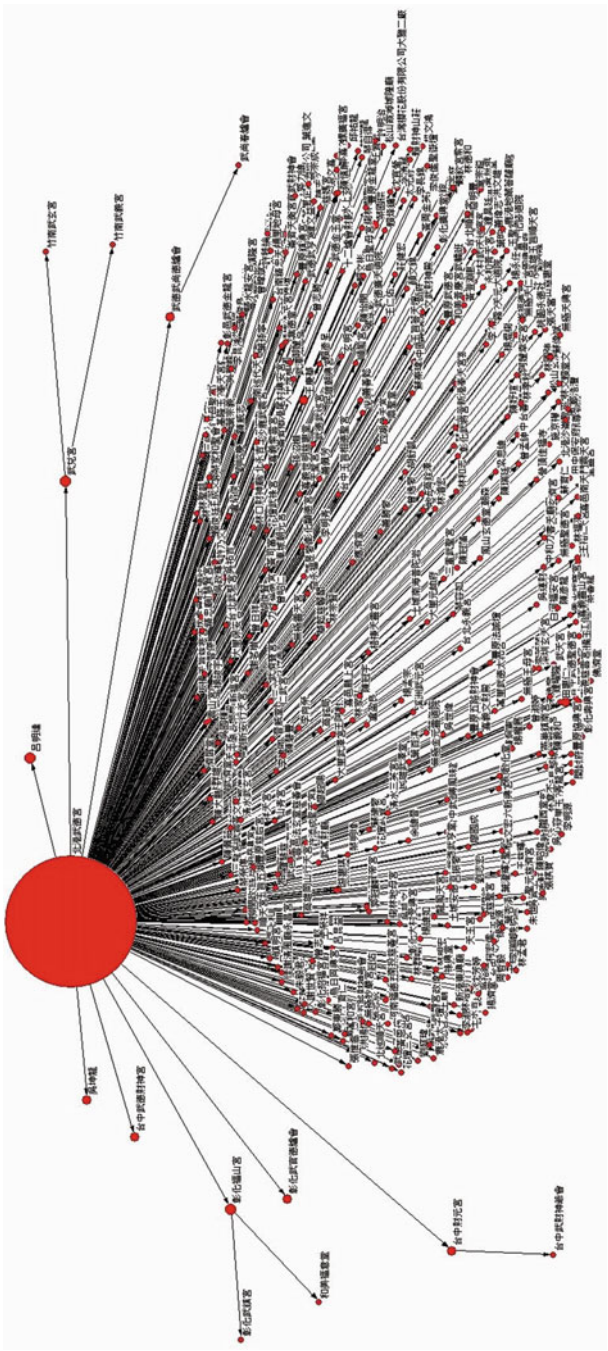


**Fig. 14.9** GIS graphic display of the Wude (Deity of Wealth) temple expansion across Taiwan with 3000 spirit separations of incense from a source temple in three generations spanning four decades to the present (compiled by Lin 2016)

data of religious activities and spaces (Lin 2016).[6] Lin's work includes instructing readers and scholars how to utilize GPS devices to record locations of temples and routes of religious processions and pilgrimages and how to create spatial and temporal datasets of religious sites and activities on GIS infrastructure.

Ann Heylen,[7] directing the Center of Taiwan Studies, National Taiwan Normal University, Taipei, is applying digital software to produce a dynamic map of seventeenth century Dutch handwritten manuscripts documenting the colonial Dutch community in Taiwan that are indispensable for a holistic understanding of Taiwan history in a global setting. Her research methodology brings to attention issues such as transcription and transliteration in view of orthography and spelling that was not standardized. This particularly pertains to place and personal names that are used as main entries. This research digitalizes a version of the church minutes of the manuscript *Kercboek, Brievenboek van Formosa, 23rd januari 1642–1644 maart* in Dutch and Chinese translations. Her observations shared through digital humanities to show how historical documents pertaining to Dutch Formosa research enables a new line of inquiry that approaches the cultural encounter between Dutch and indigenous society by paying attention to the various ways in which the encounter was expressed and represented, and also by which our current understanding is shaped (Heylen 2016).

Hsiung-ming Liao serves at the Center for Geographic Information Science (Center for GIS) of the Research Center for Humanities and Social Sciences (RCHSS) at Academia Sinica, Taipei. He is researching area studies by collecting data from fieldwork and using cross-related datasets from heterogeneous sources, of diverse types, and for multiple purposes. To conduct collaborative research works, in particular in multidisciplinary and wide-area projects, it is a norm for the various stakeholders to share and reuse their datasets. As the volume of datasets increase, so are the demands on the efficient and effective use of these datasets, ranging from data curation, storage, search methods, and visualization.

Liao specializes in collaborative research demonstrating platforms for data curation and repositories using open source and free software. His work features CKAN platforms as open source software and a powerful data management system. This software and extended spatial-related functions build the data portal for the specific research group. His goal is to assist researchers who engaged in the GIS-based data repository to share and find research resources effectively as an index so the user will not duplicate data entries. In addition, there is emphasis on the use of Web standards and open source software in this project, so that the platform can be freely disseminated and reused.

Janet Tan coordinates advanced digital humanities projects including ApSTi at the Research Innovation-Incubation Center, National Chengchi University in Taipei, Taiwan. Her projects identify regional economic data for creating dynamic cluster maps presenting graphical images. The model is based on case studies offered at

[6] For more details, visit http://crgis.rchss.sinica.edu.tw/spatial/webgis.

[7] See Ann Heylen, Chap. 4 in this volume.

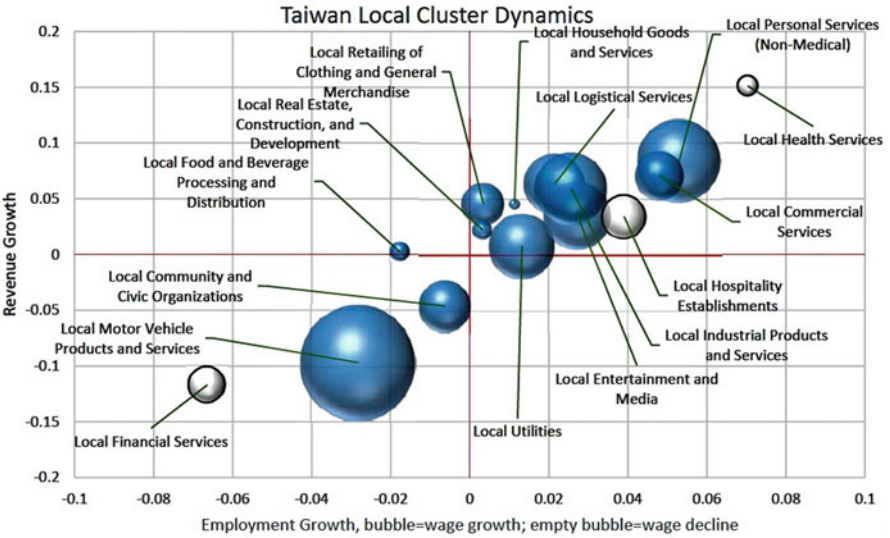



**Fig. 14.10** Taiwan local employment cluster mapping 2006–2011 (graphic by Janet Tan and Jihn-Fa Jan)

the Harvard Business School by Michael Porter on strategy and competitiveness research methodologies. The economic GIS data mapping provides viewers with a quick view of economic interpretation on relevant regional data for comparison purpose (Fig. 14.10).

The cluster mapping methodology addresses economic and sustainable developmental questions: "Why are some nations or regions more prosperous than others? What conditions enable global corporations or local businesses to innovate and grow?" (Tan and Jan 2016). Economic dynamics have been driven at a high pace, and understanding the trajectory of the economics is important. Technology advancement and globalization have triggered a paradigm shift in the recent economic growth from a supply-driven to demand-driven market. The change greatly impacted the policy makers' decision flow to make effective policies in a timely base. Three economic policy decision models: top-down, bottom-up, and interactive, are researched and compared in this chapter to reflect the differences.

This research utilizes a cluster mapping framework and growth dynamics to derive the economic landscape across Taiwan and agglomeration models. Through the analyses of cluster dynamism, challenges and potentials are pointed out, and a graphical tool is designed to unite policy makers, practitioners, and researchers for effective economic development initiatives. Porter's viewpoint is … "traditional economic theories fail to capture many of the underlying forces at work in today's global economy" (Tan 2016).

Richard Cornelisse, awarded by the Ministry of Culture, Taiwan, works on 360-degree virtual reality documentaries systems developed for exploring questions of

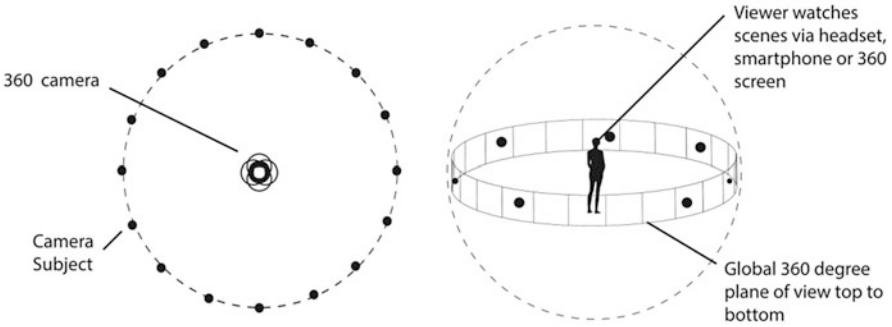



**Fig. 14.11** People stand around the 360-degree camera in circular compositions thus giving the viewer a privileged perspective within the action of each scene (courtesy of Richard Cornelisse)

identity in landscapes of a diverse multiethnic groups, the connection to perception, traditions and globalization, as well as how these questions in turn characterize the culture of Taiwan as distinct within the region (Fig. 14.11).

The virtual reality framework presented here is meant to provide a multi-aspect experience of a diverse Taiwanese milieu, which captivates and transports the viewer to a subtle, though heightened, awareness of nuances between people, place and traditions. This experience re-contextualizes how one can understand the people, culture, diversity, and landscape of Taiwan as both viewer and participant. Within this virtual reality platform, the documentary will focus on the voices, songs, dances, sounds, stories, languages, rites, and landscapes in various locations in Taiwan through 360-degree circular compositions.

Results will place the viewer in the center of an immersive nonlinear documentary that activates viewer participation and allows them to engage the subject matter in a unique individualized experience of people and place: one that not only questions how one can understand issues of identity and traditions in a rapidly changing global cultural landscape, but also how one can experience culture. This unique mosaic, in conjunction with the external pressures and influences that surround Taiwan in the area, make broad categorizations difficult to qualify definitely from any perspective. The aim is thus to make an immersive video that broadly documents distinct and dynamic cultures through an essay of fluid interactions and interrelation, which depicts cultures in flux, preserves depictions of traditions in virtual reality space, and underlines a sense of viewer experience and appreciation of Taiwan cultures and environments beyond a monolithic framework.

The subsequent immersive interaction within this realm provides the viewer with an active role in creating meaning and engendering sense of connectivity, hence empathy with the subject. Ultimately, this interaction should empower viewers to apply such meaning and presence to their own lives and opens up questions of identity, as individuals, community members, citizens of a country, and as persons in a rapidly changing globalized world (Cornelisse and Blundell 2016).

Here the goal is to recount narratives from living and historical records of religious transmissions, aesthetics, and partnerships implicit as conceptual underpinnings of advanced hermeneutics research in our qualitative tradition, critical, and able to enrich and deepen perspectives based on cultural elements seemingly unrelated, yet connected. The core interests share a commonality based on a platform of spatial humanities utilization, a subset of digital humanities. Spatiotemporal interfaces provide new methods of integrating primary source materials producing interactive visualizations.

Utilizing GIS we are able to chart the extent and dynamics of specific traits of cultural information creating layered maps. These elements are transmitted and based in places through languages and belief systems across Monsoon Asia. Through digital and spatial humanities we are reexamining the furthest extent of early historical trans-regional cultural transmissions and subsequent influences on humanity across time to the present day. To facilitate collaborative research and public participation, our vision for ApSTi is to further utilize advanced geospatial and information technologies to create a platform that furnishes researchers with digital humanities tools for data acquisition, data sharing, spatiotemporal analysis, and information visualization.

## 14.3 Spatial Humanities, GIS Databases, and Digital Tools

In terms of GIS historical maps of cultural heritage and religious sites, Academia Sinica and Dharma Drum Institute of Liberal Arts in Taiwan have established several outstanding digital databases that incorporate historical information of religious sites or activities with gazetteer-style interactive maps. The Center for GIS at Academia Sinica launched the Cultural Resources GIS to document the historical and geographic information of heritage sites, including most local temples, ancient buildings, cultural heritage sites, old trees, and significant religious sculptures (Fig. 14.12).

Local historians have helped investigate and photograph detailed information of sites, including deities, hagiographies, inscriptions on steles, histories, organizations, and buildings. As the database matures it will eventually include complete information of all heritage sites, with the aim of providing scholars and local historians a useful infrastructure to explore local history, religious beliefs and activities, and social networks in terms of spatial and temporal significance.

An excellent example of the power of GIS and spatial humanities working together to build a platform for spatiotemporal research such as the GIS platform of Buddhist Temples produced by Dharma Drum Institute of Liberal Arts. This platform has been linked to other database platforms associated with Buddhist canons, temple gazetteers, and biographies. All geographic locations in the texts can be shown on Google Earth with further information. The GIS platform for Buddhist temples in Taiwan also provides scholars a very useful tool to look up the spatial and temporal information of individual Buddhist temples on the Google Map with

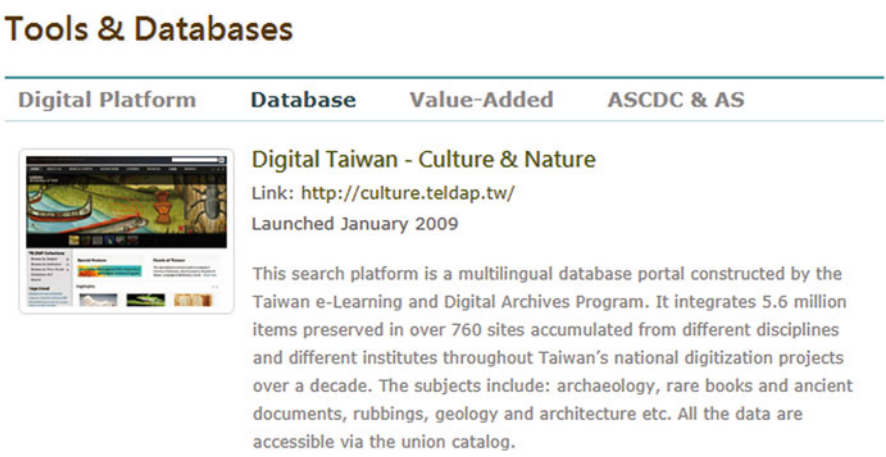



**Fig. 14.12** Academia Sinica provides researchers in the digital humanities with various platforms containing a growing catalogue of entries (screenshots from Cultural Resources Geographic Information Science, Academia Sinica, Taipei, Taiwan)

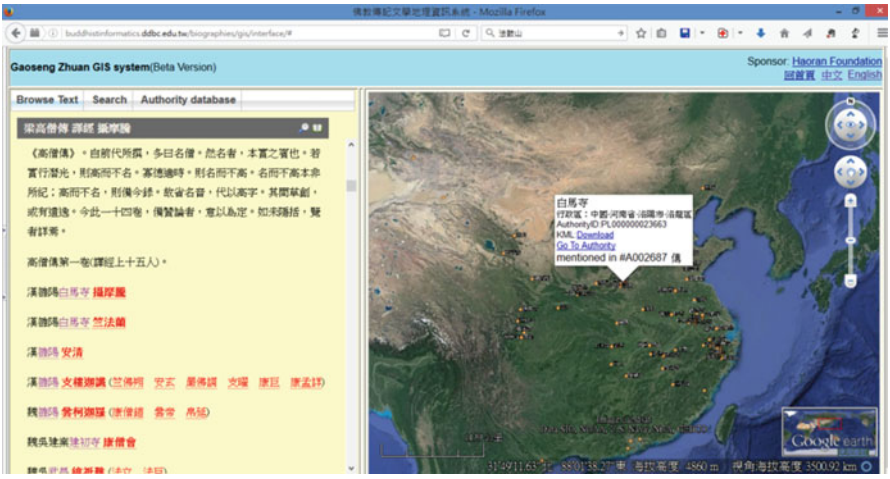



**Fig. 14.13** Screenshot of GIS Biographies of Renowned Buddhist Monks, Dharma Drum Institute of Liberal Arts

timeline (Figs. 14.13 and 14.14). To take Longshan Temple (龍山寺) as an example, the platform can visualize the locations of the 15 Longshan Temples in Taiwan from 1802 through 2009 (Fig. 14.15).

Academia Sinica has also developed a very useful platform integrating historical maps and aerial photographs. Historians can conveniently access historical spatial information for the reference of humanities research. In the past few years, several pioneer studies of local religions were done thanks to the aforementioned databases and infrastructure. Now several APPs for Android smartphones can also provide

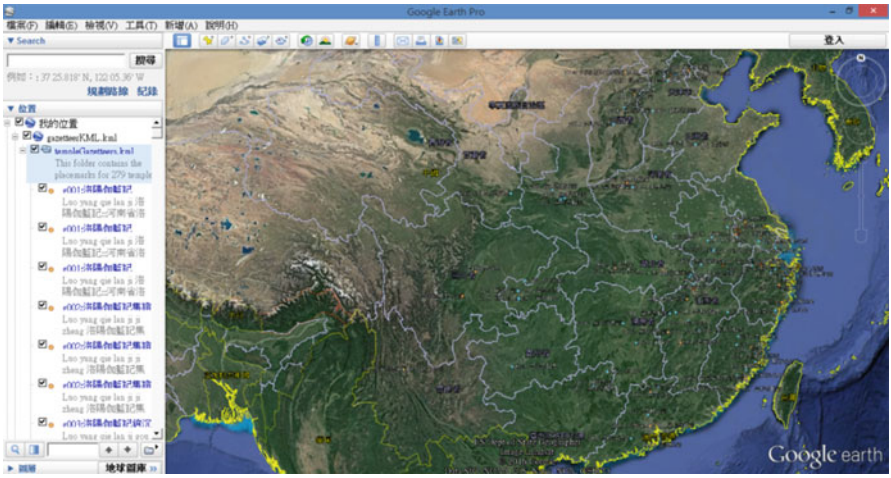



**Fig. 14.14** Screenshot of Buddhist Temples in East Asia located with Digital Gazetteers, Dharma Drum Institute of Liberal Arts

users with access to historical maps with handheld devices with great convenience (Fig. 14.16).

GPS has been applied in the study of Chinese popular religion in Taiwan. Temple processions are good examples. With GPS devices and GIS maps, scholars of Chinese popular religion can easily and accurately record and visualize the routes of temple processions and further compare routes of different years or temples. The Center for GIS at Academia Sincia and the Academia Sinica Digital Platform for Religious Culture in the Center for Digital Cultures both have collected and curated temple procession routes. Thanks to the visualization of a local temple processions, scholars at Academia Sinica have discovered that the land god temples along the routes within one township in Fujian, China, all face the township center. This phenomenon was never recognized with traditional approaches of archival studies or fieldwork.

Moreover, Academia Sinica and Chunghwa Telecom in Taiwan have also established simultaneous WebGIS to demonstrate the immediate locations of the main deities in temple processions and pilgrimages in Taiwan. GPS devices are installed on the deity's sedan chair carried in processions. Smartphone APPs are also created to make it convenient for procession followers to trace the location of a deity's sedan and then quickly join the procession. This also gives temples along the route ample time to prepare for welcoming the incoming deity's sedan and followers (Fig. 14.17).[8]

[8] All routes of temple processions recorded by Academia Sinica can be found at: http://crgis.rchss.sinica.edu.tw/spatial/atlas/roaging or http://deitygis.asdc.sinica.edu.tw/. Some technical details can be found here: http://www.godroad.tw


**Fig. 14.15** Above, screenshot of Dharma Drum Mountain GIS locator of Buddhist temples in Taiwan. Below, screenshot locating Longshan temples across Taiwan from Dharma Drum Mountain GIS Map of Buddhist Temples in Taiwan (http://buddhistinformatics.ddbc.edu.tw/taiwanbudgis/)

Researchers are utilizing GIS analysis tools and methods to visualize the spatial and temporal development of religious beliefs of several deities, and compared the relationship between specific deities and ethnic migrations.[9] Researchers also use GIS maps to analyze the process of expansions of certain deities, such as Mazu (媽祖) (Empress of Heaven) and the God of Wealth (財神). Mazu has gradually transformed her role from a fishermen's protector into multifunctional patrons as her cult expanded from coastal areas to inlandcities and mountain regions.[10] The God

[9]For example, see Hung et al. (2013a).

[10]Hung et al. (2013b).


**Fig. 14.16** APPs for Android smartphones such as the example above give mobile GIS historical map access for researchers conducting fieldwork or scholars without a desktop computer

of Wealth originated in rural and agricultural region in central Taiwan and quickly expanded with the migrants into urban and commercial areas, which demonstrates the correlation among economic development, rural-urban migration, and spread of God of Wealth temples (Fig. 14.18).[11]

[11] Ching-Chih Lin is currently working on this topic.

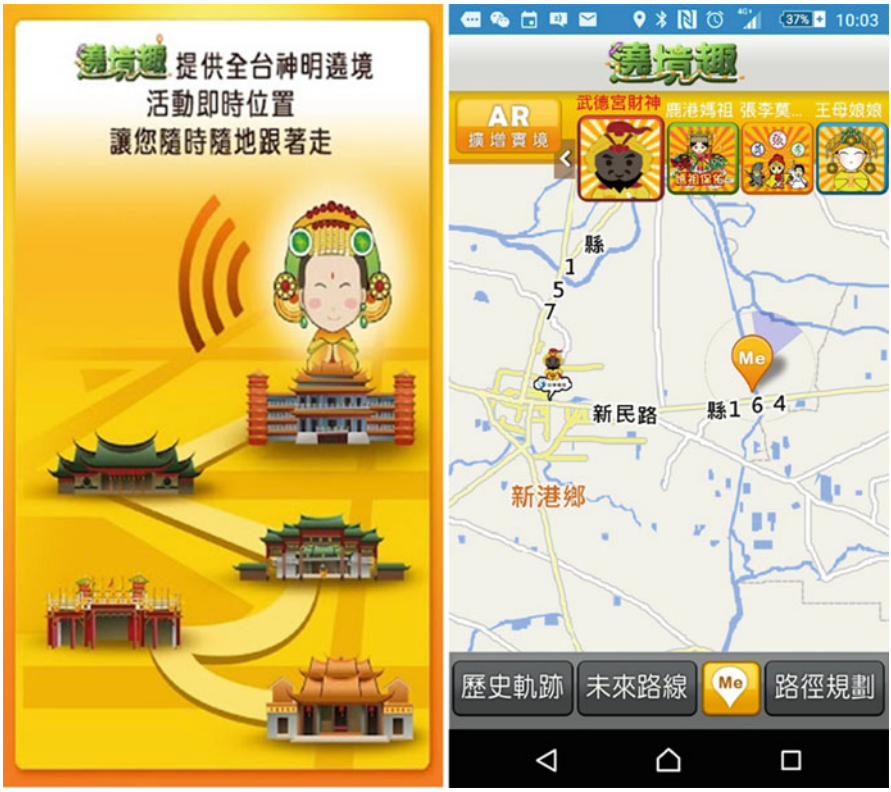



**Fig. 14.17** An example of an APP available from Chunghwa Telecom for Taiwan temple processions

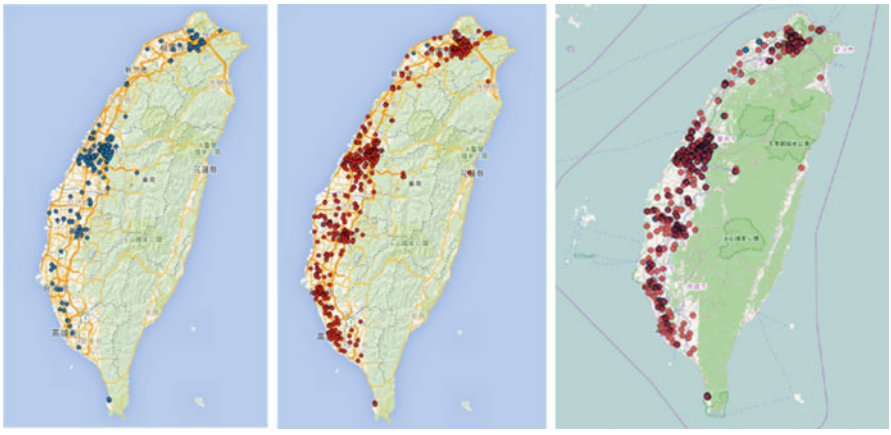


**Fig. 14.18** Left, from 1802 early spread of the God of Wealth (財神). Center, recent spread of the God of Wealth temples (2009). Right, entire historical spread of the God of Wealth temples in Taiwan (1802–2009) (maps by Ching-Chih Lin)

## 14.4 Getting into Spatial Humanities

To the reader, these projects could seem to be diverse and unrelated. Yet, with procedures and elements of the projects based on GIS mapping and formulating networks through spatiotemporal techniques, we are able to create a comprehensive atlas of maritime routes, burial sites, religious networks, volunteered geographic information (VGI) for localized points of heritage, and a visualized interactive virtual reality surround environment for a local narrative of having a 'sense of place'. These are collaborative research projects demonstrating platforms for data curation and repositories using open source and free software.

*We invite the reader to engage in utilizing spatial humanities for research.*

What is required?

See, Resources for Spatial Humanities: http://lincolnmullen.com/projects/spatial-workshop/resources.html

**Spatial Humanities Research Form**
*Fill-in data sets with profile:*

Give profile title:
Contact person:
Project abstract
Location
Short description
Latitude/longitude of location
Timeline range
Era/empire short description
Year range
Camera site? If yes:
Location short description
What is the content: What you can see from the specific point
Advantages of this site for connection to the narrative
Already collected data? If yes:
Data availability?
Proposed data—still to be collected? Brief collection plan
What is behind the site:
Writing on maps and mapmaking
http://lincolnmullen.com/projects/spatial-workshop/resources.html

**Acknowledgement** Appreciation is given to Lewis Lancaster, Jeanette Zerneke, and Michael Buckland for continuous support through ECAI. We are grateful to the kind patience and guidance of our ApSTi researchers and Shu-heng Chen, Vice President, and Director of Projects in Digital Humanities at National Chengchi University, Taipei.

## References


Aiken, C. S. (1977). Faulkner's Yoknapatawpha County: Geographical fact into fiction. *Geographical Review, 67*(1), 1–21.

Aiken, C. S. (1979). Faulkner's Yoknapatawpha County: A place in the American south. *Geographical Review, 69*(3), 331–348.

Blundell, D. (2011). World heritage: Cultural resource management giving Asia-Pacific a sense of place. *International Journal of Asia-Pacific Studies (IJAS), 7*(1), iv–ix http://ijaps.usm.my/?page_id=102.

Blundell, D. (2012). Taiwan coming of age. In D. Blundell (Ed.), *Taiwan since martial law: Society, culture, politics, economics* (pp. 2–26). Berkeley, CA: University of California, and Taipei: National Taiwan University Press.

Blundell, D. (2014). *Dharma civilization and stitched outrigger navigation: Contributions to ECAI project on maritime Buddhism. Buddhist culture and technology: New strategies for study*, *Vietnam Buddhist University series* (Vol. 26, pp. 41–63). Berkeley, CA: ECAI http://ecai.org/projects/MaritimeBuddhism.html.

Blundell, D. (2016). Indo-Pacific Austronesia: Spatiotemporal mapping points of early historical religious networks. In *International Symposium on Austronesian Diaspora*. National Research Center for Archaeology. Nusa Dua, Bali, Indonesia, July 18–23.

Blundell, D., & Hsiang, J. (1999). Taiwan electronic cultural atlas of the pacific. In *Proceedings of the 1999 EBTI, ECAI, SEER & Pacific Neighborhood Consortium (PNC) Joint Meeting* (pp. 525–540). Taipei, Taiwan: Computing Centre, Academic Sinica http://pnclink.org/annual/annual1999/1999pdf/blundell.pdf.

Blundell, D., & Jan, J.-F. (2016). Workings of spatiotemporal research: An international institute in Taiwan. In *IEEE proceedings of the 22nd international conference on virtual systems and multimedia (VSMM)* (pp. 33–38). Kuala Lumpur, Malaysia: Sunway University.

Blundell, D., & Zerneke, J. (2014). Early Austronesian historical voyaging in Monsoon Asia: Heritage and knowledge for museum displays utilizing texts, archaeology, digital interactive components, and GIS approaches. *International Journal of Humanities and Arts Computing, 8*, 237–252.

Bodenhamer, D. J., Corrigan, J., & Harris, T. M. (2010). *The spatial humanities: GIS and the future of humanities scholarship*. Bloomington & Indianapolis, IN: Indiana University Press.

Buckland, M. (2004). Histories, heritages, and the past: The case of Emanuel Goldberg. In W. B. Rayward & M. E. Bowden (Eds.), *The history and heritage of scientific and technical information systems* (pp. 39–45). Medford, NJ: Information Today.

Cornelisse, R., & Blundell, D. (2016). A Taiwan virtual reality memory project: Rituals in the circle. In *IEEE proceedings of the 22nd international conference on virtual systems and multimedia (VSMM)* (pp. 73–77). Kuala Lumpur, Malaysia: Sunway University.

Cosgrove, D. (2004). *Landscape and landschaft, lecture delivered at the "Spatial Turn in History?"* Symposium German Historical Institute. Retrieved from http://www.ghi-dc.org/fileadmin/user_upload/GHI_Washington/Publications/Bulletin35/35.57.pdf

Gregory, I. N., & Geddes, A. (Eds.). (2014). *Toward spatial humanities: Historical GIS and spatial history*. Bloomington, IN: Indiana University Press.

Guldi, J. (2011). What is the spatial turn? In *Spatial humanities: A project of the institute for enabling geospatial scholarship.* Retrieved December 30, 2016, from http://spatial.scholarslab.org/spatial-turn/the-spatial-turn-in-history/index.html

Heylen, A. (2016). Expressing power in dynamic maps through 17th century Taiwan Dutch manuscripts. In *European Association of Taiwan Studies (EATS) Conference*. Prague, Czech Republic, March 30–April 1.

Hung, Y. F., Chang, C. C., Liao, H. M., & Fan, I. C. (2013a). A preliminary spatial analysis of Zhenan Temple of Mount Maming. *Folklore and Culture*, (8), 41–64.

Hung, Y. F., Chang, Y. S., Chang, C. C., Chang, H., Fan, I. C., & Liao, H. M. (2013b). History and space: An exploration of the amounts and spatial distribution of Mazu temples in Taiwan. *Folklore and Culture*, (8), 17–39.

Jan, J. F., & Mao, W. H. (2016). *Cultural heritage inventory using Arches: Case study of a rural community in Taiwan*. Paper presented at the Pacific Neighborhood Consortium (PNC) Annual Conference and Joint Meetings, Does Data Construct Reality? The Getty Center, Los Angeles, California, August 16–18.

Lin, C. C. (2016). *GIS spatio-temporal analysis of Chinese woodblock prints (Nianhua) and religious networks in Taiwan*. Paper presented at the Pacific Neighborhood Consortium (PNC) Annual Conference and Joint Meetings, Does Data Construct Reality? The Getty Center, Los Angeles, California, August 16–18.

Meccarelli, M., & Sastre, M. J. S. (2016). *Spatial and temporal dimensions to legal history: Research experiences and itineraries*, *Global perspectives on legal history* (Vol. 6). Frankfurt am Main: Max Planck Institute for European Legal History.

Reed, R. (1974). The role of chronology in Faulkner's Yoknapatawpha fiction. *The Southern Literary Journal, 7*(1), 24–48.

Roberts, L. (2016). Deep mapping and spatial anthropology. *Humanities, 5*(1), 5. https://doi.org/10.3390/h5010005.

Robertson, S. (2012). Putting Harlem on the map. In J. Dougherty & K. Nawrotzki (Eds.), *Writing history in the digital age*. Ann Arbor, MI: University of Michigan Press.

Stadler, J. (2015). Conceptualizing and mapping geocultural space. *International Journal of Humanities and Arts Computing, 9*(2), 133–141.

Streiter, O. (2015). Carvers and epigraphic practices on the Penghu archipelago. *DRGT2015, Second Workshop on Documenting and Researching Gravesites in Taiwan*, 8–44. Taipei, December 11–13.

Streiter, O., & Goudin, Y. (2013). The tanghao on Taiwan's tombstones. *Archivi Orientalni, 81*(3), 459–494.

Streiter, O., Goudin, Y., & Huang, C. (2011). Thakbong, digitizing Taiwan's tombstones for teaching, research and documentation. In *TELDAP 2010 (the international conference on Taiwan E-learning And Digital Archives Program)* (pp. 146–157). Taipei, Taiwan: TELDAP Proceedings.

Streiter, O., & Morris, J. X. (2015). *Researching Taiwan's gravesites with ThakBong and R*. Paper presented in the ECAI workshop at the Pacific Neighborhood Consortium (PNC) Conference Annual Conference and Joint Meetings, Taking Data into the Public Domain, University of Macau, September 27–30.

Tan, J. (2016). *Economic landscape of Taiwan: Dynamic cluster models for public policy effectiveness*. Doctoral thesis, International Doctoral Program in Asia-Pacific Studies, National Chengchi University, Taipei, Taiwan.

Tan, J., & Jan, J. F. (2016). *Taiwan's industrial development models: An inductive research from Taiwan's economic dynamics by utilizing cluster mapping framework*. Paper presented at the Pacific Neighborhood Consortium (PNC) Annual Conference and Joint Meetings, Does Data Construct Reality? The Getty Center, Los Angeles, California, August 16–18.

White, R. (2010). What is spatial history? *Spatial History Lab: Working Paper*. Stanford University. Retrieved December 1, 2016, from https://web.stanford.edu/group/spatialhistory/cgi-bin/site/pub.php?id=29

Zerneke, J. (2014). The Atlas of Maritime Buddhism technical infrastructure for collaborative development. In *Buddhist culture and technology: New strategies for study*, *Vietnam Buddhist University series* (Vol. 26, pp. 31–40). Berkeley, CA: ECAI http://ecai.org/projects/MaritimeBuddhism.html.

# Chapter 15
# Cloud Computing in Social Sciences and Humanities


**Michael J. Gallagher**

## 15.1 Introduction

Fortunately many new innovations are becoming available which can bring the computer power associated with a government laboratory into the hands of anyone. A number of cloud based, virtual, work spaces have been developed which can bring enormous computing advantages to any desktop or laptop. There is an initial steep learning curve associated with developing a high performance computing environment, and this obstacle prevents many who stand to benefit the most from using the cloud. This section seeks to demystify the process and get the researcher up and running very quickly.

Cloud based computing has tremendous potential for academia. Increasingly we live in a quantitative world. There is more and more data available every moment. The ability to read, store, and manipulate this data is of paramount importance to be on the cutting edge of research. Yet, many in academia are not availing themselves of the potential. There are a couple of reasons for this.

Firstly, generally all cloud based platforms are built for the greatest audience. In most cases they are built for computer developers, and made available to everyone else, but the architecture comes from a programmer perspective. While most in academic research have extensive programming skills, we generally are not computer scientists. Because the vendors of virtual server space are targeting such a wide user base, the instructions to get up and running are often overwhelming. One must go through a seemingly infinite tree of choices and pages of instructions. This requires a significant time investment.

M. J. Gallagher (✉)
Department of Finance, St. Bonaventure University, St. Bonaventure, NY, USA
e-mail: mgallagh@sbu.edu



S.-H. Chen (ed.), *Big Data in Computational Social Science and Humanities*,
Computational Social Sciences, https://doi.org/10.1007/978-3-319-95465-3_15

Secondly, the cost may at first seem prohibitive. Generally a credit card is required, and many are reluctant to use a personal credit card, and are unable to use a department credit card. In fact, the cost initially, while one is learning, could be substantial. One must be sure to stop a virtual server when it is not being used, at the same time making sure not to delete it. More than once I have been called away, left my server running, and all the while I was dealing with some other emergency the clock was ticking. However, all of this is mute when one becomes skilled at using the service. Once you are proficient, doing computational chores that would take days on a desktop can be done in minutes, and as we all know, time is money. Another advantage of the cost structure of most vendors is that you only pay for what you use, whether your needs are computational, storage or whatever you need.

## 15.2 A Revolution for Computational Social Science and Humanities

The amount of data available today is staggering! Almost everything we do is tracked and recorded. The widespread use of social media and smart phones leaves a digital trail of all our interests, internet searches, banking activity, shopping habits, eating habits, entertainment choices, etc. Simultaneously, the computational power available to manage this data, to extract information previously unavailable, has become readily available to researchers in all the Social Sciences. Anyone who has used the popular social media site Facebook knows if you are signed into Facebook and on the same computer do a search for beach umbrellas, your Facebook experience will then be plastered with advertisements for beach umbrellas. Similarly, Google keeps track of your preferences by tracking the websites you choose to visit and then tailors the results of future searches. Quite literally, every internet user's browsing experience is shaped by their browsing history. The news we read is news with a political and philosophical bent we have demonstrated appeals to us.

All of us leave a continual trail of digital footprints. All the texts, phone calls, GPS data, mouse clicks on websites, credit card transactions, photos, automobile technology, television choices, twitter, Instagram, Facebook, supermarket barcode scanners, all this information data is collected and stored.

According to Dan Gardner in the introduction to *The Human Face of Big Data*, "But, with a vast storehouse of our past decisions to analyze, it could detect patterns of behavior we are not aware of, and those patterns could reveal the unconscious thought processes that drive the behavior. In a very real sense, Big Data could know us better than we know ourselves. In that world, not only the buildings would be made of glass; so would our skulls" [1].

The ability to capture the information contained in this data stands to revolutionize research in the Social Sciences and Humanities. Already the impact is visible in the fields of business, finance, and economics; consider the advertisements you receive from Amazon, tailored specifically for you. Long held theoretical constructs, so-called "sacred cows" of academia, will be challenged with the proliferation of information.

Fortunately, the computing performance required to tackle these challenges is simultaneously being developed. There are a variety of computing platforms available to the researcher in the Social Sciences and Humanities, what remains is to bring the expertise required to harness the power of the cloud to the researcher.

## 15.3 Cloud Based Alternatives

There are a number of viable cloud computing solutions on the market. As one would expect in a relatively new industry, vendor competition brings a wide variety of slightly different products to the market. The last section of this chapter briefly outlines the steps to develop a multicore processor environment on Amazon Web Services (AWS), hence we will leave a discussion of AWS until then.

Google Cloud Platform (GCP) is a major contender in the market. Google currently offers a 60-day free trial period. At the top of the list of products available from Google is their Application Program Interface (API) Manager. The API Manager boasts more than 100 APIs. Every imaginable program interface is available. The list begins with the Compute Engine API which provides virtual machines for large-scale data processing and analytics applications. The wide range of choices include Cloud Storage, Social media, Blogger APIs, Map applications for all operating systems, You Tube applications, Calendar APIs, Advertising and search interfaces, messaging and gaming, translation features, all with Software Developer Kit accessibility. The point and click methodology of the Google Cloud Platform makes it easy to use, and the liberal use of "how to" videos makes it a natural choice for researchers in the Social Sciences.

Microsoft offers a cloud platform. On the one hand, they offer a cloud platform which is largely targeted to businesses; they tout the benefits of increased security, and the proliferation of cyber-crime as an element of the digital age which we all must be cognizant of. On the other hand, they also offer Microsoft Azure. Azure is an integrated cloud services platform which provides analytics, computing, database, mobile, networking, storage, and web services. Azure seems to be marketed to the developer or IT professional community; and it appears to be primarily designed for the software developer. They do offer a myriad of products and especially the elasticity of applications, so it is well worth exploring in greater detail.

Hewlett Packard Enterprises are also in the market; at first glance, they seem to follow in the footsteps of Microsoft in that they seem to cater to bigger industry, not as much to the individual researcher. As a matter of fact the first page to open on their website is a cry about the dangers of cyber security. Next they announce the benefits of moving to a hybrid structure, in other words, using your existing architecture in conjunction with the cloud based environment. Having said all of this, Hewlett Packard does also proclaim the importance of embracing big data, and that big data analytics are the way of the future. It just is not clear from an initial

investigation of the products and services offered on the company website that it would be a best choice for an academic researcher.

Amazon Web Services has a plethora of cloud based applications available. Amazon was one of the first in the market, and the following section details setting up a high performance computing environment using the Elastic Cloud Compute platform offered by Amazon.

## 15.4 Setting Up a Multicore Environment on Amazon

Open an account at Amazon Web Services: http://aws.amazon.com/. There are detailed instructions for getting up and running, there is extensive documentation, and there is a myriad of resources available. Once an account is established at Amazon, the user can access preconfigured Amazon Machine Images (AMIs). An AMI is simply a virtual machine with particular operating system and preloaded software. This is analogous to going to a bricks and mortar store and requesting a computer with certain specifications and software already loaded. There are AMIs for every conceivable configuration. Preconfigured AMIs with the programming language R, R Studio, and Message Passing Interface (MPI) are available on my website (http://www.michael-j-gallagher.com/#!high-performance-computing/c1kr2) where there are links for multiple geographic locations.

Message Passing Interface (MPI) Standard is the crux of using a high performance computing environment. Normally computing environments are designed where each node, or core, operates in serial, in other words, a process moves from node to node. MPI essentially sends out the process to all available nodes at once. This is ideally suited to computational tasks which are not sequence dependent, for example, the bootstrapping requirements of nonparametric econometrics, because each repetitive evolution of a process is completely independent of the results of the previous process.

MPI was developed in the early 1990s, originally based on the Chameleon portability system, originally denoted with the acronym MPICH. The development has gone through three iterations MPICH, MPICH2, and MPICH3. Now, support for MPICH has been discontinued, and the de facto standard in the industry is the open source software, Open MPI (Open MPI is readily available at: http://www.open-mpi.org/).

To communicate with the virtual environment, it is necessary to have a secure shell (SSH) and a secure file transfer protocol (SFTP) client. I recommend PuTTY, also open source and freely available on the internet (PuTTY is available here: http://www.chiark.greenend.org.uk/~sgtatham/putty/). PuTTY must be installed on the local machine. Install all the binaries. PuTTY provides two essential ingredients. One is a secure shell from which the user will invoke the computer nodes that have been incorporated into this environment, execute command files, and otherwise fine-tune the environment for the specific purposes of the researcher. Secondly, through

the Windows Command shell, PuTTY Secure File Transfer Protocol, (PSFTP) allows a secondary method to upload/download data files to and from the server.

An encryption key pair is required to access and utilize the server. Amazon provides detailed instruction for generating keys, and saving them. Keys must be saved in two formats, a ".pem" format and a ".ppk" format. The ".pem" file will enable passwordless communication between the computing nodes, which is essential for MPI. The ".ppk" file is necessary to allow PuTTY to open a command shell to the server. Both are necessary, so both files should be carefully saved in the directory from which the server will be accessed. It is also good practice to save the keys in a folder separate from the working directory, for example a "keys folder" in the event of some corruption in the working directory

Now the server is ready to be accessed and a few remaining steps are necessary to configure the multi-core environment. Choose an Amazon Machine Image (AMI) which corresponds geographically to your region from the list on my web page. Choosing an AMI is the first step of the prompts on Amazon to set up your environment. The AMIs on my website are Linux Ubuntu operating systems. The important feature of the AMIs I have created on Amazon is that R, RStudio Server and Open MPI are all installed and ready to go. The researcher will not need to install anything, or set paths, or otherwise configure the environment. I have chosen a Linux environment because the software available, particularly open source software is written in C or C++, so a Linux environment seems to allow a more fluid transition than a windows environment.

Choose an instance type. I have been using a Compute optimized "cc2.8xlarge" which is a computing optimized instance designed for cluster computing. This instance is available under the all generations drop down choice displayed on the AWS console. This allows multiple servers to be physically located next to each other, which facilitates Message Passing Interface. Certainly, they do not need to be in proximity to each other to work, but why not make the job easier for the computer. I believe any of the newer generations should work as well; I just have not tested them. The exception to all of this is the free tier, on which I was unable to establish node to node communication.

The next step is to configure the instance; choose the number of instances. Recall instances are simply virtual machines. I usually run 4 but there is no reason not to run more, however bear in mind that each instance is additional cost, currently the spot rate has been around $0.30 per hour, but more importantly, there is a point of diminishing returns. With small amounts of data, it can take longer to push out the data to all the nodes than it can take one processor to do the work. The cc2.8 x-large instance has 16 cores, actually 16 cores and 16 hyper threads, along with 32 GB Ram, so, running four in parallel is equivalent to a 64 core machine with 128 GB RAM, pretty much your own little super computer. There are two methods to purchase time on AWS servers. "On demand" means purchasing time which will not be interrupted. Also another option is "spot" instances; this can be a significant savings over purchasing on demand instances. Spot instances allow bidding for time on the server. While this can be a large savings over the on demand rates, the researcher must be willing to allow the work to be interrupted.

**Table 15.1** Security group inbound rules

| Type | Protocol | Port | Range source |
|---|---|---|---|
| SSH | TCP | 22 | Local IP address |
| All TCP | TCP | 0–65,535 | Security group ID |
| Custom TCP rule | TCP | 8787 | Local IP address |
| HTTP | TCP | 80 | Local IP address |

Configure a security group. Here the focus is on inbound rules; only the IP address of the local machine should be able to access the server. The server should not be open to the world for just anyone to have access. A security group with inbound rules as in Table 15.1 will allow only the local machine access.

The security group rules in the above table satisfy the following: SSH on port 22 and HTTP on port 80 allow communication from the local machine to the server. A custom rule on port 8787 allows you to communicate with your server through R Studio Server. Finally when you save and name your security group you will have a Security Group ID which will allow your compute nodes to communicate with each other on all ports, this is essential for MPI.

Launch the instance. Amazon will notify the user when the instance is available, which may take a few minutes. Then the researcher will access the server through a PuTTY shell. A minimum of Linux skills, which are discussed here, are necessary. The Ubuntu command line prompt will appear like this, where XXX is the IP address of the instance.

ubuntu@ip-XXX-XX-XX-XX:~$

The instance contains a file called the hosts file which contains information about the internet addresses of each instance associated with the server the researcher is using. This file must be edited to include the other instances in this little cluster of virtual machines being created. Use an elementary Linux editor such as "nano," and the Linux command "sudo" to allow you to edit a file without changing file ownership. In other words, at the command line prompt, type:

ubuntu@ip-XXX-XX-XX-XX: ~$ *sudo nano/home/ubuntu/etc/hosts*

This will open the hosts file and allow the researcher to edit it. Add the internal IP addresses for the instances that have been created and name them, perhaps; node1, node2, etc. The IP addresses are provided on the Amazon console. Save the edited hosts file. The editor "nano" provides keystroke instructions for saving edited files, leave the name the same. Next a file must be created with information about each of the instances which are part of this cluster. Create a file with the configuration of the instances you have selected. I called mine "nodefile," for example if you have chosen a "cc2.8xlarge" in which each of the four instances in this configuration contains 16 cores, the "nodefile" will look like this:

node1 slots = 16
node2 slots = 16
node3 slots = 16

node4 slots = 16

To create such a file, simply use nano again:

ubuntu@ip-XXX-XX-XX-XX:~$ *sudo nano nodefile*

Then type the number of cores for each nodes based on the type and number of instances that were created when the instance was configured. This must be done on all nodes, or this nodefile may be pushed out from the master node using a file called "extend," which is on the image and will send out whatever files desired to all nodes. The file "extend" should be edited to reflect the number of nodes you are using. To use extend, at the command prompt, simply type:

ubuntu@ip-XXX-XX-XX-XX: ~$ *./extend/nodefile*

Passwordless SSH must be enabled between all the nodes, the easiest way to do this is to put the ".pem" file on each node. The ".pem" file was generated when the researcher opened an account at AWS. This can be done through a windows command shell using PuTTY SFTP, PuTTY Secure File Transfer Protocol. From the PuTTY directory in the windows command shell type:

*psftp.exe ubuntu@XX.XX.XX.XX -i* {*your putty-key-pair-name*}*.ppk*

where XX.XX.XX.XX is the public IP address of your instance, and {your putty-key-pair-name}.ppk is the PuTTY key pair generated when the AWS account was established, as mentioned above. When the connection is accessed, using the "put" / "get" Linux command, from "PSFTP" write: put {*your key file.pem*} /home/ubuntu/.ssh/id_rsa

This is the default Linux ubuntu password. This needs to be done on all compute nodes.

While you have PSFTP open, you may also upload data files or R script files, or you may do this from the R Studio Server GUI which is described next. R Studio Server does not currently facilitate parallel processing, therefore, parallel jobs must be run through an R command line batch file or script file from the PuTTY window on node1. The results from this batch file will of course then be available to the researcher on R Studio Server. The command line command will vary by application, but will look like this:

mpiexec -n 64 -hostfile ~/nodefile R -save -q < learnmpi.R > output.txt

This command will invoke the 64 node virtual machine which the user has previously setup, invoke the hostfile and nodefile the user has created, start and quit R, run a batch file named "learnmpi.R," save an R data file if the batch file contains a save command, and also save an output text file named "output.txt." The file "learnmpi.R" is available on my website along with a small data set text file called "624short.txt." These files can be opened and edited in any notepad or R Graphic User Interface. These files should allow the researcher to get acquainted with parallel processing in very little time. The user must run their own R code in place of "learnmpi.R," this file is meant to demonstrate the syntax required to

run a parallel job. As was mentioned the parallel job must be run from a PuTTY command line, the results of which can be accessed through a windows command line via PSFTP, or directly in R Studio Server, as discussed next.

## 15.5 Launching R-Studio Server

From the PuTTY command shell prompt:

ubuntu@ip-XXX-XX-XX-XX:~$
type: {*sudo adduser username*} *(where username is a username of your choice)*

Choose a user name, all lowercase, and a password which is typed in twice. Now the user should open a web browser window and put the public IP address of node1 followed by the default port for R Studio Server ":8787" in the browser address bar. In other words: XX.XX.XX.XX:8787 This will open an RStudio GUI sign in page, use the username and password just created in the PuTTY command window and the researcher is ready to harness the power of the cloud!

Developing a virtual high performance computing environment on Amazon Web Services, and harnessing the computational resources available, can dramatically speed up computational chores. Consider stringing together four virtual machines with 16 cores and 32 GB RAM each, to provide a 64 core, 128 GB RAM environment. Compare this to the average desktop which usually contains 8 GB of RAM. A computational task a researcher may have run for a week can now be processed in a matter of minutes. The ability to manage, analyze, and understand Big Data, by using the resources of the cloud, is now available to the academic researcher in the Social Sciences and Humanities.

## Reference

1. Smolan, R. (2012). *The human face of big data* (1st ed.). Sausalito: Against All Odds Productions.

# Chapter 16
# Analysis of Social Media Data: An Introduction to the Characteristics and Chronological Process


**Pai-Lin Chen, Yu-Chung Cheng, and Kung Chen**

## 16.1 Introduction

Mass communication scholars often refer to the impact social media has on the mass communication ecosystem. Not only would scholars like to have a greater understanding of social media content, but also the government and enterprises, who would like to see the developing trends of social consensus and the ongoing movement of consumers through social media content analysis. However, social media data analysis has characteristics of big data, which differs from the data analysis methods traditionally used by social science and previously applied by mass communication scholars. Consequently, social media data analysis is a current area of academic interest.

Even though many scholars believe the application of big data methods is a necessary trend (boyd and Ellison 2007; Mahrt and Scharkow 2013; Parks 2014), results put forth by the discipline of mass communication indicate that research articles concerning data-mining and the methods in which social media is analyzed are still rudimentary (boyd and Crawford 2012). The field of social media data analysis demands additional academic attention and contribution. This preliminary chapter seeks to describe the characteristics, elements, and the chronological process of social media data analysis from a mass communication scholar's perspective. Through case study, this chapter seeks to present ways a researcher analyzes social media data, and how that researcher poses questions and deals with the data during the process.

P.-L. Chen (✉)
College of Communication, National Chengchi University, Taipei, Taiwan

Y.-C. Cheng
Hsuan Chuang University, Hsinchu, Taiwan

K. Chen
Department of Computer Science, National Chengchi University, Taipei, Taiwan



S.-H. Chen (ed.), *Big Data in Computational Social Science and Humanities*,
Computational Social Sciences, https://doi.org/10.1007/978-3-319-95465-3_16

## 16.2 The Data Analysis of Social Media

Social media appearing at the start of the twenty-first century provides "web-based services that allow individuals to (1) construct a public or semi-public profile within a bounded system, (2) articulate a list of other users with whom they share a connection, and (3) view and traverse their list of connections and those made by others within the system" (boyd and Ellison 2007, p. 211). Facebook, Twitter, and PTT can be viewed as different types of social media. We view photographs and watch videos uploaded by our social media friends; we leave comments on social media posts; and we paste and/or share useful or interesting messages. In light of this, social media's information producing mechanism is very different from previous mediums of mass communication. It refers to platform mechanism for "a conversational, distributed mode of content generation, dissemination, and communication among communities" (Zeng et al. 2010, p. 13). The birth of social media not only revolutionizes the way people share information, but also hugely impacts mass communication research.

When a hundred thousand audiences use social media to produce, cooperate, and share every different type of message, the messages form a huge unprecedented dataset through the automatic categorization, record, and preservation by the media platform. The data generated by social media not only includes the content produced by human beings, but also comprises of metadata produced by machines. These data sets are huge in quantity and variety, which Lev Manovich (2012, p. 2) called "Social big data." The content of the data carried by social media results from the huge amount of facts, opinions, imagination, and feelings people have produced (Yang and Hhao 2016, p. 2). It provides a huge database that can be used as the target for collecting and analyzing (Stieglit and Dang-Xuan 2012). Many scholars, entrepreneurs, politicians and media workers seek to discover a social, political, cultural and/or industrial niche within the enormous data set. Tufekci (2014, p. 1) provides a vivid analogy: "the emergence of big data from social media has had impacts in the study of human behavior similar to the introduction of the microscope or the telescope in the fields of biology and astronomy." This metaphor points out how the birth of big data has brought a qualitative change to the research of social science: the thing the birth of big data changes is not just the scale of analysis, but also its vision and depth.

The birth of social media analytics is a field of knowledge that corresponds with the birth of social media. For scholars of mass communication, the birth of big data brings a profound change toward research method paradigm. The greatest significance of big data is not only the sudden multiplication of the quantity and scale of research data, but also the way mass communication scholars (boyd and Crawford 2012, p. 663).

Social media analytics is an emerging field to which scholars have provided different definitions. For example, Zeng et al. (2010, p. 14) maintain that social media analytics is a set skills for "developing and evaluating informatics tools and frameworks to collect, monitor, analyze, summarize, and visualize social media data, usually driven by specific requirements from a target application." Whereas

Stieglitz et al. 2014, p. 90 argue that social media analytics is "an approach toward research that involves multiple disciplines of knowledge." The mentioned scholars have not only provided the methodological foundation to other scholars from the perspective of different disciplines, such as business management, economics, and sociology, but also collected, mined, and analyzed the data, ultimately using large-scale social media to construct a data model intended to solve problems posed by the given academic or practical field.

Social media and data analysis can be viewed through two models when concerning contemporary society's ability to process information. A Nobel Prize winner in economics Daniel Kahneman points out that there are two thinking or decision-making process models: first, there is the model of thinking or decision-making process which generates responses immediately after the event (system I), second, there is the model of thinking or decision-making process that makes decisions through deliberation (system II). Respectively known as "fast thinking" and "slow thinking" (Kahneman 2012). On the one hand, social media of contemporary society has generated large-scale, diverse, and highly dense data which is conducive to the "reactive system" enabled by contemporary society. On the other hand, the data analysis of social media, conducted by researchers through applied methods in data collecting, filtering, and analyzing suggest a social reality rather similar to the "reflective system" Social media and its data analysis parallels the push and pull which maintains the equilibrium of social information embodied by "fast thinking" and "slow thinking."

## 16.3 The Challenge of Analyzing Social Big Data

Upon reviewing research articles from different disciplines, we discover that social big data analysis has faced various challenges including: the enormity and complexity of the data, the difficulty applying automated technologies when processing a dataset derived through human behavior or human expression, not to mention the questionable completeness and transparency of a data set still in its initial stages.

### *16.3.1 The Enormity of Amount and Scale*

The quantity and scale of social data is large, the data types are diverse and in continual output. The preferred network platform in the age of web 2.0 is social media, allowing users to both produce and share data (Brügger and Finnemann 2013, p. 78). Data aggregated on social platforms include not only digital content, but also meta-data that characterizes the data (Cheng 2014, p. 80). Therefore, the amount and scale of social media accumulated data is bigger when compared to traditional media platforms. This is why researchers often fail to process this data when applying existing and traditional research methods (Stieglitz et al. 2014, p. 91; Boumans and Trilling 2016).

### *16.3.2 Human Generation*

Social data is generated by a mechanism of human-generated computing. This type of data largely comes from the action of human languages and the human-machine interaction. This is different from other forms of instrument measured big data such as: outdoor temperature, accumulated rainfall, or atmospheric particulates observed.

Most social data includes two kinds of data, meta-data and content data. Meta-data is the "data that describes data," for example, the user account, the time of post, and the serial number of articles. They are usually in list form, and created by the computer system. Analysis of the content data is much more difficult. First, a lot of human and material resources are required to clean the data, because of the way human's use language and the diversity of data formats (a great amount of titles, keywords, tags, emotional symbols, plus the content of streamlined audio/video files can be included along with the textual content), the content's attributes are highly disparate; the connotations of the text are very complex. Second, there are several different ways of using human language: opinions, evaluation, irony, etc. One word may contain multiple meanings. Although researchers can look for the patterns of text through data science techniques (such as text-mining or machine learning), the intended meaning of a given text is still difficult to grasp. Third, although many people believe that patterns of online interaction are reflected in social media data, online interaction are interconnected with social situation, and the meanings of human interactions are complex. It is an oversimplification to interpret the meanings of complex human interactions from limiting, textually derived data accumulated from social media platforms. For example, clicking the "like" or "share" buttons on Facebook, statistics on the number of online messages, or the centrality of a social network connection. In this light, the degree to which automated techniques help people understand social media interaction is still very limited.

### *16.3.3 The Integrity and Transparency of Data*

The other challenge facing social media analysis is data integrity. Social media data is one of the most important forms of revenue for social media companies. Social media makes social media data profitable by selling it. Therefore, social media companies restrict access to all their data. The data released through the API is a very small amount of the total data. For example, Twitter only releases 1% of its total data through the API. Payment is required to access the rest. Social media owns and safeguards the big data; the only organizations who can use social media data in its entirety are the enterprises and government organizations who can pay the high fees for big data access, as noted by Lev Manovich (2012).

Data analysts from academic institutions and average size enterprises must accumulate their own materials from data released by social media platforms, or hire data mining companies to provide the materials for them. When data is collected by

a third party, the integrity of the data is often linked to the technical capabilities of that party. For instance, social media is composed of many platforms; therefore, data sources vary and weighting among the data is difficult. An ongoing challenge is whether a "cross-the-board" analysis of Facebook, Twitter, and YouTube, is credible.

Besides data integrity, the other dimension worth noticing is data transparency. One criterion for scientific research quality is whether the data can be reproduced, whether the same result can be achieved by the same procedure. Thus, the procedures of data mining or analysis must be transparent. In all the fields collecting social media data, transparency is exists as debates regarding whether the algorithmic mechanism should be open source. The owners of contemporary social media platforms use algorithmic mechanisms to mine and release data. Due to this, data suppliers (the external data collectors) are incapable of knowing the algorithmic content of social media platforms through their own data crawling, or from purchasing data from social media platforms. Social media companies often consider their algorithms a market competitive business secret. Consequently, such companies will not allow open sourcing. Therefore, the degree of data transparency is insufficient, researchers cannot measure the rationality backing the data collecting procedures nor the integrity of the data accumulated through the algorithmic mechanism.

Due to the above reasons, the analytical requirements of collecting social media data are relatively high. Therefore, there is currently relatively little research based on empirical data. According to the meta-analysis made by Felt (2016, p. 4–5), in the 294 social media articles collected in the communication studies database *Communication and Mass Media Complete*, 83% still applied traditional data collection methods, only 17% of them collected data using information science techniques. The articles using traditional data collection methods largely used content analysis methods (21%) and survey questionnaires (20%; especially the online questionnaire). Not to mention, the same research found that the social media data collected was primarily from Facebook (69%) and Twitter (46%). There was infrequent transplatform research (<10%). According to the bibliographical review of Chiang and Lin (2015), of the 39 articles they sorted from the SSCI database, only 12 articles used numerical data for empirical research (the rest were mainly articles on theoretical or conceptual analysis). There were only two articles in the domestic TSSCI database in which only one article used data analysis methods. The research mentioned above also analyzed the number of authors. It showed that big data research is often coauthored by many contributors reflecting the nascent state of this field, and the necessity of teamwork due to its transdisciplinary nature.

### *16.3.4 Summary*

As we mentioned above, traditional research approaches have difficulty handling large-scale, multi-material, multiform data filled with various kinds of noise. However, through the assistance of data science, social media related questions

posed by communication scholars can be solved. This means social media analysis possesses two processing qualities.

## 16.4 The Dual Processes of Social Media Analysis

Social media analysis bears the characteristics of a dual process. On the one hand, communication scholars need to develop problems worth probing into. Followed by discovering an answer through viewing what comes forth after data processing. On the other hand, there is a definite procedure for data processing. Researchers will collect, clean, and present data through visualization, in accordance with the posed problems. We will start by delineating the contents of these dual processes. Then we will demonstrate the relationship between them.

Data analysis is essentially the process "from problem formulation to problem solving." Academic researchers and practical workers both use social media data analysis to solve problems. While communication scholars may focus on questions like, how do messages of tremendous catastrophes communicate, disseminate, and/or converge through a social media platform (e.g., Chen and Cheng 2014)? Public relations employees in corporations or organizations use social media data to understand trends of consensus toward particular events, and take that data as a reference for maintaining the image of a brand or strategically handling a crisis (Li 2011). Although academic scholars and practical workers have different problems, question/answer strategies, meticulousness, and expectations toward analysis, both require consideration of data related questions and problems, as well as the process of finding an answer through social media data.

From the perspective of posing and positioning problems, social media data analysis is not much different from other academic or practical research processes. The real difference between other research methods and social media data analysis, which includes the subject of big data, are the details in handling the numerical data. As we mentioned above, social media data is generated by human beings. Social media data is not only in large amount, multiple in form, and filled with noise, but also incomplete. Therefore, not only does the problem formulation to problem-solving process require data science knowledge but the data processing process also requires further explanation.

### *16.4.1 Data Processing*

The second factor in is data processing. Data processing includes five elements: (1) Data/Metadata: metadata is the data about data, or the structural message of data. For instance, cookies or user ID, data generated geographical, temporal information. (2) Algorithm: algorithm means that platforms will use a formula and variables to calculate social interaction. The algorithm decides the competitiveness of the

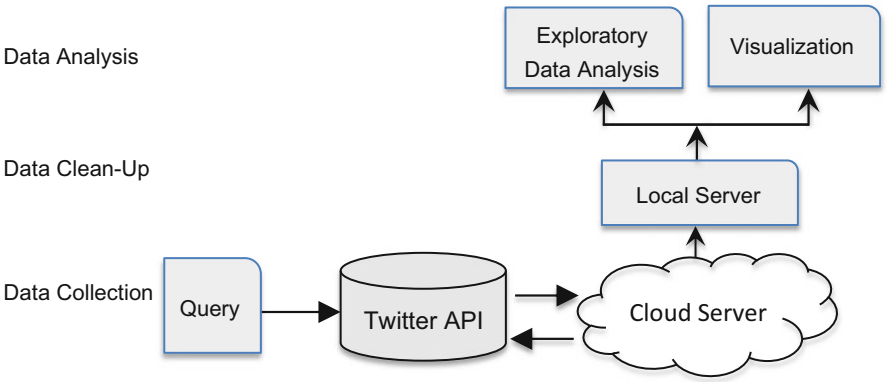



**Fig. 16.1** A basic procedure to process data from social media platform

social platforms such as business secrets. (3) Protocol: protocol unites the data formats of different systems and implicitly directs user behavior in a manager-favored direction; (4) Interface: visible interface means the end user interface which is iconized and easy to use, the invisible interface is the one which is used to connect the hardware and software, and the API is the one between the visible interface and invisible interface; and (5) Default: the software has the function of directing the user (van Dijck 2013). We can induct social media data processing as the stage of data collection, the stage of data cleanup, and the stage of data analysis as per the following (Fig. 16.1).

### *16.4.2 Data Collection*

Data collection is the initial stage of processing social media data. This stage is mainly about the process of tracking, monitoring, mining, cleaning, and arranging the digital footprints left by specific platforms or audiences, then making them into targets for analysis.

Because the quantity of social media data is big, the data-mining process often relies on scientific techniques which mine data through the automated software. The composite of specific data obtained through the information technology mining process is called the dataset. Generally speaking, there are two ways of obtaining a social media dataset: that is, you can access the dataset by API, or you can access the dataset by parsing the RSS/HTMAL.

The so-called "access by API" is the process in which researchers write the program, and log into the API created by the owners of the social media platform in accordance to specifications of different columns and limits provided by social media. One can set several vocabularies in the program, mine the data from social media servers, download and save data in a database, and wait for the subsequent data cleaning up and analysis (Fig. 16.2).

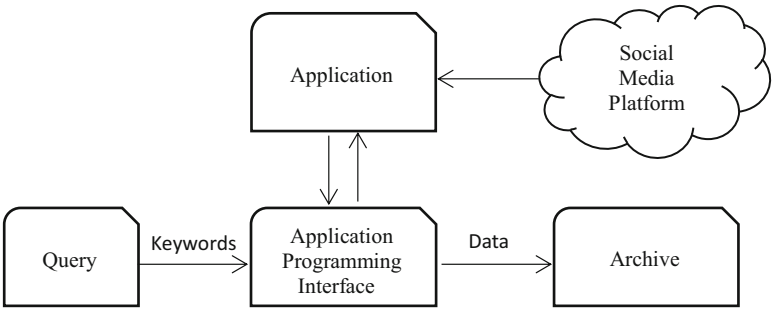



**Fig. 16.2** Assessment by API of the social media platform's application interface

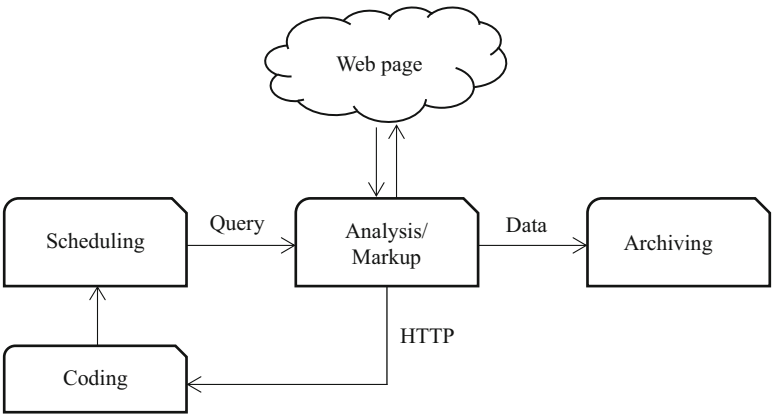



**Fig. 16.3** The process of data crawling from the social media platform

On the other hand, the so-called "RSS/HTMAL parsing" is a program that attempts to simulate the human act of browsing a webpage, or the process that attempts to mine data from webpages of social media websites. Concerning the webpage of social media platforms, users can write web crawler programs to simulate human webpage browsing behavior. The schedule of these web crawling programs can often be set, manipulated to send messages to specific websites, and used to mine data from a specific section of the website through one-pass programming. Once the web crawler program has crawled the data (text and metadata), it downloads it to the database, analyzes them, and does parse/markup before saving them in a database, awaiting the cleanup and analysis of researchers. Because web crawling involves arranging the format in accordance with the webpage, the web crawler program must be updated whenever the social media platform owners change the webpage format (Fig. 16.3).

For researchers, whether to use API or Web crawler is a difficult decision to make. APIs usually set ceilings of the data volumes; and the versions of access protocol vary from time to time. On the other hands, web crawling imitates human

actions in browsing web pages, the right to access may be taken over when its behavior is considered "hostile" by the Platform. The researchers have to choose either method and live with it.

What data to collect? Bruns et al. (2013) pointed four major types of data used in Twitter analysis, namely keywords, Hashtags, Mentions, and URL. Particularly, researchers dealing with data analytics on significant public events usually use the above data sets. Among other ways, keywords are the most common way in accessing social media data. Researchers can identify or mine the targeted data using one or several social media keywords. Keywords are usually used for data crawling specific events (e.g., the presidential election, social movements, or catastrophes). The keyword's setting either comes from researcher experience, or from judgments made after reviewing the data. For example, collecting data on airline accidents using the airline company name, flight number, or accident location. Using words like "Malaysia Airlines" or "MH370" targets the March 2014 Malaysia Airlines aircraft disappearance.

The keywords used for data crawling are not only the keywords researchers set in relation the research goal, but also the hashtags provided by social media platforms. Hashtags are phrases set by social media users in accordance with a communication goal. They are often used to arch, group, and connect social issues. For instance, users of Twitter use #Orlando to mark the Orlando, Florida gunshot incident in June, 2016. By comparing the keyword/hashtag amounts during a specific period, researchers can understand the Orlando accident's social media trend and area of discussion.

An ever-lasting challenge to the social science researchers is: how to solve their specific research questions in this limited metrics of data?.

### *16.4.3 Data Cleanup*

Every dataset can include characters, numbers, icons, or other formats. Because social media data contains a very large amount of non-structural data, researchers must clean the data.

The goal of data cleaning is to unify the format, reduce noise and convert data into an easy-to-process scale. It is the most time-consuming stage of data processing. As we mentioned above, social media data is generated by human beings. The content of social media data is mainly non-structural. Social media data must be converted into a machine-readable format before being processed by automatic information technology.

We can induct the data cleaning procedures as follows: format processing, selection/filtering, and integration or separation.

#### 16.4.3.1 Data Format

Researchers often transform or arrange the data format into a desirable state. For example, the time of a social media post is usually recorded as a 15-number string. The string is converted into year/month/day by processing different sections. Another example is that much textual information—which is saved in Big-5 format—must be converted into UTF-8 format for convenient computing.

For analysis usage, data usually requires a cut after processing the data format. The cut is the handling process of data selection, filtering, integration, separation, etc. Due to this process, selection and filtering of data are vital.

#### 16.4.3.2 Data Selection

After processing the format, the dataset is arranged like one to many matrixes. It can be saved as several tables. Each table stores its social media data in column and row. Social media data has many dimensions, for instance, the user account, the post time, and the message content. These dimensions are each stored in a vertical column. All columns are arranged from left to right. The so-called data selection is when researchers identify and preserve corresponding columns keeping with analysis requirements.

#### 16.4.3.3 Data Filtering/Sifting

Every dataset contains several forms of data. Each is presented horizontally. Each presents the variables consistent with the sequence of individual columns. All data has a horizontally fixed, corresponding to the columns' sequence. The filtering/sifting of data is when researchers give instructions to save or delete data, according to the analysis requirement and the individual column's range of variables. For example, researchers only need to delete those appearing twice in the user number columns if they wish to delete those which repeat, so each user number appears only once.

#### 16.4.3.4 Data Integration/Separation

From the start, datasets researchers usually analyze contain no corresponding or analyzable columns. Due to this, researchers must combine or separate different columns, as well as create new columns for analysis, according to the researches problematic. The so-called data integration is when researchers combine several columns into one. For example, the "like," "drop your message here," and total number of times one picture or article is shared exists in three different columns. The researchers wish to merge the three columns into one variable by the weighting of the three, and saving them as one new column. Conversely, data separation is

when researchers separate a single column into several columns. For example, the post's time column contains the year, month, and day of the original post, which are separated into three individual columns, before analyzing only the month.

#### 16.4.3.5 Data Dimensions

The above actions such as data processing, selection, filtering, integration, or separation are used by researchers manipulating data, to make the data analyzable. During the process of updating social media data, the "data dimensions" are the datasets having the same properties. In social media data, the most common types of data dimensions are temporal types, spatial types, numeric types, categorical types, and relational types. For example, the time of the post is presented in the dataset in year/month/day columns. These columns co-present temporal dimensions. The post's location co-presents spatial dimensions in longitude and latitude columns. These dimensions are cleaned up and categorized as the desired goal the researchers wish to pursue.

However, not all the data dimensions correspond to the research questions. The dimensions are not immediately suitable for data analysis. To construct the research metrics, researchers usually organize the dimensions of the data according to particular research goals.

#### 16.4.3.6 Data Metrics

The so-called "data metrics" is a function formula constructed by the dimensions of the data in accordance with the goals of research. The metrics are usually used as research variables. For example, the Facebook monitor integrates the quantity of the clicked "like" button, the number of dropped messages, and the number of "shares" generated by users into a set of functions:

The engagement rate of the fans page per day = (the quantity of the clicked "like" button + the number of dropped messages + the number of "shares")/the number of fans that day × 100%.

The researchers require the "quantity of the clicked 'like' button," the "number of dropped messages," and the "number of 'shares.'" From the data dimensions, they crop, combine, and weigh the above information forming the "engagement rate of the fan page per day," this obtains the engagement rate between the fan page and its audience.

The above mentioned "engagement rate" is constructed by researchers. Although all the data have the same amounts, scale, or dimensions. The result is different dependent on the different ways of constructing the metrics (e.g., the data will be weighted according to different extents of the clicked "like" buttons, dropped messages, shares, and participation). Thus, the effectiveness of the metrics must be further evaluated and proved.

Regarding "engagement rate" use, when researchers consider them as equal value, the amounts of the "like," "share," and critics can be aggregated and present the above mentioned functions. However, the weighting or the other ways of data processing must be considered while integrating the three values mentioned above, if the researchers are to deem the difference between the three above mentioned activities' extent of participation (e.g., Mayfield 2006; Forte and Lampe 2013) and decide that the difference should be made when the quantitative value is put forth (e.g., the researchers decide that the quantity of those clicking "share" > quantity for those of the critic>the quantity of those clicking "like"). In other words, researchers construct a set of function formulas based on his or her theories to calculate the data. Thus, the result of the analysis might be different according to the different metrics of calculation the researchers apply based on his or her theories, even though the used data set is the same.

### *16.4.4 Data Analysis*

According to the goal of the data, there are two types of data analysis: exploratory analysis and argumentative analysis.

#### 16.4.4.1 Exploratory vs. Argumentative Analysis

Exploratory Data Analysis (EDA) is when researchers discover the state or trend of data distribution through simple statistic indexes and visualization tools, and consider the state or trend as the basis of further data analysis (Tukey 1977; Seltman 2015). The explorative data analysis method is often used in the initial stage of data processing, and when the data's connotation is unclear. Since the characteristics of big data are that the sample is very similar to the pool (Schönberger and Cukier 2013), and every dataset to possess its singularity, the state of the data requires exploration, and explorative data analysis is vital during the initial stage of analysis. The process is both explorative and hypothetical, just as its surface meaning suggests. It has an explorative meaning in which the researchers can make the state of the data surface through simple statistic methods and visualization tools. It also has the hypothetical meaning in which researchers try to observe the state of data distribution and the possible irregularities, concerning highly uncertain data, to find out key points for questioning. For example, the convergent or divergent trend of the quantity of posted articles can be observed by constructing a simple temporal model in reference to the time and the number of posts, when the critical public event breaks out. The temporal relationship of the trend can be visualized, allowing researchers to understand the opinions or trends of social media, making analysis possible.

On the other hand, the "argumentative data analysis" is the process when researchers hypothesize or prove the result of the initial exploration by ways of classifying, grouping, correlating, or predicting data, including the construction of statistical models through the application of various kinds of statistical tools. Besides, researchers can reduce the data's scale and undergo qualitative analysis according to results from the initial exploration of quantitative analysis, in order to discover the data's connotation.

#### 16.4.4.2 Types of Data Analysis

The analytical tools of social media data are very different from traditional social science analytical methods. This chapter aims to discuss the following types of data analysis: temporal/trend analysis, relational analysis, numerical/type analysis, text analysis, and sentiment analysis. The future types of data analysis will be increasingly diverse with the evolution of information science technology and research methods. The connotation of the different data analysis types are:

#### 16.4.4.3 Temporal/Trend Analysis

This type of data analysis aims to observe the transformation of social media data's variables in a particular period according to sequence of time by using time and other dimensions as the materials. For instance, researchers observe the transformation of the extent of concern of social media users by referring to the number of social media posts or the migration of frequency of particular texts posted during a catastrophic event (Jungherr 2014). The transformation of the quantity's scale under the sequences of time can be illustrated by considering sequences of time as the X variable, and frequency of the post as the Y variable. For instance, concerning the Tweets of the 2012 Australian presidential election, Burgess and Bruns (2012) discuss how people aggregate and the election on Twitter. They collect 41,500 posted articles containing the hashtag #ausvote for 38 days before and after the election, and compare the frequency of the posted articles on different days. The research found that 22% of the articles were posted on the day of the vote. However, only 1900 out of 3700 people (51%) posted articles during the past 38 days, also posted articles on the day of the vote. In other words, over half of the users gave election opinions on the day of the vote.

Temporal analysis can be used to predict the trend or pattern over an issues lifetime. To predict the developing trend of social media discussion (Stieglitz et al. 2014, p. 92), this kind of data analysis survey records the life pattern of social media discussion, and stores every kind of time sequence models in the database, according to the algorithmic models constructed from information science/statistics (e.g., hidden Markov model). The cross-strait data analysis for Twitter on the 2012 presidential election is presented on the graph below. The graph shows the distribution frequency of posted articles by three different languages groups

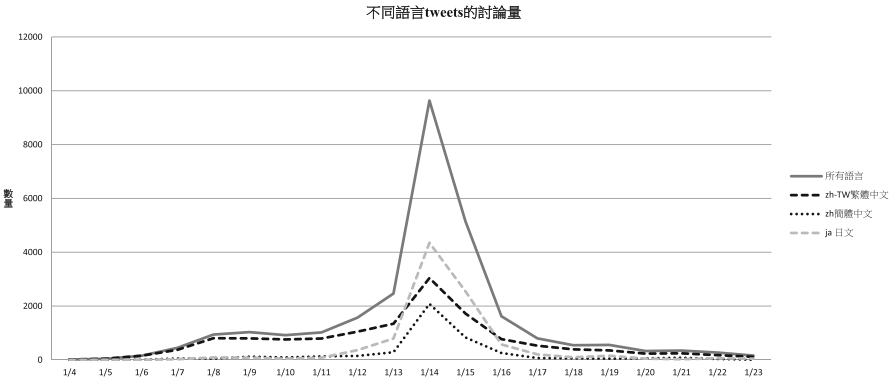



**Fig. 16.4** The time sequence analysis of tweets during the 2012 presidential election (Cheng and Chen 2014)

(traditional Chinese, simplified Chinese, and Japanese) 2 weeks before and after the election. The increase in public opinions of those who apply traditional Chinese is earlier than the other two language groups (Cheng and Chen 2014) (Fig. 16.4).

#### 16.4.4.4 Relational Analysis

Relational analysis uses relational data produced by the social media platform. The main goal is to observe the relational context among platform users. For example, when Twitter users retweet another user's article, that user's account is posted on the platform. This kind of record relates two users, and it can be considered as a node and link between them. The above type of data can be analyzed through mining and coding, in order to represent the social relation between those who post and those who receive the posts. For example, Kogan et al. (2015) collects tweets/retweets during Hurricane Sandy. They differentiate four kinds of social networks. First, by classifying users in affected areas or unaffected areas according to the tweet locations; second, by referring to tweets before, during, and after the Hurricane; third, by referring to the related data constructed by authors of the tweets/retweets, and by considering the authors of the tweets/retweets as nodes. Their research found that Twitter users in affected areas sent much more messages while under the hurricane's affect. These tweets form the tight and mutual related social networks during the Hurricane through tweeting. The data below comes from the cross-strait Twitter data during the 2012 presidential election. The analysis goal is to relate the data between Tweeters/Retweeters. Using graphic software from the network Gephi, the connotation this network graph shows is: although the three language groups (traditional Chinese, simplified Chinese, Japanese) focus on discussing the 2012 presidential election, users of the three language groups refer to different subjects. Showing that

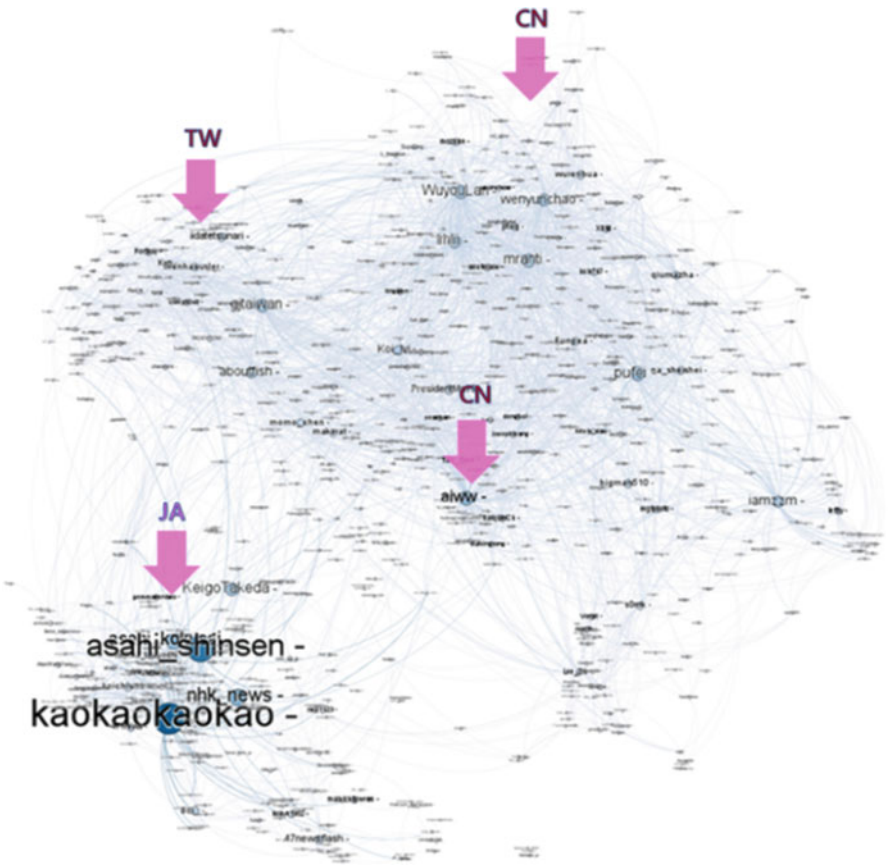



**Fig. 16.5** The network analysis of language groups during the 2012 presidential election (Cheng and Chen 2014)

during the 2012 presidential election, the phenomenon is that Twitter users of the three language groups gave opinions concerning the same issue (Fig. 16.5).

#### 16.4.4.5 Numerical/Type Analysis

Numerical/type analysis concerns the relationship between two or multiple data dimensions. This kind of analysis usually considers two groups of data dimensions as the variables, and cross analyzes them through statistical tools (such as SPSS or SAS). The data dimension can be either numerical data or type data. The researchers illustrate the relationship between data dimensions through the statistical results. For example, Bruns (2014) analyzes the relationship between the hashtags of buzzwords and reposted articles. For instance, to discuss the different states of related tweets,

in particular the interaction between different language groups, Burns and Burgess (2013) arrange user groups who speak Latin-based and non-Latin-based languages differently, and according to the quantity of their tweets, differentiate between the "active users," the "highly participant users," the "relatively inactive users" of the different systems of languages, by using millions of tweets which contain the hashtags #Egypt and #Libya, and by applying language recognition tools, concerning the protests people held in Egypt and Libya during "Arabic Spring." From observing the number of tweets with #egypt and #libya, this research finds that there are several differences concerning the number and transfer of the tweets. The most active 1% of Egyptian users usually use #egypt, particularly after the most common buzzword hashtag is transferred from #Jan25 to #egypt, and the Arabic users increased hence leading the whole discussion, even surpassing English users. This research compares the number of tweets between different language communities, and then discovers the differences in the event's transfer of messages between different language communities, according to the characteristics of different tweeted language systems.

#### 16.4.4.6 Text Analysis

Text analysis of traditional media is when researchers construct categories by considering human being as the coder and the page or single article as the sample for analysis (Chew and Eysenbach 2010). This kind of analysis is often called content analysis. The scale of social media data is huge to the degree that artificial analysis cannot handle it. Recently, data scientists analyze text through Natural Language Processing (NLP). Often referred to as "text analysis" or "computer-assisted text analysis," it first considers each word as the basic unit, and then undergoes the statistical analysis of word frequency or the relationship analysis of the vocabulary (Brooker et al. 2016). Chinese words are a little bit more complex than English words, because there is no interval among Chinese words. In order to do text analysis, the different types of words in one sentence must be cut, and the obsolete words must be reduced, by applying word segmentation technology (Chen and Cheng 2014; Cheng and Shih 2016). The below graphic example describes the algorithm of Chinese word segmentation which applies the surveillance pedagogy: the initial text needs to be cut and marked in the dictionary. First by noting the prefixes, mid-section suffixes, or single words, according to the position the single word has in the vocabulary; second, put into the language model for learning through the marked texts; third, provide for further analysis by outputting the text after the processing of word segmentation (Fig. 16.6).

Researchers can process vocabulary (such as word frequency, or the co-occurrence of vocabulary in certain paragraphs) through automatic software, after huge amounts of words in social media texts are segmented (Fig. 16.7).

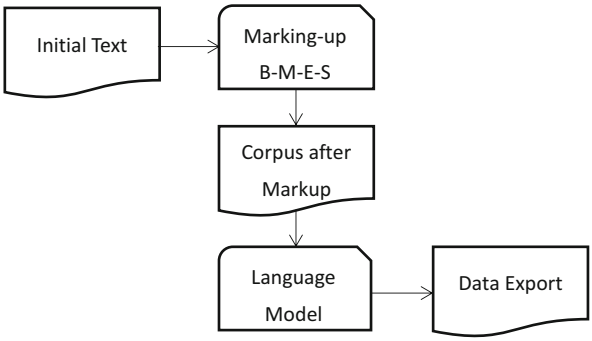



**Fig. 16.6** The process of Chinese word segmentation (the word database group, academic sinica)

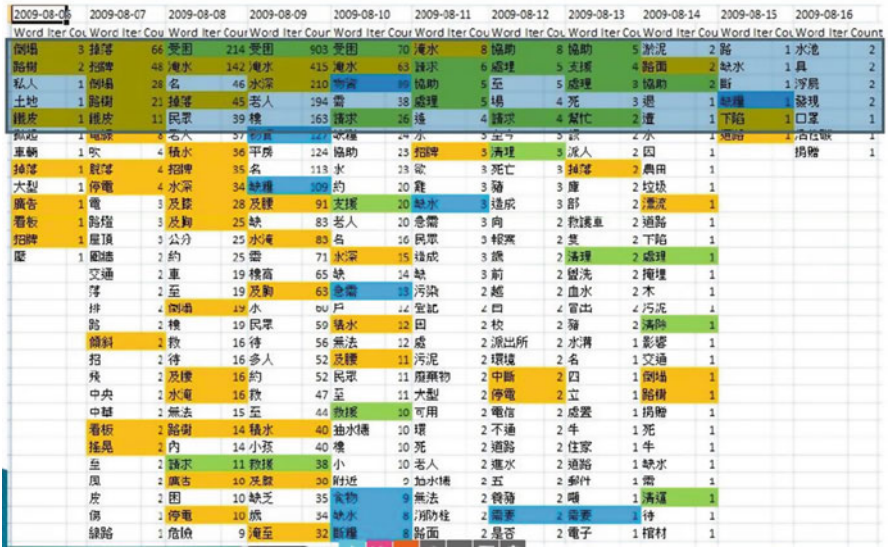



**Fig. 16.7** The distribution of the time sequence of word frequency on web sites during Typhoon Morakot in 2009 (Chen and Cheng 2014)

### 16.4.4.7 Sentiment Analysis

Sentiment analysis can be viewed as a particular kind of text analysis. In sentiment analysis, researchers will select the target vocabulary from the text, and compare the shared sentimental traits, to judge whether the text has a positive or negative sentiment (Stieglitz et al. 2014, p. 92). In artificial or fully automatic surveillance, the sentiment analysis must undergo word segmentation first, and then classify the positive/negative categories of the word according to the pre-constructed sentiment dictionary (i.e., the categorized vocabulary groups according to the sentimental properties), or by applying methods of machine learning. Through statistics and integration, sentiment analysis can predict the text's sentimental disposition. For example, American scholar O'Connor et al. (2010) analyzed the mutual referencing

between Tweets and keywords of consumer confidence by applying Twitter data analysis methods, and comparing that with the results of surveys done by survey companies. The author analyzed the sentimental dimension of the text by using software. The research found that there is a positive correlation between the word frequency of sentimental words on Twitter and the traditionally surveyed consumer confidence index. The correlation coefficient is up to 80% on specific issues. The research shows that the traditional public survey can be potentially substituted or supplemented by the sentiment analysis of the text. Thus, the author maintains that the social media data analysis can be used to measure public opinions and predict consumer confidence.

### *16.4.5 Data Visualization*

Data visualization is when researchers transform the data from numbers to graphic materials through software assistance, in order to display the distribution types or data trends. In social media data analysis, data visualization can be used in two situations: the first situation is when the visualization process is considered a tool for discovery. The second situation is when the visualization process is considered tool for narration, which usually takes place in the late stages of data analysis. The researchers will communicate with the audience through graphic display, and focus on the external manifestation through data visualization of the analysis. When the visualization process is used for narration, the researchers decide how the image is presented according to the data dimensions, analysis types, and graphic genre.

### *16.4.6 Using and Combining Types of Analysis*

The relationship between the research questions (goals) and the types of data analysis (means) is not a one-to-one correlation when discovering or solving a single question. Several different analytical tools may be applied for one research. The researchers will do the discovering analysis, for example, the relatively simple sequential or numerical analysis and describe the complete figure of event time and the discussion hotspots first, converting the key strings the users form into different times in accordance with the hotspots (in order to mine the text's contents or participants), and finally apply semantic web analysis or social network analysis of the users. For instance, in the research of Tweets in the 2012 Australian presidential election, Burgess and Bruns (2012) show the complete data first, followed by further numeric analysis of the Tweets' language groups second. Cheng and Chen (2014) use the same strategy to present the complete trend using numeric analysis first, then differentiate the groups through comparative numeric analysis according to the surfacing core issues second.

## 16.5 Data Evolution

Data evolution refers to the process relating problem to data processing. As we mentioned above, social media data analysis is both the process from questioning to answering and the process of data analysis and data cleanup. The two processes share data in common. The mined and produced data evolve between these two processes. We can illustrate how the two processes evolve data by referring to the graph below (Fig. 16.8).

### *16.5.1 Data Collection*

The social media data analysis usually originates from questions. In the initial stage of data analysis, researchers imagine how the data would look like, according to the posed research questions. For example, what kinds of platforms and what types of data do we need to find? What methods should we apply to mine the data? The research team gets into the collection stage of data processing while imagining it. It will set the key strings per the imagination, include the key strings, then mine the data from the social media platform through API-mining software (or set the crawling software), to obtain the raw materials.

### *16.5.2 Data Cleanup*

The raw materials of social media platforms may have many mojibakes, data lags, or misplaced formats. The researchers need to transform these raw materials into an intuitively usable data format, for instance, transforming the codes of the character (Big5 → UTF8) or transforming time zones of Greenwich to that of Taiwan. Researchers use many ways to structuralize the datasets and convert them

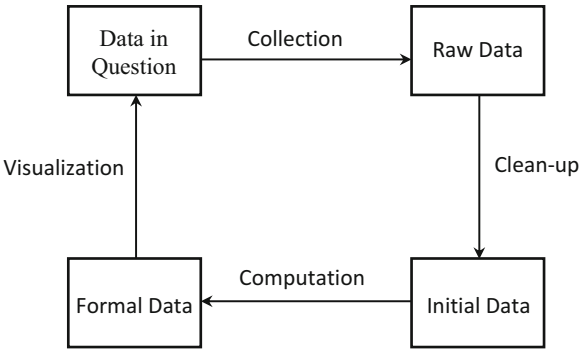



**Fig. 16.8** The process of data evolution

into initial data usable for discovery. At the same time, the researchers will use simple visualized graphs to undergo data discovery analysis (e.g., the researchers understand the life pattern of a product through collating time, posted articles, and quantity of users), or use descriptive statistics to quickly find the state of data currently presented, and to predict questions such if the data is complete? Is the data clean enough? Is there any noise? Was there a problem in the data collection process? Even the question "Is the dataset useful?" is worthy of being raised.

### *16.5.3 The Computation/Verification of Data*

The columns of the initial data still require researchers to select, integrate, or separate according to the data's dimensions. They number of variables still require filtering from researchers to be the execution target of computational calls. In the meantime, researchers will select and integrate desirable analysis types, and actually them according to the posed research questions. For example, researchers will transform the data of the reposted articles into a correlated dataset, and process the correlated data by applying the social network analysis method. Or the researchers will do one type of analysis, then do the other types of analysis from the results of the previous analysis. In other words, the analysis, as a means of discovery, is constantly occurring during the research process.

Researchers should judge if the analysis results answer the research questions after every kind of analysis has been applied. If the data can't answer the question, the researchers should evaluate and reflect upon this phenomenon in accordance with different standards, such as the research question, the data, and the results of the analysis. Researchers should also evaluate the relationship between the dataset and the posed questions, such as whether the construction of data dimensions can become its metrics, whether the construction of data dimension can be used for analysis, and whether these metrics can sufficiently answer the research questions? In other words, researchers compare the data's representation and the data's dimensions/metrics, then see if these two aspects mutually correspond.

Finally, researchers will present the data by means of visualization. Researchers select the tools and present the data comprehensively when he or she deems the data is sufficient for analysis, argumentation, and interpretation.

In the data analysis process, the discovery questions and the problem-solving process, as well as the process of data processing appear parallel. The processes are mutually related, not existing independently. Thus, researchers need to constantly care about the correlation between his or her own problematics and the data processing. This process of care may be difficult to describe in words. Therefore, we will use one data analysis case here to illustrate it.

## 16.6 Discussion

The contemporary social media data analysis is when researchers develop and evaluate every information tool and data structure, in order to collect, observe, analyze, note, or present social media data, in accordance with the requirements of the institution. The essence of social media data analysis is knowledge production/reproduction according to the data constructed by social media. This article aims to analyze the characteristics, elements, and process toward approaching social media data analysis. Based on bibliographical analysis and case studies, we have obtained several viewpoints for further research.

We have a few points to make before the end of the chapter.

### *16.6.1 Data Analysis as the Field of Activity*

This field of data analysis, which contains human behavioral data on the social media platform, is not only large in scale but also diverse in form and full of noises. Thus, social media data analysis is an activity where researchers come from different knowledge backgrounds (in particular the social and data sciences) and go through data analysis in mutual cooperation and communication.

Social media data analysis is a field of transdisciplinary knowledge in which researchers come from different intellectual backgrounds, and posess different domains of knowledge. Social scientists are good at transforming social reality into problems and interpreting messages and meanings. On the other hand, data scientists are good at processing big diverse data, and converting it into forms of visual communication.

Social media data analysis is located at an intersection between social science and information science. For communication scholars, the greatest challenge in social media data analysis is converting posed problems into data processing (Brooker et al. 2016, p.1). Thus, the key is transdisciplinary teams and communication.

### *16.6.2 Considering the Characteristics of the Data Thinking Process*

The core of social media data analysis is thinking using data. This process is a dynamic one. Data analysis is the process of problem-posing, problem-solving, and answer-seeking (Smith et al. 2014). The problems and data are continually evolving, and researchers will develop problem-solving strategies according to changes in situation, as the "Thinking in Action" Scribner (1986) described. Data analysis is not only related to the mental structure of the researchers, but also embodied

by connection between mind and body. Thus, data thinking can be viewed as the connected and collaborative process between mental state, body, and artificial products.

### *16.6.3 The Connection of Research Questions and Materiality of Social Media Data*

Social media researchers must understand the questions communication scholars pose, and frame those questions in the context of social media data. Researchers look for the connection between questions and data, and discover a problem-solving approach. As Gibson (1979) said, what social media data analysis looks for is the affordance between the posed questions (the subjective desire of the researchers) and the situation (the materiality in the situation). Researchers not only require research questions in mind but also need to understand the affordance between software tools and data. Data processing of social media data is usually handled by software tools. Every software tool has its own materiality (e.g., researchers will present the dimensions, sequence, and amounts of data by using rows and columns), in order to form the data's characterization. Consequently, researchers must use pros and cons of the material characteristics wisely, in order to find the best problem-solving strategy.

As for educators in social media data analysis, teaching/learning in a moving context is very important. The traditional pedagogy of research methods puts emphasis on method, rule, and universality. However, the essence of research activities focuses on tacit knowledge. It does not necessarily involve oral or textual explanation. The datasets are disparate. The important thing is in all cases to find the connection between the research questions and the data's materiality, and explore the best solution within the diverse problem-solving toolbox. Therefore the application of teaching strategies such as using cases as teaching materials, learning by doing, and using real-world cases to lead student thinking is necessary.

### *16.6.4 The Importance of Emphasizing Explorative Data Analysis*

Data analysis contains explorative and hypothetical processes. It contains the explorative connotation. Researchers use simple statistical methods and visualization tools reveal the different types. It contains the hypothetical connotation. Researchers tend to observe the data distribution types and the possible data singularities to find key questions in highly uncertain data areas.

The datasets must be explored, because every one of them is both stand-alone and similar to the other datasets occurring at the same time. In order to give researchers

an understanding of the data, and hence find the best solution to the problems, the functional scaffold allows researchers to represent different types of data before other researchers through simple descriptive statistics and visualized graphs.

However, traditional statistics (no matter if it is used for pedagogical purposes or research purposes) usually put more focus on verification process and model construction, rather than explorative analysis. Researchers need to observe the different data types of social media big data. Thus, the exploratory process is very important in social media data analysis. The question regarding the proportion that explorative data analysis has on future education is worthy of attention.

## 16.7 Concluding Remarks

Contemporary social media is like the "responsive system" of immediate response our contemporary society provides in the sense that it can generate very large scale, diverse, and highly dense data. On the other hand, data analysis is like the "reflective system" in the sense that researchers collect, filter, and analyze the data through meticulous and slow processes, hence representing social reality.

Social media developed at a very fast pace, while its data analysis is still in its initial stage. There is a huge gap between "fast thinking" and "slow thinking." The community of social media big data analysis is continually growing, while the reflective system of social media still lags.

Social media data analysis is an emerging field of knowledge. This field is still awaiting the inputs of different academic communities. This chapter is a preliminary analysis which attempts to draw a rough outline for social media data analysis. Owing to the format limits of this research chapter, we have selected the parts we wish to express. Complete topic coverage is impossible. We believe that the chapter would be a worthy reference for novices of social media data analysis. The goal of this chapter would be achieved if, in the future, the conversation and critique as regards social media data analysis can be advanced further.

## References

Boumans, J., & Trilling, D. (2016). Taking stock of toolkit: An overview of relevant automated content analysis approaches and techniques for digital journalism scholars. *Digital Journalism, 4*(1), 8–23.

boyd, D. M., & Crawford, K. (2012). Critical questions for big data. *Information Communications Society, 15*(5), 662–679.

boyd, D. M., & Ellison, N. (2007). Social network sites: Definition, history, and scholarship. *Journal of Computer-Mediated Communication, 13*, 210–230.

Brooker, P., Barnett, J., & Cribbin, T. (2016). Doing social media analytics. *Big Data & Society, 3*(2), 1–12.

Brügger, N., & Finnemann, N. (2013). The web and digital humanities: Theoretical and methodological concerns. *Journal of Broadcasting & Electronic Media, 57*(1), 66–80.

Bruns, A., Highfield, T., & Burgess, J. (2013). The Arab spring and social media audiences: English and Arabic Twitter users and their networks. *American Behavioral Scientist, 57*(7), 871–898.

Burgess, J., & Bruns, A. (2012). (Not) the Twitter election: The dynamics of the #ausvotes conversation in relation to the Australian media ecology. *Journalism Practice, 6*(3), 384–402.

Burgess, J., Bruns, A., & Hjorth, L. (2013). Emerging methods for digital media research: An introduction. *Journal of Broadcasting & Electronic Media, 57*(1), 1–3.

Chen, P. L., & Cheng, Y. C. (2014). From information to social convergence: Discovering emerging channels in major disasters. *Mass Communication Research, 21*, 89–125; (in Chinese).

Cheng, Y. C. (2014). The computational turn for new media studies: Opportunities and challenge. *Communication Research and Practice, 7*, 45–61; (in Chinese).

Cheng, Y. C., & Chen, P. L. (2014). Emerging communities in social media during the 2012 Taiwanese presidential election: A big-data analysis approach. *Mass Communication Research, 120*, 121–165; (in Chinese).

Cheng, Y. C., & Shih, S. F. (2016). News sources in social media during the 2012 presidential election in Taiwan. *Chinese Journal of Communication Research, 29*, 107–133; (in Chinese).

Chew, C., & Eysenbach, G. (2010). Pandemics in the age of Twitter: Content analysis of tweets during the 2009 H1N1 outbreak. *PLoS One, 5*(11). online14118.

Chiang, Y. & Lin, T. T. C. (2015). Big data in communication studies: A systematic review. In Peng, Y. (Ed.), The Proceedings of "2015 Big data, New Media & Users" Conference, Taoyun, Taiwan: Yuan-Tze University (pp. 355–368) (in Chinese).

Felt, M. (2016). Social media and the social sciences: How researchers employ big data analytics. *Big Data & Society, 3*(1), 1–15.

Forte, A., & Lampe, C. (2013). Defining, understanding, and supporting open collaboration: Lessons from the literature. *American Behavioral Scientist, 57*(5), 535–547.

Gibson, J. J. (1979). *The ecological approach to visual perception*. Boston: Houghton Mifflin.

Jungherr, A. (2014). The logic of political coverage on Twitter: Temporal dynamics and content. *Journal of Communication, 64*, 239–259.

Kahneman, D. (2012). *Thinking: Fast and slow*. New York: Penguin Books.

Kogan, M., Palen, L., & Anderson, K. M. (2015). Think local, retweet global: Retweeting by the geographically-vulnerable during Hurricane Sandy. Paper presented at the *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work*, Social Computing, Vancouver, BC, Canada.

Li, B. (2011). *The rising wind of opinion mining: On spatial and temporal structures in the diffusion of online hotspot events*. Beijing: People's Daily Press; (in Chinese).

Mahrt, M., & Scharkow, M. (2013). The value of big data in digital media research. *Journal of Broadcasting & Electronic Media, 57*(1), 20–33.

Manovich, L. (2012). Trending: The promises and the challenges of big social data. In M. K. Gold (Ed.), *Debates in the digital humanities* (pp. 460–475). Minneapolis, MN: The University of Minnesota Press.

Mayfield, R. (2006, April 27). *Power law of participation*. Ross Mayfield's Blog. Retrieved from http://ross.typepad.com/blog/2006/04/power_law_of_pa.html

O'Connor, B., Balasubramanyan, R., Routledge, B. R., & Smith, N. A. (2010). From tweets to polls: Linking text sentiment to public opinion time series. *ICWSM, 11*(122–129), 1–2.

Parks, M. R. (2014). Big data in communication research: Its contents and discontents. *Journal of Communication, 64*, 355–360.

Schönberger, V., & Cukier, K. (2013). *Big data: A revolution that will transform how we live, work, and think*. New York: Houghton Mifflin Harcourt.

Scribner, S. (1986). Thinking in action: some characteristics of practical thought. In R. Sternberg (Ed.), *Practical intelligence*. New York: Cambridge University Press.

Seltman, H. (2015). Exploratory data analysis. In *Experimental design and analysis* (pp. 61–98). Pittsburgh, PA: Carnegie Mellon University.

Smith, A., Molinaro, M., Lee, A., & Alberto, G. (2014). Thinking with data. *The Science Teacher, 81*(8), 58–63.

Stieglitz, S., & Dang-Xuan, L. (2012). Social media and political communication: A social media analytics framework. *Social Network Analysis and Mining, 3*(4), 1277–1291.

Stieglitz, S., Dang-Xuan, L., Bruns, A., & Neuberger, C. (2014). Social media analytics: An interdisciplinary approach and its implications for information systems. *Business & Information Systems Engineering, 6*(2), 89–96.

Tufekci, Z. (2014). Big questions for social media big data: Representativeness, validity and other methodological pitfalls. In *ICWSM '14: Proceedings of the 8th International AAAI Conference on Weblogs and Social Media*, Ann Arbor, Michigan, USA, June 1–4, 2014.

Tukey, J. W. (1977). Exploratory data analysis. Reading, PA: Addison-Wesley.

van Dijck, J. (2013). *The culture of connectivity: A critical history of social media*. New York: Oxford University Press.

Yang, L. W., & Shao, K. H. (2016). *Social big data: Listening & Analysis*. Taipei: Future Career Publishing; (in Chinese).

Zeng, D., Chen, H., Lusch, R., & Li, S. (2010). Social media and intelligence. *IEEE Intelligent Systems, 25*(6), 13–16.

# Chapter 17
# Big Data and Research Opportunities Using HRAF Databases


**Michael D. Fischer and Carol R. Ember**

## 17.1 Introduction

The database known as the "HRAF Files" in paper was for its time (in the 1930s and 1940s) a technological breakthrough giving scholars unparalleled access to a large quantity of textual and graphic information about the cultures of the world. This chapter briefly discusses the initial innovations, the enhancements with online searching (now eHRAF World Cultures and the newer eHRAF Archaeology), and future developments planned. As of the spring of 2018, the two eHRAF databases contain almost three million "paragraph" units from over 8000 documents describing over 400 societies and archaeological traditions.

Before the advent of computers, academics at Yale's Institute of Human Relations, convinced that scholars should study humans in all their variety not just those closest to home, were interested in producing data about cultures of the world that could be rapidly retrieved by scholars in many different disciplines. The basic data was primarily ethnographic in nature, that is, largely text information about cultural and social life based on participant observation and interviewing. The pre-computer organized information systems developed at the Institute were a technological breakthrough and provided a backdrop for the computerized (and now online) versions that supplanted it. It involved the following principles (Ember 2012): (1) use original text so that researchers could make their own evaluations;

M. D. Fischer
University of Kent, Canterbury, UK

Human Relations Area Files at Yale University, New Haven, CT, USA
e-mail: m.d.fischer@kent.ac.uk

C. R. Ember (✉)
Human Relations Area Files at Yale University, New Haven, CT, USA
e-mail: carol.ember@yale.edu



S.-H. Chen (ed.), *Big Data in Computational Social Science and Humanities*,
Computational Social Sciences, https://doi.org/10.1007/978-3-319-95465-3_17

(2) organize the materials by systematically classifying subjects as well as cultures; (3) use human intelligence to subject-classify at the paragraph level and sometimes sentence-level; (4) make the materials available in one place as a discrete collection; and (5) physically put materials on the same subject by all authors together for each culture.

The first order of business was to develop a topic classification system that could help scholars find similar types of information despite vast differences in custom and terminology used in different regions and cultures, including normalizing the many alternative names for most cultures. To take a simple example, all societies we know of have some kind of dwelling where families live, but ethnographers could use native terms (such as the Navajo term "hogan") or they could use alternative words like "hut," "house," "tent," "pit-house," etc. The HRAF staff decided to create a number of different subject categories pertaining to residences. One, "Dwelling," a subcategory of "Structures" describes residential structures with an emphasis on their physical attributes, such as mode of construction, shape and size, the durability or portability of the structures, or their seasonal uses. In contrast, the subject "Household" focuses on the social aspects of family units, such as typical and varying composition of households and whether household members live in one dwelling or a group of buildings within a compound. Other categories cover how buildings are constructed or what the interiors are like. Of interest is that the creators of the Outline of Cultural Materials report that they found it difficult to develop a system based on theoretical or preconceived categories; rather they noted that it was necessary to develop the system more inductively through trial and error, that is, after reading a variety of ethnographic materials seeing how anthropologists and other observers organized their materials (Murdock et al. 1950, p. xix). The result, the Outline of Cultural Materials (OCM), first published in 1938 (Murdock et al. 1938), was revised in print 12 times (6 editions and 6 editions with modifications—the latest print edition is Murdock et al. 2008). The OCM provides over 700 categories of controlled vocabulary to characterize subjects. As a shorthand, all subjects were given three-digit numbers, with the first two digits representing the more general category. (For example, Dwellings is 342, under the broader Structures category of 34*.) The OCM categories are not just used by HRAF; museums use it to classify their materials and individual ethnographers have used the subjects to classify their own field notes. Although controlled vocabularies are not unique, what is unique to HRAF is the fine level of subject-indexing to the search and retrieval elements (SREs—typically paragraphs). The other classification system, the Outline of World Cultures (OWC) provides a standardized list of the cultures of world; cultures were given alphanumeric identification numbers, generally reflecting the region and country location of the culture. (The first print edition appeared in 1954 (Murdock 1954) and the sixth edition in 1983). The HRAF staff concluded that some of Murdock's regions were problematic, particularly the grouping of Muslim cultures together as "Middle East," even though many were in sub-Saharan Africa. Therefore, in eHRAF World Cultures new broader terms for region and subregions were introduced that were based more on geography.

The actual process of producing the "files," originally called the Cross-Cultural Survey (later referred to colloquially as the "HRAF files," but we use the more appropriate name, HRAF Collection of Ethnography), was extremely labor-intensive because it predated not only computers but also copy machines. So each paragraph had to be retyped using onion-skin paper and carbon paper. If a paragraph was about more than one subject it had to be duplicated as many times as it had subjects because of the need to put all the same subject together. Because of duplication needs, the paper version of the files had about 4,000,000 pages of information.

The dilemma in any searching method, whether using an analog or digital means, is how to balance the breadth of coverage (i.e., the number of retrieved elements) with the efficiency of a smaller search result. Even with pinpointed OCM subjects, the number of results can be quite large. For example, the category "Dwellings" yields almost 23,000 paragraphs for 301 cultures in eHRAF World Cultures; "Households" has almost 40,000 paragraphs. A few strategies can help narrow the scale. For example, cross-cultural researchers generally test specific hypotheses on smaller samples of societies claimed to be representative and limit their reading for each society to a focal community in a particular time period (Ember and Ember 2009, pp. 76–78). Narrowing to a focal community and time period usually means that only those documents pertaining to the right foci are perused. HRAF has provided aids for this, such as marking the documents that match the foci for the Standard Cross-Cultural Sample, a commonly used cross-cultural sample (see http://hraf.yale.edu/resources/reference/sccs-cases-in-ehraf/). Specific hypotheses usually mean that a researcher can narrow the scope of a search to fewer paragraphs. So, for example, a researcher may be interested in the size of dwellings and how size varies with different aspects of social structure (Divale 1977; Ember 1973; Porčić 2010, 2012).

Another strategy (possible in the online eHRAF databases) is to narrow searches by combining subject categories or narrowing by adding keywords. For example, if you want to know about family household and dwellings and you believe they are likely to be described in the same paragraph you can ask for both categories in the same advanced search (this narrows the number of paragraphs to about 1900 paragraphs). Adding keywords to the search works well if there are only a few commonly used distinct words or phrases. If you are interested in the size of dwellings you can add the keywords "feet" or "meters" to narrow the Dwelling subject search. However, keyword searching is problematic when too many multiple terms describe the same construct, making it almost impossible to include all the appropriate words.

Also problematic is when a subject of interest does not fit neatly into one or two OCM subject categories. In a current project, we (Ember and colleagues) are trying to measure the "tightness" or "looseness" of cultures from ethnographic descriptions. This is a broad concept that involves assessing the degree to which there are strong and pervasive norms as well as expected punishment for norm violations (Gelfand et al. 2011). Many broad subject domains have to be examined (e.g., offencss and sanctions, norms, sexuality, marriage, gender, socialization) and

the volume of material is so large that researchers sometimes have to spend a week reading the material for each society before they can assign specific values to the various "tightness/looseness" measures (a process called "coding" in cross-cultural analysis).

Although the HRAF files greatly facilitate qualitative and quantitative comparative research, especially compared to the time it would take to collect all the books, articles, and manuscripts and then find relevant material, the quantity of data returned in search results is still often problematic. It is relatively easy to retrieve relevant text, and being able to collect together all relevant text from multiple sources greatly expanded the capability of researchers to do meaningful comparative cross-cultural research. However, HRAF is embarking on using "big data" methods to develop a range of postprocessing tools and methods for the returned text to expand researcher capacity once again, and thus make detailed cross-cultural research attractive to a wider range of researchers within and outside anthropology. Although the size of the HRAF collection is not extraordinary with respect to some "big data" datasets, just a few gigabytes, the structure is heterogeneous and complex and most of the relevant information must be extracted from ordinary document text.

## 17.2 Addressing Problems of Scale in New Ways

As indicated above, one of the main problems that scholars face is having too much material even when it is narrowed by OCMs and/or keywords. HRAF is currently developing a system where researchers can store a personalized set of preliminary results in one or more "notebooks," that can be returned to as often as needed. An initial search for a subject may be large, but the notebooks will have additional tools of selection besides deleting or adding paragraphs or adding keywords to refine a search. These include (1) searching within collected materials in personalized notebooks with topic maps and summarization services; (2) after identifying some critical paragraphs using computer algorithms to find "paragraphs like this" . . . ; (3) developing computerized auto-coding or interactive computer-assisted coding that might assist in developing post hoc subcategories (variables) with normalized values for analysis.

One approach to improving this situation that HRAF is experimenting with is leveraging the OCM classification to produce more nuanced topic-maps for the documents that can be used not only to expand capacity to search for information, but also to interpret the information within. The OCM was developed to support particular approaches to research, and has, indeed changed over the years to reflect changing research priorities. However, it is more pragmatic than theoretical in design, beyond the broad theoretical principle of comparison that stimulated the original collection of the data. The categories used in the classification represent the broad topics that anthropologists and others have found productive, based on the theoretical and practical aspects of the ethnographic literature and its applications. Topic-mapping is local to whatever specific collection of ethnography it is applied

to, and will vary depending on the corpus. A topic map of the entire collection will be different from topic maps of individual documents (or groups of documents), which will be different from topic maps derived from the results of a search. These differences can be leveraged to identify the gravity of particular topics at the different levels of mapping (terms gain "gravity" when these also appear in prior and/or subsequent search units).

Among other things, topic mapping improves results from searching for keywords, but also helps identify sections that are strongly correlated with a keyword. For example, in Fig. 17.1 a query for "oxen" is made in Paul Stirling's field notes (not currently in eHRAF, but smaller in scale and useful for developing services for dynamically identifying topics) in an application which returns results based on topic maps associated with the English term "oxen." This expands the results from an initial five instances to 65 instances, including several instances where Turkish terms for oxen are used. Of the 65 results the majority (52) were directly relevant to the search term, although the term did not actually appear in the text, and the remainder was of secondary relevance in the context of the whole search results in that they answered questions that arose from the other results relating to land usage, alternatives etc. The additional 60 notes are produced because there are topical relations in the oxen notes that can be satisfied by these additional notes. As the topic map in Fig. 17.1 shows, these contextualize the specific notes relating to oxen. Although primitive, this illustrates the potential applications of text mining for secondary research from archived resources.

Topic-mapping, combined with the returned text and publication metadata, can also assist in creating narrative or structured summaries of a set of search results, where different parts of the summary link back to the search results responsible for that part of the summary. This leads to the prospect of a "search by abstract,"


**Fig. 17.1** Topic search for Oxen in Paul Stirling's fieldnotes

where a single abstract is produced for each search, and subsequent searches can take place based on selected portions of an abstract which retains a record of the sources contributing to each portion. Related, but more easily achieved, topic maps in conjunction with OCMs can support a "search by example" approach, where selected entries are used as a basis for identifying similar entries from across the database.

Another development planned is the ability to do some data mining of the eHRAF corpus. We will develop topic extraction techniques ("text mining") interactively, that is, in conjunction with directed research questions. Text mining refers to the use of computational methods to extract significant terms and relations between terms from segments of text. Textual words are compared with those from a larger corpus of texts to derive measures of similarity. (The most common method is to use vector comparison methods, such as cosine similarity. These are tuned by transformations, stemming, and other techniques to find similar text segments by the closeness of the match.). This can be further improved through leveraging the OCMs, which make it possible to "mine" specific categories in the collection, and by using topic maps, either those local to each OCM, or collectively.

Why is interaction important? Although general ideas about the content in terms of contextualized "themes" can be "automatically" derived, these just provide an overview of themes and their relationship to other themes. Research inquiry usually requires more directed, contextualized searches. Our approach is to work from these more general topic maps to identify segments of text (paragraphs, sections, pages, user selections, etc.) that the researcher identifies as pertinent, to classify and transform these into a pattern that can be matched against other text segments for similarity. Effectively, we leverage the intelligence of the researcher to identify textual environments of interest; we can then identify similar segments without needing a high-level semantic interpretation. The researcher can also refine the search by pointing to subsets of passages that are most relevant. False positives are possible, but the aim is to reduce the matches to a manageable number. Results may be improved by the use of stemming algorithms to reduce words to their core part—for instance, "food," "foods," and "foodstuffs" become "food."

We have applied this approach successfully to text mining academic articles from the *American Anthropologist* (AA) for 1950–2000. Figure 17.2 shows examples of topical domains from the AA produced in a prior EPSRC/ESRC project, Genealogical Relations of Knowledge (GROK). However, articles are highly contextualized around a topic, have fairly clear sections that can be identified, and progress thematically. As we discovered in recent exploratory work, ethnographic monographs, although formally having general sections, cover many topics repeatedly throughout the text without distinct sections or progressive development, which is even more characteristic of fieldnotes. Fortunately, we have developed an algorithm for segmenting ethnographic texts that appears to overcome some of these problems by looking for "runs" of basic relevant patterns in paragraphs within the text and then applying the previous algorithms to these newly derived segments.

With a text-mining resource we could ask questions such as: How has topical interest changed over time? Do different types of authors describe different things?

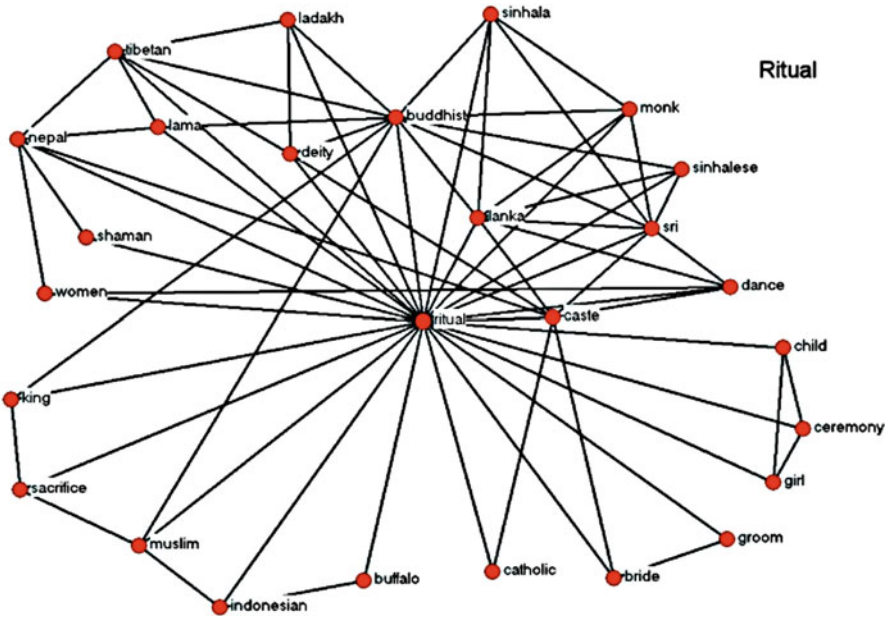



**Fig. 17.2** Concept network for ritual in a sample of American Anthropologist articles 1940–1960. Nodes can be clicked to reveal source segments (Zeitlyn et al. 2008)

Do subjects described vary by region of the world? If we added additional metadata to the text search and retrieval elements, such as the gender of authors, their nationality, their research training, then we could ask: Do male and female ethnographers describe different topics? How does nationality of ethnographer affect areas described? Does research training matter?

## 17.3 Moving Forward on on Reuse of Older and New Ethnographic Data

Databases like eHRAF World Cultures are aimed at facilitating reuse of largely previously published ethnographic information about the cultures of the world, but HRAF, which relies primarily on institutional memberships, has only been able to produce a subset of the ethnographic information potentially available. A major limitation is the costliness of human subject-indexing to the paragraph-level. There is much more information out there—published, unpublished, and on-going—that could be studied. Frankly, the reuse value of ethnographic data is high but hard to achieve. There are efforts, which we will describe shortly, that we will try to undertake at HRAF to be able to more efficiently process materials and maximize reuse of existing material, but we also need to enable others to process their own

ethnographic materials (see "Towards a Services Platform for Broader Reuse") if we are ever going to be able to scale-up.

Before describing those efforts, we will discuss some general issues regarding the goal of putting disparate ethnographic materials together. The main issue is arriving at interoperability across varied ethnographic sources and other kinds of data, quantitative and qualitative. Ethnography, as a form of reporting physical and cultural data relating to societies, was founded around strong comparative principles and data collected across populations and/or societies. Thus, ethnography was always intended to have considerable value for reuse purposes as scholars and policy makers ask broad or narrow questions about similarities, differences, and changes in human populations. But although the goal was broadly comparative, the heterogeneity of data across time and within and across the various cultural branches of knowledge is considerable and interoperability presents considerable challenges. Understanding the range of data and the logical possibilities for interaction is critical and yet is under-addressed in the literature.

Broadly speaking, there are three commonly understood types of interoperability: (1) structure and/or format conventions—the form in which data is represented (syntactic); (2) meaning—what does the data represent and why and how does it represent it (semantic); and (3) what we can do with a data set—the models and/or interpretations a data set supports and how these are constructed (pragmatic). After explaining these types of interoperability, we outline some steps HRAF is taking to increase interoperability.

Syntactic interoperability requires that we can identify individual records, groups of records (if any) and their relationship, and individual data items and the encoding of data items. Any metadata associated with individual data items or groups of items would be included. The interpretation of items or records is not an aspect of the syntactic specification. Specific algorithms to do many of these syntactic conversions are becoming widespread. However, much irreplaceable legacy research data is buried in unsupported, and often one-off, file formats and data layouts. Development of more general tools and services that "open" these for future use will make valuable and often irreplaceable data accessible, and help ensure future pathways to automate "rolling over" data from format to format as digital infrastructure inevitably changes. Likewise media dependencies, for example, legacy data held on media no longer well supported (e.g., paper, card decks, tape of any kind, floppy disks, older hard disks, some CD formats) are a matter of urgency and must be transcoded in the very near future.

Much of the data that anthropologists collect is not suited for representation in simple flat files; indications of complex contexts are as much a part of a dataset with respect to description and analysis as the data items themselves, and complex structures and relationships cannot be easily represented as a simple flat "row and column" type database. So data organization, and related metadata, is often multidimensional, taking forms that can be represented in trees, graphs or other relational abstractions. Additionally, metadata is used to relate how higher order classifications, inferences or transient references might be represented and

processed. Semantic services are usually built on new layer of metadata applied to a "flatter" syntactic metadata layer.

Semantic interoperability for datasets is not as straightforward as syntactic interoperability. Whereas syntactic conventions allow us to differentiate data items and recover metadata relating to these, semantic operations are required to identify similar data items between datasets. One traditional method uses a codebook describing each variable and its possible instantiations in the data set. Similarly, the most common means of supporting semantic interoperability between digital datasets is through associating metadata with each data item and value. Metadata is usually (but not always, as for XML) maintained apart from the data itself, serving as a kind of template for a class of data sources, rather than a descriptor for a single dataset or individual data items. Most people are familiar with simple metadata, such as that for a publication in bibliographic record, where slots are specifically designated for particular roles or states; author, title etc. Metadata includes syntactic information regarding how data will be organized in addition to providing a label for a slot. The goal for semantic interoperability is to relate the items between datasets to link similar items. For interoperability, information must also be available relating the possible values for data and how these relate to each other. For example, for the simple variable "age" one dataset might relate this in years, another years and months, and a third corresponding to the set {child, adolescent, adult, elderly}. With conversion, the latter will clearly not be equivalent to the former, just comparable.

In the case of ethnographic data, such as that processed by HRAF, while there are often tables of synthesized aggregate data, the majority of the data is natural language text (the majority in English) organized into conventional structures associated with published material. Syntactically, this makes it fairly easy to put a range of material from different datasets (individual documents) into a common form, particularly with sufficient metadata relating to elements of publications, and a relatively robust XML schema that can represent a wide range of publication formats.

Although some of the publication metadata has a semantic definition, relating time, place, and society, the substantial semantic metadata HRAF has added are the Outline of Cultural Materials (OCM) subject classifications applied by anthropologically trained HRAF analysts. These allow, for example in searching for information about a particular topic such as how marriages are arranged using the OCM code "Arranging a Marriage" in an Advanced Search which ensures that the text of a given search unit contains information relating to the topic. However, in the current version of eHRAF, the researcher has to make his or her own decisions about the forms of the arrangements. Are arranged marriages customary? If so, how are they arranged? Or do individuals decide on whom to marry? If so, are there customary patterns of courtship? In other words, after finding the passages with information, researchers are then on their own.

Pragmatic interoperability recognizes that for standards to be adopted by the anthropological community these must have clear benefits, have useful levels of partial adoption, and be researcher-extensible and open to application for legacy data sets. Identifying the pragmatic requirements for data reuse in interdisciplinary

research goes beyond simply matching up elements syntactically, structurally and semantically. For example, relating the interaction of ethnographic research with genetic research is not a simple matter of identifying and structuring data, but requires some form of common reference or context, such as applying to a common population or site.

### *17.3.1 Steps HRAF Is Taking to Improve Interoperability*

In HRAF development we are using several approaches to promote syntactic interoperability. The most important change dates to the first online version of the database when SGML was used to mark up the text, structure of the text, and associated metadata, which was then converted to XML in the mid-2000s and published in XML in 2008. XML provides the following advantages: (1) it is far easier to perform a range of transformations of the document text and context; (2) it makes a wider range of queries much easier to perform, including queries that reference the context of matching entries; and (3) it promotes transformation into more specialized outputs that can be further processed for further analysis and customized reporting. Even though each document has a unique structure and content, these are syntactically encoded using the same XML schema, and makes interoperability from these heterogeneous sources at least possible.

To increase semantic interoperability in the future, we plan to base further services on searching a transformation of the present HRAF Production XML schema into what we call the HRAF vDoc XML schema (vDoc stands for "virtual document"). In contrast to the Production schema, which tries to reproduce the structure of the original documents, the vDoc schema reduces some of the heterogeneity between works by standardising relevant document structure and context, making relevant metadata available at the level of an SRE or search and retrieval unit (usually a paragraph) instead of having to retrieve it from the document context. We also plan to add multiple versions of the text in each record including the original markup, a text only version, a record of the parts of speech for the plain text content. We will also add statistics relating to the text, ranging from simple frequencies within the search unit, and an indication of which terms have "gravity" (also appear in prior and/or subsequent search units). There will be other additions to support semantic interoperability. The vDoc XML structure will provide a uniform way to represent both the base data and references to the base data. Within the HRAF services framework each service will produce a vDoc as output, excepting a few whose purpose is to render vDocs into forms for display or interchange with other tools and platforms. However, most services transform or aggregate other vDocs into a new composite vDoc. vDocs are flexible enough to embed most other data formats, so can serve as an all-purpose media for compositing different data streams and types.

The use of topic-mapping in conjunction with OCM classifiers will be at the intersection of semantic and pragmatic interoperability as it will leverage the OCM

and the decisions analysts make with respect to the OCM in conjunction with topics emerging from the texts, hopefully reflecting the ethnographers' intentions. Pragmatic interoperability supports the researcher's capacity to answer specific questions while drawing across data from different sources. The OCMs alone are a powerful tool in this respect, and the basis for the degree of success the HRAF databases enjoy. But the extent of manual labor required to utilize the material once located is quite onerous. In its most basic use topic-mapping will help reduce effort by helping the researcher to reduce the set of results examined, since topic-mapping provides much more detail regarding the contents than the OCM classification.

Pragmatic interoperability is primarily about integrating across platforms and disparate models. We will turn to this next in the context of a plan for a service platform to incorporate new ethnographic data from others.

### *17.3.2 Towards a Service Platform for Broader Reuse*

Beyond the present HRAF resources, we are working towards expanding access to and reuse of other researchers' underlying and published ethnographic and other data, without compromising confidentiality or other constraints, to promote reuse of data generally in a new services platform. There are considerable problems with publishing most of the material that an ethnographer collects during fieldwork in a given society. Most of the information is highly personal and often sensitive, and agreements for use often involve per person and occasion agreements,. Even then, the ethnographer has a duty of care that goes beyond the agreement of the people involved. The topics researched often contain highly confidential material that has a high potential for personal, political, or even legal damage. Because the record is cumulative and detailed, even redaction and anonymizing is inadequate to protect interests, as it is often quite simple to work out the identity of specific individuals.

One partial solution to the confidentiality problem is to search within underlying text, other than for embargoed terms (such as personal names and place names) and topics (such as potentially subversive activities) designated by the contributing researcher, but to not return the matching underlying text. Rather, a range of transformations of the results are made available to be reported (as permitted by the contributor). These include basic information, such as word frequencies for the results or subunits of the results and also the embedding document context, but also topic maps for the results, and various approaches to narrative or more structured summaries of the content. Further leveraging the HRAF collection, examples from HRAF similar to those for the undisclosed text results would be available so that researchers can see the range of material across a designated range of HRAF cultures that corresponds to the themes in the confidential material. In addition to their exemplary value, these will assist in interactive auto-coding of the unseen material.

Drawing on correlations between HRAF derived text which includes OCM classifiers, and texts submitted by other researchers which do not, a graph can be

created with will facilitate assignment of OCM codes to the external texts, either automatically, or with some assistance by the depositing researcher. A services platform for detailed topical analysis of underlying texts together with outer textual metrics will be configurable to different topical graphs and, potentially, ontologies other than the OCM. Tools will be developed to assist researchers in situating their own results with respect to others for comparative purposes using material submitted by other researchers in addition to the HRAF materials. These services will operate on sources directly submitted by researchers to the repository and published material, for which underlying text will not be quoted, and additionally the expertly analyzed, classified and curated ethnographic data corresponding to a number of accepted samples (the HRAF Collection of Ethnography) using standardized topical tags from the Outline of Cultural Materials.

Some of the material we will want to integrate with the HRAF corpus will have been developed with other schemas and models. To achieve pragmatic interoperability, we will be exploring a method analogous to the notion of "docking" originally proposed by Axtell et al. (1996) for agent based modelling. Docking is a method of establishing connections between apparently disparate models which have some matching or related classes, variables or parameters. Collaborating researchers find common ground where their respective descriptive understanding of their datasets (perhaps as a model) are sensitive directly or indirectly to each other, collectively building a new layer that "glues" the two together. However detailed knowledge of the others' model is not necessary for a given researcher, only the agreed model that connects the two. Pragmatically, a "docking" approach helps researchers focus mainly on their own research problems but allows them to further explore the impact of the larger context on their research problem. For example, by relating marriage, the distribution of a relatively rare allele, and mortality the social anthropologist, biological anthropologist, and demographer each gain insight into how context refines more general understanding.

## 17.4 Conclusion: Expanding HRAF Research Services

Currently the eHRAF application available to the membership has a fixed set of options for search and reporting, fairly typical of current generation Web applications. But we are repurposing search and retrieval operations as a variety of services that are independent of any specific Web application. Our present HRAF XML schema is oriented towards reproducing the original appearance of a publication, and has a very complex structure due to the heterogeneity of the 8000 or so sources we use, spanning over 100 years of ethnographic evolution. We will retain this structure for production and archival purposes, but to facilitate large scale search and retrieval services we are normalising the structure to focus more on associating key metadata and precompiled statistics with each paragraph so that it can be more easily evaluated within a given search, and a broader range of search criteria used.

Our principal goal is to develop new tools and infrastructure to support primary and secondary ethnographic research using the data resources available at HRAF.

The new services will leverage attributes that identify pertinent metadata such as culture, region, time of description, time of publication, type of author (e.g., ethnologist, geographer, missionary), in addition to the present text and associated analyst-supplied OCM subjects. We are working on auto-classification capabilities based on the HRAF collection that will enhance conventional topic extraction (ontological classification), tools to support coding materials for value (epistemological assignment) and work towards auto-coding techniques to promote broad consideration of comparative analysis and situation of human practices and behaviors for basic and applied research. To support data mining we are developing services with which we are experimenting with different approaches to producing topic maps suitable for paragraphs in context and with auto-classifying paragraphs and larger sections with OCM categories in a manner consistent with our professional analysts. We are also exploring methods to apply similar auto-classification to other sources, ranging from academic publications in anthropology to newspaper articles. Finally, beyond the present HRAF resources, we are working towards expanding access to and reuse of other researchers' underlying and published ethnographic and other data, without compromising confidentiality or other constraints, to promote reuse of data generally in a new services platform that will enable many researchers to add their own materials to enhance the ethnographic corpus and promote reuse of ethnographic data within the bounds of well-established and well-founded ethical constraints.

## References

Axtell, R., Axelrod, R., Epstein, J. M., & Cohen, M. D. (1996). Aligning simulation models: A case study and results. *Computational & Mathematical Organization Theory, 1*(2), 123–141.

Divale, W. T. (1977). Living floor area and marital residence: A replication. *Behavior Science Research, 26*(2), 109–115.

Ember, C. R. (2012). Human Relations Area Files. In *Leadership in science and technology: A reference handbook* (Vol. 2. William Sims Bainbridge (Ed.), pp. 619–627). Los Angeles: Sage.

Ember, C. R., & Ember, M. (2009). *Cross-cultural research methods* (2nd ed.). Lanham: AltaMira.

Ember, M. (1973). An archaeological indicator of matrilocal versus patrilocal residence. *American Antiquity, 38*(2), 177–182.

Gelfand, M. J., Raver, J. L., Nishii, L., Leslie, L. M., Lun, J., Lim, B. C., et al. (2011). Differences between tight and loose cultures: A 33-nation study. *Science, 332*(6033), 1100–1104.

Murdock, G. P. (1954). *Outlines of world cultures*. New Haven: Human Relations Area Files.

Murdock, G. P., Ford, C. S., & Hudson, A. E., Kennedy, R., Simmons, L. W., & Whiting J. W. M. (1938). *Outline of cultural materials*. New Haven, CT: Institute of Human Relations, Yale University.

Murdock, G. P., Ford, C. S., Hudson, A. E., Kennedy, R., Simmons, L. W., & Whiting, J. W. M. (1950). *Outline of cultural materials* (3rd. revised ed.). New Haven: Human Relations Area Files.

Murdock, G. P., Ford, C. S., Hudson, A. E., Kennedy, R., Simmons, L. W., & Whiting, J. W. M. (2008). *Outline of cultural materials* (6th with modifications ed.). New Haven: Human Relations Area Files.

Porčić, M. (2010). House floor area as a correlate of marital residence pattern: A logistic regression approach. *Cross-Cultural Research, 44*(4), 405–424.

Porčić, M. (2012). Effects of residential mobility on the ratio of average house floor area to average household size: Implications for demographic reconstructions in archaeology. *Cross-Cultural Research, 46*(1), 72–86.

Zeitlyn, D., Bagg, J., Ryan, N., & Fischer, M. (2008). Making sense of anthropological synonyms. *Anthropology Newsletter, 49*(1), 34–35. https://doi.org/10.1525/an.2008.49.1.34.

# Chapter 18
# Computational History: From Big Data to Big Simulations


**Andrea Nanetti and Siew Ann Cheong**

## 18.1 Introduction. The Vision for Computational History

Do historians need computational history to better understand the actual history? Sir Arthur Stanley Eddington (1882–1944), in his 1927 *Gifford Lectures* said that "the contemplation in natural science of a wider domain than the actual leads to a far better understanding of the actual" (Eddington 1929, pp. 266–267). Before the advent of computational technologies, the value of thought experiments, of which Albert Einstein was very fond, was to present scenarios different from the ones humans observe. The physical scientist would then follow the scenarios through their logical ends to identify what we might have missed and realize what else could be possible if we had lived in a different universe, and ultimately understanding the physical laws that we have at a much deeper level (Eddington 1929).

We believe this can be equally true for simulations in historical sciences. The historical accounts work on "what happened" (i.e., the factual), while computer simulations tell us "what could have happened" (i.e., the counterfactual). Only by combining both the most accurate assessment of what actually happened and what could have happened, we can address the question if in history there are such things as universal laws, from which we cannot deviate in a cause and effect "mechanism-based understanding" (Paolucci and Picascia 2011, p. 135) of historical phenomena. The power of computer simulations can support historical sciences

A. Nanetti (✉)
School of Art, Design and Media, Nanyang Technological University, Singapore, Republic of Singapore
e-mail: andrea.nanetti@ntu.edu.sg

S. A. Cheong
School of Physical and Mathematical Sciences, Nanyang Technological University, Singapore, Republic of Singapore
e-mail: cheongsa@ntu.edu.sg



S.-H. Chen (ed.), *Big Data in Computational Social Science and Humanities*, Computational Social Sciences, https://doi.org/10.1007/978-3-319-95465-3_18

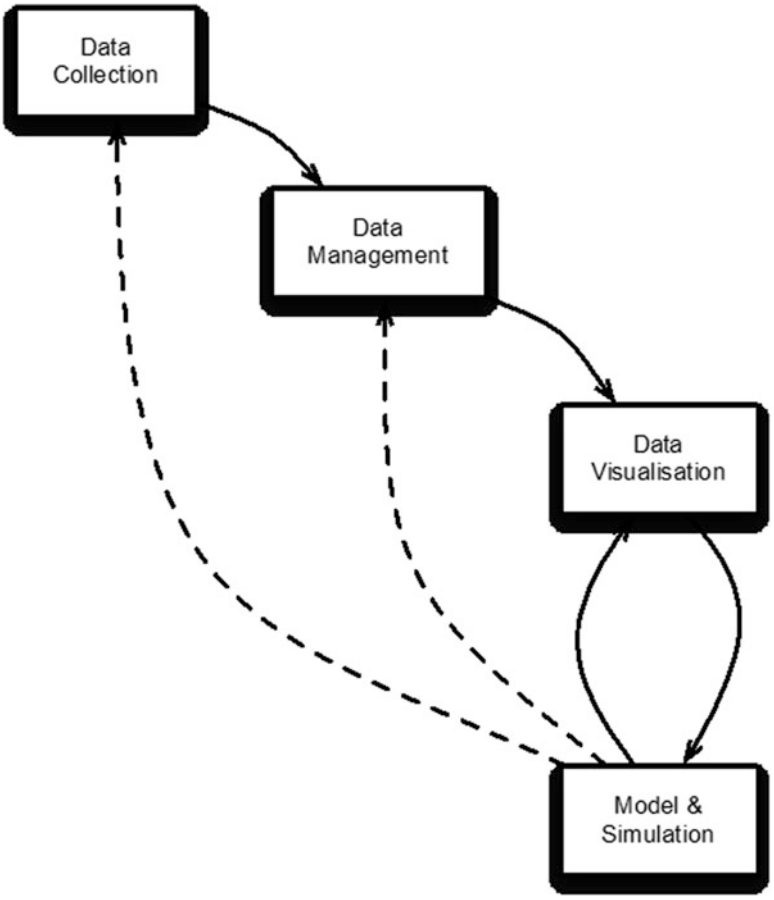



**Fig. 18.1** The Stages a discipline must progress through to become computational

to develop a shared prescriptive mode of inquiry in the assessment of primary and secondary sources. It will also provide new freedom in the historian's subjective and descriptive identification and assessment of problems to be investigated. Figure 18.1 illustrates the stages, through which history can improve as a computational discipline.

In general, it is known that to improve this kind of advancement of learning, historians need to develop specific ontologies to parse data and recognize entities from historical sources. These data can then be mapped into an electronic database and used in analytical environments to build linkages between parsed texts and recognized entities from other heterogeneous sources (e.g., Wikipedia, Open Street Map, etc.) and search engines (e.g., Google Scholar, Microsoft Academic, etc.). For this to happen, historical data also need to be published in online and open access databases, so that they can be properly shared. Historians, as a collective whole, have

big digital data, organized in databases but they are not very useful because most of them sit with some kind of organization on the hard disk of individual researchers.

Scholars partially share their data via published books and journal papers, in which data are manipulated in descriptive narratives and need a reverse-engineering process to be used again for a different kind of thinking. If citations and notes are the "procedures intended to communicate an effect of authenticity" (Ginzburg 2012a, p. 21), since Modern times historians normally use the footnote as "the one form of proof supplied in support of their assertions" (Grafton 1994, 1995, 1997). However, over time these footnotes can become an unwieldy web that takes considerable effort to navigate. Superhuman efforts are thus required to take all the pieces, and put them together into a recognizable whole.

Therefore, not only the interface to the databases must be properly designed so that it is user friendly, but also and most importantly the data must be curated and tagged by experts using the same identified ontologies and vocabularies, in order to aggregate the data, for example, into a graph database and make it publicly accessible to the international scholarly community, so that any researcher who needs a particular piece of data can find it easily and quickly (e.g., on MSRA Graph Engine, Linx Analytics, etc.). The same identified ontologies and vocabularies can be used to model historical data from historical sources as Linked Data (i.e., best practices to export, sharing and connecting pieces of data in the Semantic Web) and generate, for example, graph representations of the data (e.g., RDF using JSONLD-JavaScript Object Notation for Linked Data), among other solutions (Grinin and Korotayev 2010; Graham et al. 2016).

Unfortunately, nearly all historical databases were designed to be the end products of research projects or programs. To further proceed, databases need to be constantly expanded with the addition of new data sets. Among others, examples of such excellent historical databases include the *Digital Atlas of Roman and Medieval Civilizations* directed by Michael McCormick, the *Seshat: Global History Databank* initiated by Peter Turchin, the *Big History Project* conceived by David Christian, *Trismegistos* founded by Mark Depauw, and *Pelagios* coordinated by Leif Isaksen, and the *Collaborative for Historical Information & Analysis* (CHIA) for creating a world historical dataset initiated by Patrick Manning with support from the US National Science Foundation (Manning 2013, Manning 2015). We believe that these databases, as well as others, can become portals of historical knowledge, if they also offer functionalities to combine data with metadata, show visualizations of this combination, and run simulations based on insights gained from such visualizations.

Beyond the mandatory identifying metadata associated with each piece of historical data, databases should also record the interactions between researchers from different disciplines and the data, in the form of metadata. Clearly, these forms of interactions between experts could not happen easily without the computer database, because most of the expert assessments are pre-publication level and conjectural, so we will not see them in journal publications or books, however long we wait. In this sense, having very diverse data made available on a database, and having metadata to augment the data sets themselves is one way the digital computer is revolutionizing the study of history, by allowing historians more

intimate interactions with the data, and consequently closer interactions amongst each other.

However, if we stop at this stage, then data sets and metadata will accumulate, and very quickly the volume of data and metadata available will be so large that no one expert can comprehend them anymore. Therefore, to take advantage of the third wave of 'really' computational history's opportunities, historians can be helped by the computer to better comprehend the collection of data and metadata, i.e., to go from simply data management aided by the computer (Graham et al. 2016, pp. 73–111) and to more sophisticated topic modelling and data visualisations ("deforming, compressing, or otherwise manipulating data in order to see them in new and enlightening ways", Graham et al. 2016, pp. 113–158; pp. 159–194), and network analysis (Hitzbleck and Hübner 2014, pp. 7–15; Graham et al. 2016, pp. 195–264).

In this Data Visualization stage, the historian will borrow various machine learning strategies from the computer scientist to discover patterns in the data. Because historians traditionally spend long hours working directly with data, they become very good at formulating hypotheses, and thereafter finding from memory other pieces of data that would support such hypotheses. However, it is highly likely that they miss many other patterns in the data that do not fit into their modes of theorizing. The suite of data visualization and machine learning methods developed by computer scientists over the years can help discover most of these patterns. We feel such methods have been under-utilized because (1) the historical databases are fragmented, and therefore, patterns across different data sets cannot be detected, and (2) the methods are not traditionally included in the training of historians. More importantly, the historical databases are designed for human query, and not necessarily structured for machine query and thus machine learning.

The final stage that history must reach to become a full-fledged computational discipline, is Modelling and Simulation to explore big historical data in big simulations, algorithmically—as John Holland would say (Holland 1975; Mitchell 1996, pp. 2–3). Models can be top-down (equation-based) or bottom-up (rule-based), and can be analysed (by following the chain of logic in the equations or rules until we arrive at conclusions) or simulated (by letting the computer follow the chain of logic, so that we can interpret the conclusions). Models help us understand the big picture, by functioning (in conjunction with analysis, and/or more likely, simulation, when the model becomes too complex) as a *macroscope* that synthesizes our fragmentary knowledge and insights into a complete whole.

As summarised by Shawn Graham, Ian Milligan, and Scott Weingart, the term *macroscope* was first used by de Rosnay (1979) to discuss complex societies. In literary criticism, a similar concept was called 'distant reading' by Moretti (2005) and 'macroanalysis' by Jockers (2013). As for cultural history, an exemplar demonstration of "data-driven macroscopic" approach is given by Maximilian Schich and his research team (Schich et al. 2014, p. 562). Murray Gell-Mann pointed out in his keynote lecture *A Crude Look at the Whole: A Reflection on Complexity* given at the homonymous international conference hosted by Nanyang Technological University Singapore from 4 to 6 March 2013, to increase the

understanding of historical processes, we should improve the approach pioneered by the British historian Toynbee, rather than simply criticizing and marginalizing. In his twelve-volume magnum opus *A Study of History*, Toynbee presented the development of major world civilizations starting from a history of the Byzantine Empire (Toynbee 1934–1961; Gell-Mann 1997, p. 9; Schäfer 2001, p. 301). Others, like Aiden and Michel (2013) also wrote about "a [macro]scope to study human history" (Graham et al. 2016, p. 2).

In Sect. 18.3 we will explain the limitations of equation-based modelling, however powerful, when applied to historical inquiry, and why it is more natural and appropriate to adopt agent-based modelling (ABM). We will explain how we would go about developing agent-based models, and how we can use their simulations to add to our understanding of history. In spite of it being critical to *macroscope* approaches, ABM, as a computational practice, remains largely unfamiliar to digital historians, despite signs of increasing interest (Gavin 2014). In the historical landscape, ABM, like in other disciplines, would explain general trends and offer a complementary, but very different path to macroscopic knowledge. Joe Guildi and David Armitage in their *Historical Manifesto* (2014) argued the importance of macroscopic thinking. Shawn Graham, Ian Milligan, and Scott Weingart gave to their monograph on *Exploring Big Data* (Graham et al. 2016) the subtitle *The Historian's Macroscope*.

Ultimately, the purpose of having models is to do predictions, and these can be qualitative or quantitative. If we re-simulate the past, we can end up with a simulated world (Gavin 2014, p. 24), interwoven by counterfactual histories. If we simulate into the future, we will be exploring different scenarios. Counterfactualism and the debate over contingency versus inevitability have been explicit themes in modern evolutionary biology since Stephen Jay Gould's book about evolution and how to interpret evidence from the actual past (Gould 1989). The discussion became relevant for history of science, in general (Radick 2005), and Osvaldo Pessoa Jr has been exploring the role for computer models in assessing history of science counterfactuals (Pessoa 2001). This discussion fits in the discourse of "The Social Logic of the 'Text'", as discussed in 1997 by Gabrielle Spiegel, who argued that "while cultural anthropology and cultural history (together with the New Historicism ...) have successfully reintroduced a (new) historicist consideration of discourse as the product of identifiable cultural and historical formations, they have not been equally successful in restoring history as an active agent in the social construction of meaning" (Spiegel 1997, p. 9). But, before we explain how simulated histories can help historians, let us first link simulations to the historian's key problematics.

### *18.1.1 History's Chase for Truth*

According to the New Oxford American Dictionary, the Greek word ἱστορία/*historia* comes from *histōr*, which means 'the one who saw, the testimony > learned, wise

man', and comes from an Indo-European root shared by wit/vit (to know) that gave Sanskrit *veda* 'wisdom' and Latin *videre* 'see', as well as the Old English *witan* of Germanic origin, and is related to Dutch *weten* and German *wissen* (Joseph and Janda 2003, p. 163). Thus, history is a kind of knowledge acquired by investigation with the intent to generate wisdom, and implies the action of 'inquiring/examining', which is a requirement to move from knowledge (knowing how to do something) to wisdom (knowing under which situations to act).

If one agrees with Aristotle (Poetics, 51b), the historians speak of that which exists (of truth), the poets of that which could exist (the possible). In a computational modelling perspective, Michael Gavin (2014) notes that "on the surface, computational modelling has many of the trappings of science, but their core simulations seem like elaborate fictions: the epistemological opposite of science or history". He proposes "that these forms of intellectual inquiry can productively coincide" (2014, p. 1). But, it is not as simple as that. Let's give a few significant examples. Roland (1967)—following the structural linguistics of Ferdinand de Saussure and its anthropological extension made by Claude Lévi-Strauss—rephrased this key speculation arguing if "the narrative of past events, subject usually in our culture, from the Greeks onward, to sanction of the historical 'science', [...] is really different, for some specific trait and an indisputable relevance, from the imaginary narration, which we can find in epics, novels, drama?"

On an opposite interpretative angle, we have Carlo Ginzburg. "Under the influence of structuralism, historians oriented themselves towards the identification of structures and of relationships. This identification rejected the perceptions and the intentions of individuals, or turned them into independent experiences, thus separating knowledge from subjective consciousness. In parallel, the number, the series, the quantification, which Carlo Ginzburg has called Galilei's paradigm [1986, 96-125 and 200-213], drove history towards a rigorous formulation of structural relationships, the establishment of whose laws became its mission" (Vendrix 1997, p. 65). The synopsis provided by the publisher for Carlo Ginzburg essay collection (Ginzburg 2012a), states that he "takes a bold stand against naive positivism and allegedly sophisticated neo-scepticism. It looks deeply into questions raised by decades of post-structuralism: What constitutes historical truth? How do we draw a boundary between truth and fiction? What is the relationship between history and memory? How do we grapple with the historical conventions that inform, in different ways, all written documents?"

Bernard Williams' famous statement that "the legacy of Greece to Western philosophy is Western philosophy" (Williams 2006, p. 3) is particularly true in this circumstance, because Plato's iconic quote from the *Apology of Socrates* (399 BCE) still provides the exact framework: *The unexamined life is not worth living* (Ὁ δὲ ἀνεξέταστος βίος οὐ βιωτὸς ἀνθρώπῳ, Apology of Socrates, 38a). Life is not worth living without ἔλεγχος/*elenchus*, that is examination, argument of disproof or refutation, dialogue; cross-examining, testing, scrutiny especially for purposes of refutation. Such is the Socratic *elenchus*, often referred to also as *exetasis* or scrutiny and as *basanismus* or assay (Vlastos 1983).

Since Herodotus of Halicarnassus (c. 484–c. 425 BCE) in Classical Antiquity, Lorenzo Valla (c. 1407–1457) in the Renaissance, Leopold von Ranke (1795–1886) in Modern Times, and Marc Bloch (1886–1944) in the twentieth century, the critical assessment of the authenticity and reliability of historical sources is the basic and fundamental tool that historians have been using as a *condicio* sine qua non to acquire their data and establish relations such as cause-effect among them (Galasso 2000, pp. 293–353, Ginzburg 2012a, pp. 7–24). While the "procedures used to control and communicate the truth changed over the course of time" (Ginzburg 2012a, p. 231), and the use of the same data can be dramatically different in various accounts bearing on the same past events across time, space, and cultures as well (Grafton and Marchand 1994; Guldi and Armitage 2014; Wang 2016).

Thus, the historians' key problematics have endured for a long period of time. In 1986, Carlo Ginzburg, in his seminal essay on *Clues: Roots of an Evidential Paradigm*, highlighted how history shares with two pseudo sciences, divination and physiognomics, not only roots but also their derivative sciences, law and medicine, that "conducted their analysis of specific cases, which could be reconstructed only through traces, symptoms, and clues. For the future, there was divination in a strict sense; for the past, the present, and the future, there was medical semiotics in its twofold aspect, diagnostic and prognostic; for the past, there was jurisprudence" (Ginzburg 1989, pp. 104–105; Momigliano 1985).

### *18.1.2 The Historians' Big Data in a Computational Perspective*

The electronic computer radically changed at all levels the ways our society and economy work (Robertson 1998, 2003). Historians are fully aware of the importance of this technological turn for the advancement of historical research (Ladurie 1973–1978; Galasso 2000, pp. 311–315; Ginzburg 2001; Cohen and Rosenzweig 2005). In principle, the historian is not refractory to new technologies: all historians went digital, in one way or another. They "have been actively programming since the 1970s as part of the first two waves of computational history" (Graham et al. 2016, p. 58).

Today, computers can do for historians what they did, for example, for mathematicians and chemists in the twentieth century, both at the level of capacity of observation and theoretical speculation (Robertson 1998). For example, chemists used to create models of molecules using plastic balls and sticks. Today, the modelling is carried out in computers. In the 1970s, Martin Karplus (Université de Strasbourg, France and Harvard University, Cambridge, MA, USA), Michael Levitt (Stanford University School of Medicine, Stanford, CA, USA), and Arieh Warshel (University of Southern California, Los Angeles, CA, USA) laid the foundation for the powerful programs that are used to understand and predict chemical processes. Computer models mirroring real life have become crucial for most advances made

in chemistry today, and on 9 October 2013, the Royal Swedish Academy of Sciences decided to award the Nobel Prize in Chemistry for 2013 to them "for the development of multiscale models for complex chemical systems".

However, after the "Digital Humanities Moment" (Graham et al. 2016, pp. 37–72), when historians started delving into data management and experimenting with various software to shed new light on their data sets, they seem to find it more difficult to take full advantage of the fact that computation itself is again *morphing*, as William Brian Arthur would say (Arthur 2009, pp. 150–151). Machine learning algorithms, one of computation's key technologies, underwent radical change and have now opened new horizons to the automation and speed of discovery (Domingos 2015). In this third wave of computational history the barriers of entry to powerful computing and big data have never been lower for the historian (Graham et al. 2016, p. 58). So, it should be more attractive and easier for historians to step in. But, in practice, it is more complicated because the question of the sources—which keeps on being of the essence to the historian's craft at each dramatic technological turn (oral-to-written, handwritten-to-printed, analog-to-electronic, and now from mathematical to algorithmic computation)—is acting as a bottle-neck. Let us explain why and how.

These expanded research capacities can allow new computational-driven research questions (and new answers): What shall the historian do having *all* data available in a digitalized form accessible in any language? What are the implications when *all* research materials are digitized and searchable through metadata in any language? Can we understand the mechanisms of convergence/divergence between local communities and international networks? How can the same networks/people bring new wealth and development, or generate war and poverty? Which dynamics and mechanisms operate in the world systems of individuals, families, cities, and countries? When we know the relationship between *all* (past) facts, *all* their (still present) traces/evidence, and *all* historiographical interpretative accounts, what kind of wisdom can be built on them? Is it possible to model bottom-up universal laws to influence the future? (Nanetti and Cheong 2016, p. 8).

Since the introduction of punch cards to enter data into computers, historians started to create large data sets that may be analysed computationally. In the 1970s, the French historian Emmanuel Le Roy Ladurie was the first to foresee the implications of the use of the computer in historical studies: "History based on computers/information technology is not limited to a very specific category of research, but also leads to the establishment of an 'archive'. Once transferred to tape or punched cards, and after having been used by a first historian, the data can in fact be stored for future researchers, who want to find non-experimented correlations" (Ladurie 1973–1978, p. I, 3).

Since then, in their daily research activities, historians are producing and accumulating extremely large digital datasets, in different languages and formats. More and more historical databanks are becoming available on the Internet. Thus, big data are becoming part of the historian's craft, worldwide. As more historical databases come online and overlap in coverage, historians and history as a discipline needs more and more big data approaches to cope with the increasing volume of

available sources and interpretations. Despite these big data, so far, big results are at the horizon but not yet clearly visible. Why?

### *18.1.3 What Prevented 'Big' Results from Emerging so Far?*

Cognitive computing borrows methodologies from two other disciplines, artificial intelligence and signal processing, for the simulation of human thought processes, while computational history aims to simulate the historian's craft, in a computerized model. Being at the very birth of artificial intelligence and automatic signal processing, current scholarship and technology may have science fiction dreams, but cannot have the presumption to automatize history as a whole, because its data volume and complexity are still far beyond any available digital storage system capacity and machine learning capability (Pavlus 2015). Nonetheless, computational history can be extremely relevant to develop a new and more efficient study of primary sources and secondary literature supporting the perennial historical chase for truth.

The bottle neck is the exegesis of the sources, because before dealing with big outputs, we need to work on big inputs. The ontology adopted for the definition of the entities and properties of databases is at the heart of the visualisation processes that can allow agent-based modelling to shed new light on historical records. Thus, computational history, before getting into the debate on the laws and purpose of history (Gilbert 1990; Popper 1999, pp. 105–115), is called to agree upon standardized methods to define machine-readable ontologies for both data (items known or assumed as facts) and the relationships among data (i.e., information, facts provided or learned about something or someone), which can be automatically extracted from primary and secondary sources, and possibly allow to expand, quantitatively and qualitatively, historical evidence, that is the available body of facts that the historian uses to judge whether a belief or proposition is true or valid.

In a cognitive computing perspective, this process can be rephrased as provenance-based validation (Wong et al. 2005). In the adoption of such a practice, historical records need to be comprehensively decomposed into unambiguous fields in order to be able to feed machine learning algorithms, which, firstly, can engineer evidence-fact-event relationships in both primary sources and secondary literature, and, secondly, build models of historical phenomena accounts in local, regional, and global historical scenarios (e.g., in our case study, trade-conflict-diplomacy relationships).

Hence, this paper (re)address the question of the sources and aims to provide some solutions and facilitate this new 'macroscopic' computational turn in historical studies. The solution that we propose to fill the gap comes in two stages: (1) to restructure the computation of sources using big data automatic narratives to extract facts from them and see their potential interconnections; and (2) to look at intensity in the flow of facts to identify events as tipping points (Gladwell 2000) in societies' natural nonlinear life using agent-based big simulations.

Firstly, historical data are seen by computer science people as unstructured, that is, historical records cannot be easily decomposed into unambiguous fields, except for the population and taxation ones, which are rare and scattered throughout space and time till the nineteenth century. This fact, in a computational perspective, prevent taxation and population databases to be scalable and aggregated with other datasets. An evident demonstration for taxation records is the *Online Catasto of Florence*. It is a searchable database of tax information for the city of Florence in 1427–1429 (c. 10,000 records uploaded till 1969) based on the work by David Herlihy and Christiane Klapisch-Zuber, Principal Investigators, *Census and Property Survey of Florentine Dominions in the Province of Tuscany, 1427–1480*.

Secondly, machine-learning tools developed for structured data cannot be applied as they are for historical research. Both the exegesis of primary historical sources, and the analysis of how those same primary sources have been selected and interpreted in various historiographical narratives are of the essence in this issue. The historians are required to shift from generalization to conceptualization, because univocal distinctions among theoretical units (e.g., evidence, fact, event) and historical phenomena (e.g., trade, conflict, diplomacy) become necessary conditions to generate new computational ontologies for databases (Guarino et al. 2009) and their application in agent-based modelling for historical simulations (Gavin 2014).

## 18.2 Big-Data Automatic Narratives as a Prerequisite for Big Simulations

According to Thomas R. Gruber (1993, 1995), a computational ontology requires a research domain to share an explicit formal specification of the domain terms themselves and their reciprocal relationships. Following Gruber's methodology, Andrea Nanetti extracted from the Morosini Codex (1205–1433) a coherent set of indexing terms (Nanetti 2010, pp. xvii–xix; pp. 1853–2274) to aggregate data for the interactive study of global histories. This research project, started from the world as seen from Venice, is creating an international research team with the ambition to engage the scholars of all other coeval chronicles written in Chinese, Arab, Russian, Persian, etc. (Nanetti and Cheong 2016).

This Venetian beginning is highly relevant in global context for three main reasons. Firstly, the Morosini codex was the model for the subsequent Venetian vernacular historiography leading to the famous 58-volume *Diarii* (1496–1533) by Marin Sanudo the Younger (1879–1902). These primary sources, providing information on all the empires and cities having marketplaces in the inhabited known world (the oecumene), represent one of the most important international texts for late medieval European and Mediterranean history. They deal with innumerable political and economic records taken mainly from merchants' (news)letters and the Venetian council deliberations (Nanetti 2010, pp. xi–xvii). Secondly, the Mediterranean basin has the longest and best-studied record of the ways in which

human activities have transformed the world (Abulafia 2011, pp. i–xxxi). Thirdly, in a computational perspective, the time period between 1205 and 1533 provides just enough but not overwhelming data to imagine big simulations (Nanetti and Cheong 2016, pp. 22–25).

The system, to which this interactive study of global histories refers, is the intercontinental Afro Eurasian communication network, which was first investigated in a scholarly and comprehensive way by the German geographer Ferdinand Freiherr von Richthofen (1833–1905) in his magnum opus *China* (1877–1912). In 1876 and 1877, baron von Richthofen anticipated the results of his work in two lectures given in Berlin, at the German Geological Society (Waugh 2007, p. 3). On 6 May 1876, he significantly chose to dedicate the first one to the sea routes (Richthofen 1876). The second, given in 1877, was about the communications over land (Richthofen 1876).

In this system, the actions (i.e., key relationship among events) have been identified in trade, conflict, and diplomacy. The agents (i.e., the historical actors) chosen for the simulations are in first instance the governments, which allow us to analyse continuity and change patterns in trade-conflict-diplomacy relationships among events at a world scale. On a higher level, this automatic extraction of key narratives from a historical database allows historians to formulate hypotheses on the courses of history, and also allows them to test these hypotheses in other actions or in additional data sets.

### *18.2.1 Automatic Source Provenance Identification and Facts-Evidence-Event Validation*

As the name implies, the past is an era gone by: it is no longer with us in the present. Historians use traces (the poor remains, still extant in the present) of what happened as clues to select, investigate, and judge events of the past. We think and speak of a past event as *factual* if someone or something we trust provides evidence for it. We consider accounts of such events *truthful* if trustworthy people wrote them down. By the time we read the accounts, we can have a variety of different evidence, from one single record written once in an otherwise proven *truthful* chronicle to chains of endorsements by *trustworthy* people, and therefore we consider such accounts *trustworthy*.

We frequently find two accounts that are highly similar in two or more sources, but with noticeable differences between them. Do these then refer to the same event, or to separate events? For events that appear in some sources but not in others, how would we know they are real? Similarly, for accounts that are highly similar, how would we know if they refer to the same event? Historians learn to judge the authenticity of historical records as part of their training, and become better over time. However, in the era of Big Data, the amount of data and records

will overwhelm historians. Therefore, we need the computer to help us validate the historical records if we want the process to be scalable.

To do so, we (re)propose to decompose historical records into their elementary constituents: who, what, when, where, why, and how. All elements must be demonstrably factual before the record can be considered factual. In other words, if the actor reported in an account appears also in other accounts (especially competing ones), the actor is likely to be a real person or a real institution in the past. On the other hand, if the actor appears only in one account, and is imbued with incredible or inconsistent attributes, there is a good chance it is made up. It turns out that checking the consistency in the profile of an actor is non-trivial, because different accounts may refer to the same actor using different names that may sound similar. Similarly, consistency in different accounts can also help us establish the validity of events, locations, motives, and actions.

In the validation process described above, we see that 'who', 'what', 'when', 'where', 'why', 'how' are the basic building blocks of our knowledge about the world. By themselves, they do not amount to much. For example, 'Marco Polo' may appear in multiple accounts, and based on this consistency we thus suspect his existence in the past as factual (Orlandini 1913). The consistent accounts thus provide *evidence* for the existence of 'Marco Polo'. Similarly, 'Catai' appears in multiple accounts, in manners that suggest that it refers to a place (Yule and Cordier 1913–1916). We thus establish 'Marco Polo' and 'Catai' as factual data. This is to be distinguished from non-factual data, which can refer to beliefs, whose contents may not be factual, but their existences are not in doubt.

By themselves, data are not very insightful. As we learn more about the world around us, we start to draw relationships between data. For example, 'Marco Polo in Catai' tells us more about 'Marco Polo' and 'Catai', more than the what we can infer from the separate factual existence of 'Marco Polo' and 'Catai' (Orlandini 1926). In the same way, we can understand when relationships are counterfeit. For example, 'Jacob of Ancona'—the supposed author of a book of travels, in which he was assumed to have reached 'Catai' in 1271, 4 years before Marco Polo—ceased to exist in the historical landscape when his account of 'Catai' was demonstrated to have been forged in the twentieth-century by David Selbourne (Halkin 2001).

We call this level of knowing about the outside world 'information', which allows us to say something about factual data and their relationships. In this classification scheme, historical facts are information decomposable into data entities and their relations. A historical event, though seemingly more complex than historical facts, is a collection of interrelated historical facts, but remains at the level of 'information' that is structured into a narrative. Here let us warn that the 'why' element of a narrative is extremely difficult to validate and establish as fact, because motives frequently depend on the actors interpreting them, while the actor responsible for an action may not provide a written account of its motive, truthful or otherwise. Motives are also notoriously susceptible to reinterpretation in subsequent accounts, for reasons that are difficult to uncover. Establishing the factual status of a motive

is thus a major challenge, since the consistency criterion for validation frequently fails.

In the Data, Information, Knowledge, Wisdom (DIKW) hierarchy popularized by Ackoff (1989), knowledge lies above information in our knowing of the world, and wisdom represents the highest level of knowing. In the DIKW hierarchy, knowledge is a collection of information, organized into a procedure, for acting on the world to solve problems. Wisdom is knowing when to act and when not to, because there may be no value in solving some problems, or because we need to prioritize which problem we solve first. The historian's goal for studying history is ultimately wisdom, but to acquire it we must pass through the knowledge stage. To get to this prescriptive and proactive stage of knowing, modeling and simulation is necessary.

But before we describe how to build ABMs based on historical events, and how to simulate these ABMs to obtain counterfactual histories, let us highlight the different ways historians and physical/computer scientists define data, information, facts, evidence, and events. This comparison is shown in Table 18.1.

### *18.2.2 Complex Networks Visualisations of Historical Datasets. Trade-Conflict-Diplomacy Relationships as a Key Case Study*

Assuming that we have solved the problem of provenance and validation, and have successfully created a database of historical events in narrative format ('who', 'what', 'when', 'where', 'why', 'how'), we now have the problem of extracting insights from such a database. After we have also created an ABM, this would be much easier, because we can use simulations to fill in the gaps and create an animation of the events. However, insights are precisely the ingredients needed to create the ABM, so we are faced with a chicken-and-egg problem. Therefore, in place of the ABM animation, we turn to model-free visualization strategies to discover the most important stories that can emerge from the database. We do this using a complex-network approach.

First, we create a multigraph where the nodes represent actors (which in our preliminary study are governments), and nodes can be connected by three different types of weighted links, one for conflict, another for diplomacy, and the last for trade. Other actions can also be included in follow-up studies. We start by setting the weights of all links to zero. For a given event, we identify the actors involved, and also the action. For example, in the record of Venice going to war with Rome, the actors are the governments of Venice and Rome, and the action is war. Therefore, we add one to the weight of the conflict link between Venice and Rome. After we have gone through the full list of events in the database, we end up with a time-integrated multigraph (see Fig. 18.2). We can use community detection methods in the complex network literature (Girvan and Newman 2002; Blondel et al. 2008;

**Table 18.1** How historians and other humanities scholars define data, information, evidence, facts, evidence, and events differently from physical and computer scientists

| | History/Humanities | Physics/Computing |
|---|---|---|
| Data | Things known or assumed as facts, making the basis of reasoning (philosophy) | The quantities, characters, or symbols, on which operations are performed by a computer |
| Information | Facts provided or learned about something or someone | Data as processed, stored, or transmitted by a computer |
| Facts | – A piece of information used as evidence or as part of a report or news article<br>– The truth about events as opposed to interpretation (law)<br>– The available body of facts or information indicating whether a belief or proposition is true or valid<br>– Information given personally, drawn from a document, or in the form of material objects, tending or used to establish facts in a legal investigation or admissible as testimony in court | Synonymous with information |
| Evidence | – The available body of facts or information indicating whether a belief or proposition is true or valid<br>– Information given personally, drawn from a document, or in the form of material objects, tending or used to establish facts in a legal investigation or admissible as testimony in court (law) | Collection of data demonstrating the reproducibility (consistency) of an information/fact |
| Event | A thing that happens, especially one of importance | A single occurrence of a process (physics) |

Alvarez et al. 2015) to identify groups of nodes that are persistently at war, at peace, or trading with one another. If such groups exist, historians must then find cultural and geopolitical reasons to explain them.

Using this database, it is possible to work on the identification of transient groups of nodes at war, at peace, or trading with one another. To discover them, we need to construct a timeline of complex multigraphs representing different periods in history. Again, patterns that we discover here represent the coarse flow of history, and explanations are called for. Finally, we can perform a *time-resolved analysis* of the database at the event level, to identify the *key events* that form the ingredient for our explanations (Fig. 18.3), and also *key periods* that historians should focus their attentions on (Fig. 18.4).

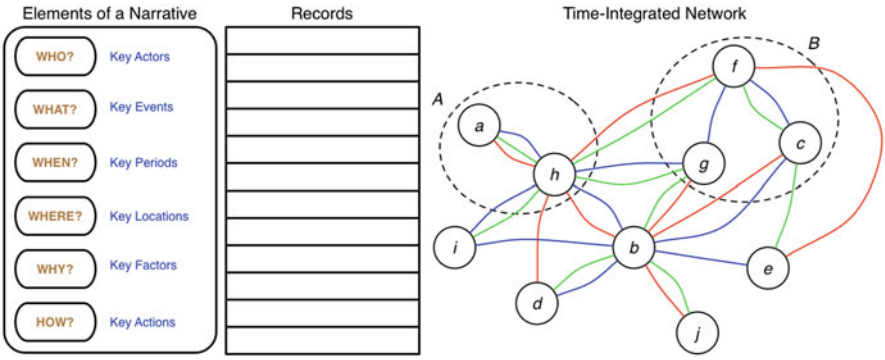



**Fig. 18.2** Constructing a time-integrated complex multigraph using the list of records in a historical database. In the database, records are organized into narratives (with elements 'who', 'what', 'when', 'where', 'why', 'how'). In the complex network, nodes represent governments, while links represent actions (red for conflict, blue for diplomacy, and green for trade). The dashed circles represent schematically persistent groups of nodes discovered using community detection methods

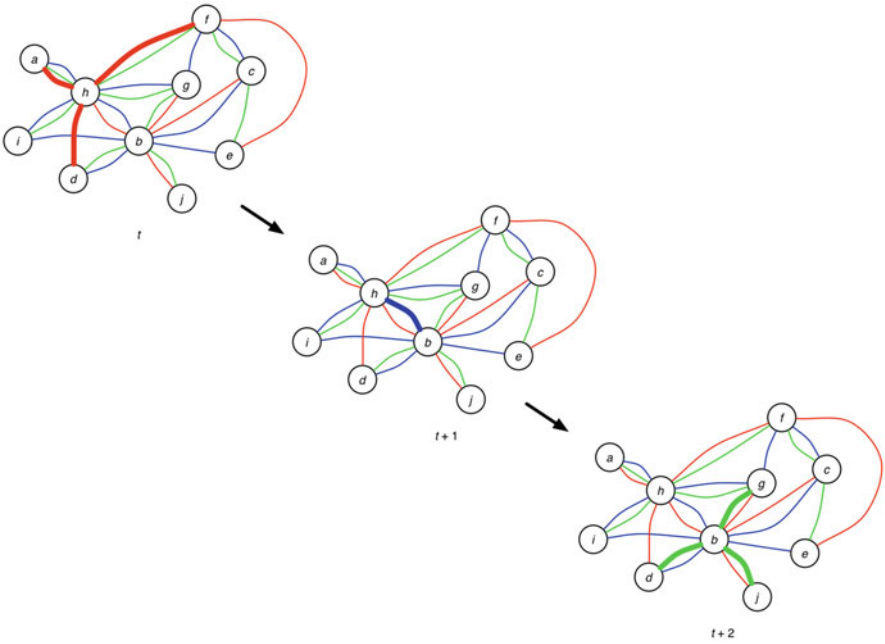



**Fig. 18.3** In this figure, we highlight active events during successive time windows $t, t+1, t+2$ by showing them as thick links. The convergence of events on $h$ in time window $t$ and divergence of events from $b$ in time window $t+2$ points to the event involving $h$ and $b$ in time window $t+1$ as a *key event*

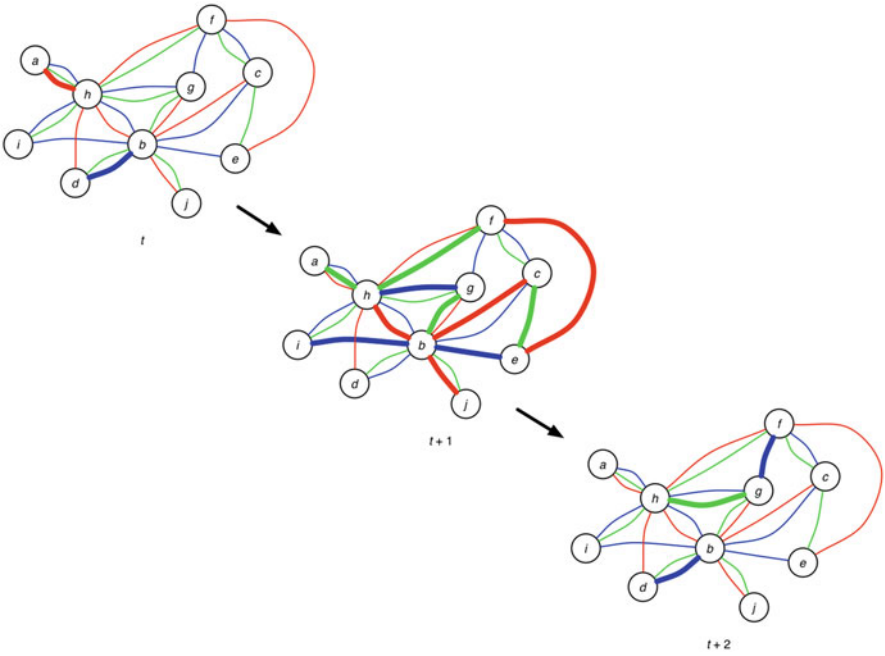



**Fig. 18.4** In this figure, we highlight active events during successive time windows $t, t+1, t+2$ by showing them as thick links. The dramatic increase in number of events in time window $t+1$ relative to time windows $t$ and $t+2$ points to time window $t+1$ as a *key period*

### *18.2.3 Formal and Informal Models. Going from Patterns to Models*

From the time-integrated complex network and the time-resolved analysis based on it, we can identify many patterns that we can use to develop large-scale historical narratives. This will feel easy, and the narratives compelling, because we have extracted key narrative elements from the historical database. Without the method of automatic narratives, this historiography would take considerably more effort from the historian. Unfortunately, historical databases today are not designed to support inquiries through such machine learning strategies, however well designed their search tools for enquiries by human historians. To support pattern discovery by automatic narratives and other forms of data visualization, existing historical databases must be systematically reorganized.

However, we must not stop at data visualization and finding patterns, which in some sense represent informal models. In our quest for historical understanding, such patterns are what economists today call stylized facts (Arthur 2014). They can be compared to Kepler's laws of planetary motion, discovered from the astronomical 'Big Data' collected by Tyge Ottesen Brahe (1546–1601, first critical

edition by Rawlins 1993). Informal models produced by such macroscopic research methodologies (e.g., Schich et al. 2014) can be articulated using words, and are good as scaffolding for organizing thoughts. However, they have neither explanatory nor predictive powers, because we know what they are, but not why they are the way they are. To be able explain historical phenomena and predict when they will recur, we need the historical equivalent of Newton's laws. These are formal models, which can be stated either in equation form or as a set of rules (ABMs). To emulate how Newton's laws explain Kepler's laws, and understand the role of change in historical studies, Peter Turchin built top-down models describing how civilizations expand, through agriculture or military conquests (Turchin 2003). By extracting a few parameters from highly aggregated historical data, Turchin was able to show how closely his technology-driven *cliodynamics* follow the historical trajectories of the major civilizations in the world (Turchin and Nefedov 2009).

Encouraging as it may seem, *cliodynamics* overlook the role of human agency. Human societies always have the need to make decisions, whether it is to trade, to go to war, or to sue for peace. Unfortunately, if we follow the *cliodynamics* approach to its logical conclusion, no part of history would have turned out differently, i.e. history is inevitable. This is contrary to what Gordon Woo is theorizing in his calculation of catastrophes (2011). To put human agency back into history, and allow history to be contingent upon the decisions made (and therefore admit counterfactual histories), our ultimate goal remains the creation of historical ABMs.

## 18.3 Agent-Based Modelling and Simulations (ABMS)

Compared to equation-based modelling, which goes back as far as Newton in the seventeenth century, agent-based modelling and simulation (ABMS) has a very short history. While there may be early thinkers who contemplate the collective consequence of decisions made by many individuals, as a mode of inquiry ABMs can only be regarded to have started in the middle of the twentieth century. This is because the history of ABMS cannot be divorced from the development of the electronic computer. As early as 1971, Nobel Laureate in Economics Thomas Crombie Schelling developed a toy model of segregation, in which happy agents stay put while unhappy agents move (Shelling 1971). Agents are happy if more than a certain fraction of their neighbours are similar to themselves, and are unhappy otherwise. Using coins and graph paper to run the simulations, Schelling was able to show that any level of preference for neighbours similar to themselves will lead to segregated neighbourhoods.

In the 1980s, when the electronic computer was starting to become popular as a research tool in universities, political scientist Robert Axelrod hosted a tournament for computer programs to play the Prisoner's Dilemma against each other (Axelrod 1980). In this first true agent-based simulation, the agents were the computer programs that had to decide what strategy to use when playing against other

computer programs who are also capable to deciding on or changing their strategies. Later in the 1980s, we also saw the development of ABMs called *Boids* by Craig Reynolds to explain the flocking of birds and the schooling of fishes (1986). In these models, the agents follow three simple rules: (1) move in the average direction of neighbours, (2) stay close to neighbours, and (3) avoid collisions with neighbours, and adjust their velocities accordingly.

Around the same time, computer scientist John Holland and economist Brian Arthur were also developing the world's first artificial stock market, where adaptive agents in the form of computer programs buy and sell stock according to their predictions of how the stock price will change (Palmer et al. 1994). In this Santa Fe Institute's Artificial Market model, agents trade with their best prediction model, out of a list of prediction models they maintain. These prediction models then evolve over time by random mutations or by mating between models. They found that the market and the agents never settle down, and are constantly generating booms and busts like in the real market. In 1991, John Holland and John Miller published a paper referring to their model as an 'agent-based model', and the name stuck (Holland and Miller 1991).

Since then, the field of ABMs expanded rapidly. While economists were the first adopters of this new computational methodology, ABMs quickly spread to other social sciences. In particular, Joshua Epstein and Robert Axtell created the *Sugarscape* ABM to explain the rise and fall of a large North American Indian settlement, and popularized ABM for social scientists by writing their book on this project (Epstein and Axtell 1996). As of now, ABM has become a fairly mature technology. There are now major conferences on ABM, and also summer/winter schools on ABM attended by postdoctoral researchers and PhD students, taught by leading experts in the field. However, the spread of ABM as a tool has not been uniform across the social sciences and humanities, history being a late adopter. In this section, we will first describe how historians can build ABMs, what they can learn from ABMs, and how ABMs can help them transform history as a discipline.

### *18.3.1 Requirements and Prescriptions to Build Data-Driven Simulations*

According to Cain (2014, p. 1), "a mathematical model is an attempt to describe a natural phenomenon quantitatively. Mathematical models in the molecular biosciences appear in a variety of ways: some models are deterministic while others are stochastic, some models regard time as a discrete quantity while others treat it as a continuous variable, and some models offer algebraic relationships between variables while others describe how those variables evolve over time". ABMs, though rule-based instead of equation-based, share many of the above character-

istics. Historians wishing to reap these benefits must first learn how to build ABMs. To support the development of such models, there are specific requirements besides time and space that historians must take into consideration, when they construct their databases or re-structure them accordingly.

Firstly, in the ontology of the entities that are used in the database engineering, historians should identify:

- Necessarily, agents, that are the entities, considered as individuals and/or collective wholes (i.e., governments, families, etc.), capable of setting goals, interact with other agents, and react to the environment and its changes;
- Necessarily, events, from which they can extract the actions needed to achieve goals;
- Possibly, conditions (due to the environment and/or other agents), that are the most important external factors influencing the decision-making process of the agents;
- Possibly, preferences (built-in conditions), that are the arbitrary choices made by the agents to pursue their goals.

For example, in the Engineering Historical Memory (EHM) project, as first basic entities, besides time and space (always included), we decided to identify:

- Governments and families as agents;
- Trade, conflict, and diplomacy as filtering categories for actions like single treaties, embassies, travels from one place to another, buy/sell/loan/stockpiling of goods, battles, shipwrecks, sieges, wars, etc.;
- Non-agent entities (goods, coins, ships, etc.) as conditions or preferences.

Secondly, the database needs to facilitate or at least allow for the retrieval of agent-action-condition (who did what and why) triplets, so that historians can visualise the frequencies of actions taken under specific conditions by agents and how they depend on time and space.

We then codify the most frequent or most important agent-action-condition triplets as our rules for the ABM. If preferences can be inferred, these will also be included. Otherwise, we may—through consulting human experts—endow our agents with heterogeneous preferences consistent with behaviours of peoples of that time and space, as input parameters for our ABM. At this point, we are ready to write the computer program to simulate the ABM.

If what is missing in computational history is the macroscopic modelling, which grasp big data "through a process of compression, by selectively reducing complexity until once-obscure patterns and relationships become clear" (Graham et al. 2016, p. 1), we propose to fill this crucial gap following the *Annales* experience, and the consequent development of microhistory from the *histoire événementielle* (Le Goff and Nora 1974, 1985; Burguière 2006, 2009). In our vision, macroscopic models can be inferred by microhistory. In this perspective, big history is what emerges from *all* microhistories interactions.

Microhistory (2012a, b, pp. 193–214) studies well-defined single historical units/events to ask—as defined by Charles Joyner—"large questions in small

places" in contrast with large-scale structural views (Joyner 1999, p. 1). The most famous example being Carlo Ginzburg's *Il formaggio e i vermi* (in Italian 1976 and in English 1980). In the book, which is considered to have initiated this research field in historical studies, the author wrote: "The historians have long since learned that history is the history of men, not of the "great," and the closer you get up to everyday reality the better you decipher the past, and then grasp the sense of immediacy with the problems, the connections with today's present, i.e., history".

### *18.3.2 Under What Conditions can We Learn from Big Simulations?*

In his 2014 position paper, Michael Gavin argued for the adoption of ABM in history, as a means of encapsulating the complexity of historical events in terms of a small number of rules. Following Epstein and Axtell (1996), Gavin calls this feature of ABM 'generative simplicity'. However, Gavin does not explain how ABMs are to be built starting from data. We explained in Sect. 18.3.1 how ABMs can be built in a data-driven process (which we want to promote). More importantly, it is not clear whether the small set of 'generative' rules are static, or whether they are adaptive and can change in time in response to evolutionary pressures. As John Holland had demonstrated, while the meta rules are the same (mutation, mating, selective reproduction), the rules themselves never settles down, and we have "perpetual novelty" in the system (Holland 1989, Introduction). Can we say we have an understanding of historical processes in terms of this ever-changing set of rules?

Also, when we simulate the ABMs, we end up with a large number of simulated histories. What then do we mean, when we say that we understand the observed history with the aid of simulated histories? To unpack this, let us suppose we understand much of human preference and behavior, but we do not have complete data on the population. After building an ABM, we would need to make assumptions on the preferences of the agents. Naturally, different assumptions will give us different simulation outcomes. However, if we believe that history can be 'understood', then the number of outcomes that are qualitatively different must be small compared to the number of assumptions we simulate. This means that a large number of simulations with different assumptions will give rise to the same qualitative outcome. There are a few inevitable results we can discuss.

First, some outcomes can emerge from a huge number of assumptions, whereas other outcomes appear only for a much smaller number of assumptions. We say that the former outcomes are robust, and the latter outcomes are fragile. Outcomes are not equal in this sense, and we can classify them using ABMS. Second, since many different assumptions give rise to the same qualitative outcome, many aspects of our assumptions must be unimportant (for otherwise they would have changed the simulation outcome). It may be that only a few aspects of the assumptions are

important. This realization means that the outcome may be explained by a few key factors. Third, the historical trajectories leading to two qualitatively different (for example war versus peace) outcomes may follow each other closely until some point in time where they diverge. This point in time is when the simulated histories cross a tipping point. By comparing the key factors leading to the two different outcomes, and understand how they are different, we begin to have a better understanding of tipping points and regime shifts in history.

### *18.3.3 Tipping Points. A Scalable Solution to Investigate Change (i.e. The Fundamental and Nonlinear Force of History)*

Finally, to derive causal narratives of world history, and identify causative mechanisms and processes, we need to better understand what tipping points are and are not. In the discussion, above, we have already mentioned that a tipping point separates two qualitatively different set of historical trajectories. Certainly, a tipping point can be an event, and so the action associated with the event can be understood as a cause. However, let us make clear here that the action in the tipping point event is merely the cause of the event, but not the separation of historical trajectories. To understand this, we should think of 'the straw that broke the camel's back'. The laying of this straw onto the camel's back is clearly the tipping point, but it is no more causal than all the other strands of straw on the camel's back when it broke.

Ultimately, the causal narrative we would like to take away from this is that we have been adding load onto the camel, and thus drive the camel closer and closer to the tipping point. From the historical narrative extracted from the database, and the bundle of counterfactual trajectories that it is grouped with, the causal factor must be identified with the chain of key factors along these historical trajectories.

How then do we understand tipping points? Shortly after historical trajectories diverge, we can extract the chains of causal factors along each bundle of trajectories. We can compare these chains to identify the main difference in the chains of causal factors that lead to one bundle of trajectories going to one outcome, and another bundle of trajectories going to another outcome. This difference in causal factors, compared to the actions in the tipping point event, then tells us how small decisions that seemed inconsequential eventually turn out to have large impacts on the outcomes.

Finally, because we produced these counterfactual histories using a microscopic ABM, we can test in simulations what kind of changes to the preferences and behaviors of agents as the simulations is in progress will change the outcomes. Naively, if we have the preferences and behaviors of agents change continuously from the key factors of one outcome to the key factors of the other outcome, we should be able to change the outcomes for some of the simulations leading to an undesirable outcome.

## 18.4 Conclusions. Learning from Computational History

We repeatedly call for people to "learn from history". *Historia vero testis temporum, lux veritatis, vita memoriae, magistra vitae, nuntia vetustatis, qua voce alia nisi oratoris immortalitati commendatur?* By what other voice, too, than that of the orator, is history, the evidence of time, the light of truth, the life of memory, the directress of life, the herald of antiquity, committed to immortality? (Marcus Tullius Cicero, *De Oratore*, II, 36). If we read this famous Cicero's quote through the lens of the thermodynamic paradigm, which holds that a perfect description of a given moment or set of conditions in history would provide a knowledge of future conditions—and assume that "the new society comes into being in the womb of the old" (Lechte 2003, p. 106), our increasingly complex world should cherish as much as possible the treasure of human experiences (*the* data), to increase resilience and sustainability and to nurture innovation (Nanetti et al. 2013, pp. 104–105).

However, the circumstances surrounding historical episodes are never identical, and certainly not the same as the circumstances we find ourselves in. If war have been averted in the past because of certain diplomatic gambles, we may not be able to reuse them in the present day, because circumstances have changed, and also the actors have changed. Nevertheless, if we believe that history of our society is the result of selection between a small number of outcomes when it is presented with complex inputs, then ABM can help us understand the relationships between different outcomes, and transitions between outcomes is only possible for neighbouring outcomes in some sort of phase diagram of outcomes. More importantly, ABM simulations can help us identify the key factors driving the simulations to a particular outcome, and what is the nature of the tipping point separating this outcome from a neighbouring one.

Computational analysis can borrow models from ecology, evolution, dynamical systems, and complexity theory (e.g., Holland 1989, 2000, 2012), and relate them to historical processes to extend the humanities capabilities and give new strength to the framework and prescriptions of century-old philological, historical, art historical, and anthropological methodologies. In doing so, we can aggregate knowledge for a better understanding of the present and transmit knowledge to influence the future values we desire more.

By producing counterfactual histories from ABM simulations and comparing these against the recorded histories, we can detect tipping points. If we are to learn from history and not make the same mistakes that our forebears did, we must understand the reasoning and decision processes that led to these tipping points. Only then can we acquire the wisdom to steer human civilisation towards desirable outcomes, and master the art of living together, with the consciousness of how false beliefs can change history (Eco 1998).

In our vision, this narrative-driven analysis of historical big data can lead to the development of multiple scale agent-based models, which can be simulated on a computer to generate ensembles of counterfactual histories that would deepen our

understanding of how our actual history its related historiographies developed the way they did.

It entails the creation and advancement of databases (relational, graph, and hybrid), algorithms, computational, statistical, and complexity techniques and theories to solve formal and practical problems arising from the study, and the interpretation, conservation, and management of historical data and information.

In this way, the historians' major strength, their training in special units, can overcome its major weakness, the practical and ideological narrowness of specialised expertise, by adding to their data sets other historian's datasets and test their theories with multinational additional types of approaches that were not part of their training.

**Acknowledgments** This study has been funded by 2016 Microsoft Research Asia Collaborative Research Program and 2016 Microsoft Azure for Research. The research project, called Engineering Historical Memory (EHM), was first theorized by Andrea Nanetti in 2007, when he was Visiting Scholar at Princeton University. The actual web development initiated in 2012 when he was Visiting Professor at the University of Venice Ca′ Foscari, and since 2013 has been carried out at Nanyang Technological University (NTU Singapore), where he is Associate Chair (Research) in the School of Art, Design and Media, and has been funded among others by an NTU Start-up Grant (2014–2016 M4081357), 2014 Microsoft Research Asia Collaborative Research Program, 2015 Microsoft Azure for Research, and 2016 Microsoft Research Internship Program. This study contributed to the application and kick off of the NTU TIER 1 Grant (2017–2019) on ''Data Consolidation for Interactive Global Histories (1205–1533) within the NTU National and International Research Network: Towards an NTU Interdisciplinary Laboratory for Data-Driven Agent-Based Modelling and Simulations for Historical Sciences'' (2017–2020 M4011828). The domain of EHM (www.engineeringhistoricalmemory.com) is administrated by Meduproject S.r.l., an Italian Pte Ltd. company established in 2002 by Andrea Nanetti as academic spin-off of the University of Bologna (Department of Histories and Methods for Cultural Heritage Conservation), after having been awarded in 2001 a prize in the first Italian business plan competition devoted to projects with high content of knowledge and having been financially supported by the Italian National Agency for New Technologies, Energy and Environment.

## References

Abulafia, D. (2011). *The great sea: A human history of the Mediterranean*. New York: Oxford University Press.

Ackoff, R. L. (1989). From data to wisdom. *Journal of Applied Systems Analysis, 16*, 3–9.

Aiden, E., & Michel, J.-B. (2013). *Uncharted: Big data as a lens on human culture*. New York: Riverhead.

Alvarez, A. J., Sanz-Rodríguez, C. E., & Cabrera, J. L. (2015). Weighting dissimilarities to detect communities in networks. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, 373*, 2056. https://doi.org/10.1098/rsta.2015.0108.

Arthur, W. B. (2009). *The nature of technology. What it is and how it evolves*. New York: Free Press.

Arthur, W. B. (2014). *Complexity and the economy*. Oxford: Oxford University Press.

Axelrod, R. (1980). Effective choice in the prisoner's dilemma. *The Journal of Conflict Resolution, 24*(3), 379–403.

Big History Project. (2012–Present). Big History Institute, Macquarie University, Sydney, Australia. Retrieved January 15, 2016, from https://www.bighistoryproject.com.

Blondel, V. D., Guillaume, J.-L., Lambiotte, R., & Lefebvre, E. (2008). Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment, 2008*(10), P10008. https://doi.org/10.1088/1742-5468/2008/10/P10008.

Brahe, T. (1993). In D. Rawlins (Ed.), *Tycho's star catalog: The first critical edition* (Vol. 3, pp. 3–106). DIO.

Burguière, A. (2006). *L'École des annales: Une histoire intellectuelle*. Paris: Odile Jacob.

Burguière, A. (2009). *Annales school: An intellectual history*. Ithaca, NY: Cornell University Press.

Cain, J. W. (2014). Mathematical models in the sciences. *Cellular and Molecular Life Sciences.* https://doi.org/10.1007/978-1-4614-6436-5_561-1.

CHIA. (2011). http://www.chia.pitt.edu/. Accessed 15 Jan 2017.

Cohen, D. J., & Rosenzweig, R. (2005). *Digital history: A guide to gathering, preserving, and presenting the past on the web*. Philadelphia: University of Pennsylvania Press.

de Rosnay, J. (1979). *The macroscope: A new world scientific system*. New York: Harper & Row.

Digital Atlas of Roman and Medieval Civilizations. (2007–Present). Harvard University. Retrieved January 15, 2016, from http://darmc.harvard.edu.

Eco, U. (1998). Serendipities: Language and lunacy, (W. Weaver, Trans.). New York: Columbia University Press.

Eddington, A. S. (1929). *The nature of the physical world. The Gifford lectures 1927*. New York, Cambridge: Macmillan, Cambridge University Press See also the edition Annotated and Introduced by H. G. Callaway. Newcastle upon Tyne: Cambridge Scholars Publishing 2014.

Engineering Historical Memory. (2012–Present). Meduproject S.r.l. (spin-off company of the University of Bologna) and Microsoft Azure. Retrieved January 15, 2016, from http://www.engineeringhistoricalmemory.com.

Epstein, J. M., & Axtell, R. L. (1996). *Growing artificial societies. Social science from the bottom up*. Washington, DC: Brookings Institution, MIT Press.

Galasso, G. (1984). Fonti storiche [Historical sources]. In *Enciclopedia del Novecento, VII* (pp. 198–212). Roma: Istituto dell'Enciclopedia Italiana.

Galasso, G. (2000). *Nient'altro che storia [Nothing but history]*. Bologna: Società Editrice Il Mulino.

Gavin, M. (2014). Agent-based modeling and historical simulation. *Digital Humanities Quarterly*. Retrieved from digitalhumanities.org/dhq/vol/8/4/000195/000195.html#p2

Gell-Mann, M. (1997). The simple and the complex. In D. S. Alberts & T. J. Czerwinski (Eds.), *Complexity, global politics, and national security* (pp. 2–12). Washington, DC: National Defense University.

Gilbert, F. (1990). *History: Politics or culture? Reflections on Ranke and Burckhardt*. Princeton, NJ: Princeton University Press.

Ginzburg, C. (1980). *Il formaggio e i vermi. Il cosmo di un mugnaio del '500* [The Cheese and the worms: The cosmos of a sixteenth-century miller] (J. Tedeschi, A. Tedeschi, Trans.). Einaudi, Baltimore: Torino, Johns Hopkins University Press (Original work published 1976 (Italian)).

Ginzburg, C. (1989). Miti, emblemi, spie [Clues, myths, and the historical method] (J. Tedeschi, & A. Tedeschi, Trans.). Torino Einaudi. Baltimore: Johns Hopkins University Press (Original work published 1986 (Italian)).

Ginzburg, C. (2012a). Il filo e le tracce. Vero falso finto [Threads and Traces. True False Fictive] (A. Tedeschi, J. Tedeschi, Trans.). Bologna, Berkeley, CA: Feltrinelli, University of California Press (Original work published 1986 (Italian)).

Ginzburg, C. (2012b). Microhistory, two or three things that I know about it. In Idem 2012. Op. cit., 193–214.

Ginzburg, C. (2001). Conversare con orion. *Quaderni Storici, 23*(3), 905–913.

Girvan, M., & Newman, M. E. J. (2002). Community structure in social and biological networks. *Proceedings of the National Academy of Science of the United States of America, 99*(12), 7821–7826.

Gladwell, M. T. (2000). *The Tipping Point: How Little Things Can Make a Big Difference*. New York: Little Brown.

Gould, S. Jay. (1989). *Wonderful Life: The Burgess Shale and the Nature of History*. London: Hutchinson Radius.
Grafton, A. (1994). The footnote from de Thou to Ranke. In: A. Grafton, & S. L. Marchand (Eds.), Proof and persuasion in history. History & theory (Vol. 33, pp. 53–76), Hoboken, NJ: Wiley.
Grafton, A. (1995). *Die tragischen Ursprünge der deutschen Fußnote*. Berlin: Wagenbach.
Grafton, A. (1997). *The footnote: A curious history*. Cambridge: Harvard University Press.
Grafton, A., & Marchand, S. L. (1994). Proof and persuasion in history. In *History & theory* (Vol. 33). Middletown: Wesleyan University Press.
Grafton, A., Goeing, A., Michel, P., & Blauhut, A. (Eds.). (2013). *Collectors' knowledge: What is kept, what is discarded/Aufbewahren oder wegwerfen: Wie sammler entscheiden*. Leiden: Brill.
Graham, S., Milligan, I., & Weingart, S. (Eds.). (2016). *Exploring big historical data. The Historian's macroscope*. London: Imperial College Press.
Grinin L, A Korotayev (eds.). 2010. History & mathematics. Trends and cycles. Volgograd: 'Uchitel' Publishing House.
Gruber, T. R. (1993). A translation approach to portable ontologies. *Knowledge Acquisition, 5*(2), 199–220 tomgruber.org/writing/ontolingua-kaj-1993.htm.
Gruber, T. R. (1995). Toward principles for the design of ontologies used for knowledge sharing. *International Journal of Human-Computer Studies, 43*(4–5), 907–928 tomgruber.org/writing/onto-design.htm.
Guarino, N., Oberle, D., & Staab, S. (2009). What is an *ontology*. In S. Staab & R. Studer (Eds.), *Handbook on ontologies* (pp. 153–176). Berlin: Springer Verlag.
Guldi, J., & Armitage, D (2014). The history manifesto. Cambridge University Press, Cambridge. New updated version 5th February 2015, Retrieved January 7, 2017, from doi: https://doi.org/10.1017/9781139923880.
Halkin, H. (2001). The strange adventures of Jacob d'Ancona: Is a memoir of China purportedly written by a thirteenth-century Jewish merchant authentic? And if not, what then? *Commentary Magazine, 111*(4). https://www.commentarymagazine.com/articles/the-strange-adventures-of-jacob-dancona.
Hirschi, C. (2011). *The origins of nationalism. An alternative history from ancient Rome to early modern Germany*. Cambridge: Cambridge University Press.
Hitzbleck, K., & Hübner, K. (Eds.). (2014). *Die Grenzen des Netzwerks 1200–1600*. Ostfildern: Jan Thorbecke Verlag der Schwabenverlag AG.
Holland, J. (1975). *Adaptation in natural and artificial systems. An introductory analysis with applications to biology, control, and artificial intelligence*. Ann Arbor: University of Michigan Press.
Holland, J. (1989). *Induction: Processes of inference, learning, and discovery*. Cambridge: MIT Press.
Holland, J. (2000). *Emergence: From chaos to order*. Oxford: Oxford University Press.
Holland, J. (2012). *Signals and boundaries: Building blocks for complex adaptive systems*. Cambridge: MIT Press.
Holland, J. H., & Miller, J. H. (1991). Artificial adaptive agents in economic theory. *American Economic Review, 81*(2), 365–371.
Jockers, M. L. (2013). *Macroanalysis: Digital methods & literary history*. Champaign: University of Illinois Press.
Joseph, B. D., & Janda, R. D. (2003). On language, change, and language change – Or, of history, linguistics, and historical linguistics. In B. D. Joseph & R. D. Janda (Eds.), *The handbook of historical linguistics*. Oxford: Blackwell Publishing.
Joyner, C. W. (1999). *Shared traditions: Southern history and folk culture*. Urbana: University of Illinois.
La, M. L. V. *François de. 1669. Œuvres* (Vol. 15). Paris: Libraire Louis Billaine [The fifth (and final) edition has been published in Dresde by Michel Groll between 1756 and 1759: Œuvres de François de La Mothe Le Vayer, conseiller d'État ordinaire, etc. Nouvelle édition revue et augmentée. Imprimée à Pfœrten, et se trouve à Drede chez Michel Groell, 7 tomes en 14 volumes in-8].

Ladurie, E. L. R. (1973–1978). *Le Territoire de L'Historien* (Vol. 2). Paris: Gallimard.
Le Goff, J., & Nora, P. (Eds.). (1974). *Faire de l'histoire* (Vol. 3). Paris: Gallimard.
Le Goff, J., & Nora, P. (Eds.). (1985). *Constructing the past: Essays in historical methodology*. Cambridge: Cambridge University Press [This book presents a selection of ten of the most significant contributions to the three-volume Faire de l'histoire, 1974].
Lechte, J. (2003). *Key contemporary concepts. From abjection to zeno's paradox*. New York: SAGE.
Manning, P. (2013). *Big Data in History*. Basingstoke UK: Palgrave Macmillan.
Manning, P. (2015). A World-Historical Data Resource: The Need is Now. *Journal of World-Historical Information* 2–3/2:1–6.
Mitchell, M. (1996). *An introduction to genetic algorithms*. Cambridge: MIT Press.
Momigliano, A. (1985). History between Medicine and Rhetoric. *Annali della Scuola Normale Superiore di Pisa. Classe di Lettere e Filosofia. Serie III* XV/3:767–780.
Moretti, F. (2005). *Graphs, maps, trees: Abstract models for literary history*. New York: Verso.
Nanetti, A. (2010). *Il Codice Morosini: Il Mondo Visto da Venezia (1094–1433) [The Morosini codex: The world as seen from Venice (1094–1433)]* (Vol. 4). Spoleto: Foundation CISAM.
Nanetti, A., & Cheong, S. A. (2016). The world as seen from Venice (1205-1533) as a case study of scalable web-based automatic narratives for interactive global histories. *The Asian Review of World Histories, 4*(1), 3–34. https://doi.org/10.12773/arwh.2016.4.1.003.
Nanetti, A., Cheong, S. A., & Filippov, M. (2013). Interactive global histories: For a new information environment to increase the understanding of historical processes. In *2013 culture and computing* (pp. 104–110). Los Alamitos: IEEE Computer Society.
Nanetti, A., Lin, C.-Y., & Cheong, S. A. (2016). Provenance and validation from the humanities to automatic acquisition of semantic knowledge and machine reading for news and historical sources indexing/summary. *The Asian Review of World Histories, 4*(1), 125–132. https://doi.org/10.12773/arwh.2016.4.1.125.
Olstein, D. (2015). *Thinking history globally*. New York: Palgrave Macmillan.
Online Catasto of Florence. (1969). In: D. Herlihy, C. Klapisch-Zuber, R. Burr Litchfield, A. Molho (Eds.), Brown University (pp. 1427–1429). Retrieved January 2, 2017, from http://cds.library.brown.edu/projects/catasto/overview.html.
Orlandini, G. (1913). *Origine del Teatro Malibran. La Casa dei Polo e la Corte del Milion*. Venezia: Nozze Alverà-Trevisanato.
Orlandini, G. (1926). Marco Polo e la sua famiglia. *Archivio veneto-tridentino, 9–10*, 1–68.
Palmer, R. G., Arthur W. B., Holland J. H., LeBaron B., Tayler P. (1994). Artificial economic life: A simple model of a stock market, Physica D 75:264–274. Retrieved January 15, 2017, from https://www.phy.duke.edu/~palmer/papers/arob98.pdf.
Paolucci, M., & Picascia, S. (2011). Enhancing collective filtering with causal representation. In *2011 culture and computing* (pp. 135–136). Los Alamitos: IEEE Computer Society.
Pavlus, J. (2015). A new map traces the limits of computation. A major advance reveals deep connections between the classes of problems that computers can—and [yet?] can't—possibly do. *Quanta Magazine* Retrieved from quantamagazine.org/20150929-edit-distance-computational-complexity.
Pelagios. (2011–Present). Pelagios commons. Retrieved January 15, 2016, from http://commons.pelagios.org.
Pessoa, Osvaldo Jr. (2001). Counterfactual histories: The beginning of quantum physics. *Philosophy of Science* 68, 519–530.
Popper, K. (1999). All life is problem solving. (P. Camiller, Trans.). New York: Routledge.
Popper, K. (1994). *Alles leben ist problemlösen*. München: Piper Verlag.
Progetto Cronache Veneziane e Ravennati. Fondazione Casa di Oriani. (2016). Retrieved January 15, 2016, from http://www.cronachevenezianeravennati.it.
Radick, G. (2005). Other Histories, Other Biologies. In Anthony O'Hear (ed.), *Philosophy, Biology, and Life*. Cambridge: Cambridge University Press, 21–47.
Reynolds, C. (1986). *Boids* [An artificial live program which simulates flocking birds in 2D]. Retrieved January 15, 2017, from http://www.red3d.com/cwr/boids.

Richthofen, F. V. (1876). Über den Seeverkehr nach und von China im Altertum und Mittelalter. *Verhandlungen der Gesellschaft für Erdkunde zu Berlin, 1876*, 86–97.
Richthofen, F. V. (1877). Über die zentralasiatischen Seidenstrassen bis zum 2. *Jh. n. Chr. Verhandlungen der Gesellschaft für Erdkunde zu Berlin, 1877*, 96–122.
Richthofen, F. V. (1877–1912). *China. Ergebnisse eigener Reisen und darauf gegründeter Studien* (Vol. 5). Berlin: Reimer.
Robertson, D. S. (1998). *The new renaissance: Computers and the next level of civilization*. Oxford: Oxford University Press.
Robertson, D. S. (2003). *Phase change: The computer revolution in science and mathematics*. Oxford: Oxford University Press.
Roland, B. (1972). *Critical essays* (R. Howard, Trans.). Evanston:Evanston Northwestern University Press.
Roland, B. (1967). Le discours de l'histoire. *Social Science Information, 6*(4), 63–75.
Sanudo, M. (1879–1902). In R. Fulin, F. Stefani, N. Barozzi, G. Berchet, & M. Allegri (Eds.), *Diarii* (Vol. 58). Venezia: Stabilimento Visentini cav. Federico Editore.
Schäfer, W. (2001). Global civilization and local cultures. A crude look at the whole. *International Sociology, 16*(3), 301–319 Also in Rethinking Civilizational Analysis, eds. Saïd Amir Arjomand and Edward A. Tiryakian, 71-86. London: SAGE Publications.
Schich, M., Song, C., Ahn, Y.-Y., Mirksky, A., Martino, M., Barabási, A.-L., & Helbing, D. (2014). A network framework of cultural history. *Science, 345*, 558–562. https://doi.org/10.1126/science.1240064.
Seshat: Global History Databank. (2011–Present). The Evolution Institute. Retrieved January 15, 2016, from http://seshatdatabank.info.
Shelling, T. C. (1971). Dynamic models of segregation. *Journal of Mathematical Sociology, 1*(2), 143–186.
Spiegel, G. M. 1997. *The Past as Text*. Baltimore and London: The Johns Hopkins University Press.
Toynbee, A. J. (1934–1961). *A study of history* (Vol. 12). London: Oxford University Press.
Trismegistos. (2004–Present). University of Leuven and University of Cologne. Retrieved January 15, 2016, from http://www.trismegistos.org.
Turchin, P. (2003). *Historical dynamics: Why states rise and fall*. Princeton, NJ: Princeton University Press.
Turchin, P., & Nefedov, S. A. (2009). *Secular cycles*. Princeton, NJ: Princeton University Press.
Vendrix, P. (1997). Cognitive sciences and historical sciences in music: Ways towards conciliation. In I. Deliège & J. Sloboda (Eds.), *Perception and cognition of music* (pp. 64–74). Hove, UK: Psychology Press.
Vlastos, G. (1983). The socratic elenchus. *Oxford Studies in Ancient Philosophy, 1*, 27–58.
Wang, G. (2016). Heritage and history. In: Third Singapore heritage science conference, Nanyang Technoogical University, Singapore, 25 January 2016. Retrieved January 6, 2017, from youtube.com/watch?v=-0wXSnqcAlM&list=PLasWJveXPWTE02EHZ2zsxzPxU6m3-GQli&index=3.
Waugh, D. C. (2007). Richthofen's "Silk Roads": Toward the archaeology of a concept'. *The Silk Road, 5*(1), 1–10.
Williams, B. (2006). *The sense of the past*. Princeton, NJ: Princeton University Press.
Wong, S. C., Miles, S., Fang, W., Groth, P., & Moreau, L. (2005). Provenance-based validation of E-science experiments. In: Y. Gil, M. Enrico, V. Richard Benjamins, & M. A. Musen (Eds.), The semantic web—ISWC 2005. Proceedings of the fourth international semantic web conference, Galway, Ireland, November 6–10, 2005 (pp 801–815). Berlin:Springer Verlag.
Woo, G. (2011). *Calculating catastrophe*. London: Imperial College Press.
Yule, H., & Cordier, H. (1913–1916). *Cathay and the way thither* (Vol. 4). London: The Hakluyt Society.

# Chapter 19
# A Posthumanist Reflection on the Digital Humanities and Social Sciences


**Chia-Rong Tsao**

## 19.1 Digital Forms of Life

The effects of information and digital technologies are frequently discussed. Lash (2002) states that we live in "the technological forms of life" and "we make sense of the world through technological systems" (p. 15). Assuming that the mode labeled as technological systems by Lash is currently referred to as "digital" (e.g., the Internet, mobile devices, and types of ubiquitous computing), we can argue that people typically perceive and comprehend their surroundings through digital technologies.

Digital technologies mediate not only daily life but also studies in the humanities and social sciences (Berry 2012). Evans and Rees (2012) indicate that "digital technology, as a concrete and pre-existing thing in the world, is unavoidably affecting the way humanities scholars conduct research (or think about the world)" (p. 36). The emergence of interdisciplinary studies in the digital humanities and social sciences is relevant to the development of digital technologies. Schnapp and Presner (2009) argue that the digital humanities explore a universe wherein "digital tools, techniques, and media have altered the production and dissemination of knowledge in the arts, human [*sic*] and social sciences" (p. 2).

Although the digital humanities were often seen as only a technical support to the "real" humanities studies in the early days (Berry 2012: 2), the definition of "digital" changed with the advent of the Internet in the 1990s. Burdick et al. (2012) define this change as an accelerated transition in digital scholarship, from processing to networking, and refer to it as "the first wave of the digital humanities". Hayles (2012) indicates that the first wave of the digital humanities has become "a

C.-R. Tsao (✉)
Department of Social Psychology, Shih Hsin University, Taipei, Taiwan



S.-H. Chen (ed.), *Big Data in Computational Social Science and Humanities*,
Computational Social Sciences, https://doi.org/10.1007/978-3-319-95465-3_19

genuinely intellectual endeavor with its own professional practices with its own professional practices, rigorous standards, and exciting theoretical explorations" (p. 43).

If the advent of the Web and digitalization technologies comprises the "first" wave, then what makes the "second" wave possible is the proliferation of "born digital data." According to Presner (2010), the primary practices of the second wave of the digital humanities are "producing, curating, and interacting with knowledge that is 'born digital' and lives in various digital contexts" (p. 6). In addition, Presner (2010) argues that the second wave introduced "entirely new discipline paradigms" (p. 6). Thus, researchers in the digital humanities and social sciences currently experience a fundamental transformation of epistemology, in addition to the introduction of new tools and methods.

Evans and Rees (2012) reveal that the main manifestation of this epistemological transformation is that "we move towards a situation where knowledge is co-produced from encounters between humans and machines" (p. 29). Frabetti (2012) argues, "A deep understanding of the mutual constitution of technology and the human is needed as an essential part of any work undertaken within the digital humanities" (p. 161). Therefore, this article investigates this situation from the "posthumanist" theoretical perspective and examines the human–technologies relationship as a type of symbiotic relationship.

The "posthuman" is a theoretical concept associated with Haraway's (1991) notion of the "cyborg." The "post" refers to the end of the modern human. Latour (1993) points out how the modern subject constructed and assured his central and exceptional status through the absolute division of nature and society. The appearance of the posthuman as a cybernetic organism (cyborg) signified a crisis, or even the end, of the modern subject (Gunkel 2000, p. 339) because it challenged the boundaries between the human and nonhuman, the subject and object, which were demarcated by modernity.

This article argues that posthumanist theories can elucidate how researchers and their digital tools coproduce knowledge. In other words, from the posthumanist perspective, this article points out that digital technologies inevitably affect current research practices and knowledge production and, more importantly, researchers also experience fundamental transformations in this coconstitution process.

## 19.2 "The Computational Turn": Expectations and Reflections

The most consequential result of the epistemological transformation of the digital humanities and social sciences is the development of "big data" as a major research method and data type. Anderson (2008, as cited in Boyd and Crawford 2012, p. 666) asserts, "This is a world where massive amounts of data and applied mathematics replace every other tool that might be brought to bear… With

enough data, the numbers speak for themselves." Hayles (2012) also indicates, "Perhaps the single most important issue in effecting transformation is scale" (p. 45). Furthermore, Kitchin (2014) contends, "Big Data analytics enables an entirely new epistemological approach for making sense of the world; rather than testing a theory by analyzing relevant data, new data analytics seek to gain insights 'born from the data'" (p. 2).

Big data analysis is popular among researchers in the digital humanities and social sciences because the quality of such data differs from that of data used in traditional humanities and social science studies. Manovich (2011) argues, "The emergence of social media in the middle of [*sic*] 2000s created opportunities to study social and cultural processes and dynamics in new ways. For the first time, we can follow imaginations, opinions, ideas, and feelings of hundreds of millions of people" (p. 2). In other words, "what is really distinctive about social media data is the production of naturally occurring or 'user-generated' data at the level of populations in real or near-real-time" (Edwards et al. 2013, p. 247).

Traditional social researchers must choose between two research methods: the first is that defined by Edwards et al. (2013) as "extensive and punctiform research which captures variation at the level of populations but only at specific moments and only in retrospect," and the second comprises "intensive research strategies that capture social processes but only in very specific social contexts or amongst particular social groups" (p. 249). Big data analysis by using social media data is a new method, which record the naturally occurring actions of users. Therefore, research on social processes is no longer limited to a specific context and can be conducted at varying population levels. Manovich (2011) asserts, "We no longer have to choose between data size and data depth. We can study exact trajectories formed by billions of cultural expressions, experiences, texts, and links" (p. 3).

Although big data analysis has the advantage of combining the size and depth, it has several limitations and problems. For example, Boyd and Crawford (2012) criticize several aspects of big data. First, the data cleaning process, which involves determining attributes and variables for analyses, is inherently subjective. In addition, Kitchin (2014) argues, "even if the process is automated, the algorithms used to process the data are imbued with particular values and contextualized within a particular scientific approach" (p. 5). Thus, the expectation of "letting the data speak for themselves" is not entirely objective.

Second, although big data analysis has "bigger" population-level data, Boyd and Crawford (2012, p. 669) contend that such data do not represent all people. We cannot equate one account to one user. Some users have multiple accounts, and some accounts are used by multiple people. In addition, we have "zombie accounts" that are operated by automated programs.

Third, because of the policies of social media companies, researchers can only access to a limited portion of data. Even if researchers acquire the required data, they are bound by ethics of social media researches. According to Boyd and Crawford (2012), "data may be public (or semi-public) but this does not simplistically equate with full permission being given for all uses" (p. 673).

In addition to the aforementioned concerns, Manovich (2011) contends, "We need to be careful of reading communications over social networks and digital footprints as 'authentic'" (p. 6). Edwards et al. (2013) illustrated, "Research into the riots of August 2011 . . . demonstrated the low fidelity of social media communications as representations of popular sentiment about the riots and their causes" (p. 247). In other words, neither the connection between accounts and users nor the authenticity and seriousness of the information spread on social media can be confirmed. Therefore, researchers question the notion that "data can speak for themselves."

The preceding reflections on big data analysis mostly focus on the problems in methods (or methodology), such as the representativeness, objectivity, and authenticity of data. However, an in-depth understanding of the changes in "knowledge" or "knowing" in the epistemological transformation of the digital humanities and social sciences is essential. To understand the altered interpretations of the "human," who is typically considered "the knowing subject," we must examine the fundamental changes caused by "the digital."

Berry (2012) explains the knowledge and knowing changes by introducing the concept of the "computational turn." He states, "I propose to look at the *digital* component of the digital humanities in the light of its medium specificity, as a way of thinking about how medial changes produce epistemic changes" (p. 4). Moreover, Boyd and Crawford (2012) contend, "Big Data creates a radical shift in how we think about research . . . it is a profound change at the levels of epistemology and ethics. Big Data reframes key questions about the constitution of knowledge, the processes of research, how we should engage with information, and the nature and the categorization of reality" (p. 665). In other words, the "computational turn" not only represents the challenges of new methods but also involves additional reflections on epistemological problems.

Thus, the computational turn depicts the increasing inclusion of computational devices in our daily lives; such devices have become an indispensable part of our knowledge of the world and our reliance on them is becoming taken for granted. According to Evans and Rees (2012), "we are comported towards the world computationally by the fact that the world is made up increasingly of computational things. As we use computational devices daily, new ways of interpreting and evaluating the world . . . become apparent to us through that usage" (p. 37). In other words, digital (computational) technologies affect the coconstitutive human–technology–world interrelationship. Therefore, users' acting and knowing are influenced by the transformation of augmentation and reduction, rather than by the operational functions of devices.

Thus, I assume that researchers in the digital humanities and social sciences rely on digital technologies in a radically decentered manner. Berry (2012) indicates that we, researchers, depend on digital technologies "to fill in the blanks in our minds and to connect knowledge in new ways" (p. 10). Furthermore, Hayles (2012) explains, "Human cognition is collaborating with machine cognition to extend its

scope, power, and flexibility" (p. 57). It is exactly in this decentered dependence and collaboration that human and digital (computational) technologies coconstitute each other.

Although Berry, Hayles, and other scholars have emphasized the need to investigate the epistemological transformation of the digital humanities and social sciences in the context of the computational turn, researchers have not attempted such investigation. I apply posthumanist theories to elucidate further the relationship between researchers and digital technologies and identify the problems and challenges in the relationship.

## 19.3 The Posthuman Condition: Double Decentralization

Before investigating the relationship between researchers and digital technologies, I explain the concept of the posthuman and how it can facilitate investigating the relationship.

Rather than simply criticizing the traditional humanist subject, posthumanist theories suggest new alternative subjects (Braidotti 2013). Considering these theories, Tsao (2016) contends that we can understand the possible image of the posthuman subject through the process of "double decentralization." He argues that the posthuman subject emerges when we stop regarding humans as the center of the world and the mind as the center of knowing and acting.

Latour (1993) asserts that the constitutional foundation of the modern society is the absolute separation between humans as subjects and the manipulated objects; modern people use purification to create two distinct ontological zones (human and nonhuman). In addition, they ensure the centrality and exceptionality of the human and the exploitability and disposability of the nonhuman. However, Latour (1993) argues that this ontological separation is a deception of modernity, stating that "the moderns think they have succeeded in such an expansion only because they have carefully separated Nature and Society (and bracketed God), whereas they have succeeded only because they have mixed together much greater masses of humans and nonhumans" (p. 41). Furthermore, it is evidenced by the proliferation of "hybrids" (e.g., the hole in the ozone layer and unstable nuclear power plants), which are increasingly difficult to classify.

The first decentralization in the "nonmodern world," proposed by Latour (1993), reveals that humans are no longer (or never) the superior and exceptional masters. Latour (1993) identifies the concealed mediation processes and indicates that both humans and nonhumans, subjects and objects, not only mediate the construction of social and natural facts but also coconstitute each other. In other words, nonhumans are not void, passive, or powerless. By contrast, Latour (1993) argues that the nonhuman as a mediator "creates what it translates as well as the entities between which it plays the mediating role" (p. 78). In addition, humans (subjects) are not the only actors or controllers. Latour (1993) asserts, "If, instead of attaching it to one constitutional pole or the other, we move it [human] closer to the middle, it

becomes the mediator and even the intersection of the two [human and nonhuman]" (p. 137). In other words, the absolute separation and opposition between humans and nonhumans no longer exist in Latour's nonmodern world.

However, Latour did not discuss the image of subjects or what he referred to as quasi-subjects. Instead of developing a posthuman theory, Latour focuses on objects and the "actor-network." Latour's reluctance to conduct a complete study on posthuman subjectivity might be attributed to his insistence on a "generalized symmetry principle." According to Ihde (2002), although Latour and other researchers (e.g., Callon 1986 and Law and Callon 1992) have extended the symmetry principle to humans and nonhumans, they have reduced the ontological variety in a united system of variables. In other words, when Latour moves both subjects and objects closer to the middle and suspends their ontological status, the exploration of changed images of subjects is also blocked. Researchers are compelled to analyze the interaction and association between quasi-subjects and quasi-objects in a network, rather than understand whether and how the quasi-subjects can be "alternative" subjects.

Therefore, Tsao (2016) contends that we require "the second decentralization" to explore further the image of posthuman subjects. The second decentralization has relevance to those theories that consider humans as conscious subjects. In other words, they focus on how to shift the subject out of the consciousness center. For example, the metaphorical concept "cyborgs" discussed by Haraway (1991) and Hayles (1999) elucidates that the boundaries between humans and technologies are increasingly blurred with the development of modern technologies, and their symbiotic relationship has also become more apparent.

More specifically, cyborgs as hybrids of humans and technologies challenge our understanding of subjects. Clark (2003) argues that humans are "natural-born cyborgs" because of the unique open nature of the brain, which enables the development of strong and complex relationships with nonbiological tools and technologies. In other words, contrary to the individualized subject, the natural-born cyborg, as a subject, is a hybrid of the brain, the body, and technologies.

Clark (2003) states, "We are, in short, in the grip of a seductive but quite untenable illusion: the illusion that the mechanisms of mind and self can ultimately unfold only on some privileged stage marked out by the good old-fashioned skin-bag," and he seeks "to dispel this illusion, and to show how a complex matrix of brain, body, and technology can actually constitute the problem-solving machine that we should properly identify as ourselves" (p. 27). In other words, Clark proposes that human knowing and acting operate because of the circuit of the brain, the body, and technologies, and not because of an individualized mind or self within a "skin-bag."

Clark (2008) terms this as "the extended mind thesis." He emphasizes that we must reject the notion that "all these various neural and nonneural tools need a type of privileged user" (Clark 2003, p. 136), which stems from traditional humanism theories that regard modern humans as individualized subjects. Clark (2003) asserts, "We, meaning we human individuals, just *are* these shifting coalitions of tools. We are 'soft selves,' continuously open to change and driven to leak through the

confines of skin and skull, annexing more and more nonbiological elements as aspects of the machinery of the mind itself" (p. 137).

Therefore, the extended mind thesis replaces the traditional conscious subject with the cognitive system constituted by hybrid elements including the brain, the body, and tools. The conscious subject was regarded as the actor with the exclusive privilege of knowing and acting. This replacement can consequently be considered the second decentralization of subjects, wherein neither humans have an exceptional and superior ontological status nor subjects are the "core" located in individuals' minds; rather, the subject emerges from the operation of the cognitive system constituted by hybrid elements.

Finally, I propose that the double decentralization highlights two crucial aspects of the changing relationship between humans and technologies. First, technologies are mediators of hybrid networks rather than simply tools and objects; they act jointly with humans. Second, technologies constitute the "extended cognitive system." They participate in the construction of knowing and acting through the "task-specific agent-world circuit" (Clark 2008, p. 16). Using these two aspects, the relationship between researchers and digital (computational) tools in the digital humanities and social sciences can be further explored.

## 19.4 Types of Researcher–Digital Tool Relationships

As previously discussed, researchers (e.g., Berry and Hayles) have emphasized the need for reflecting on the epistemological transformation of the digital humanities and social sciences. However, the reflections have not been systematically conducted. In this article, I apply posthumanist theories to argue that two types of researcher–digital tool relationships require reflection: The first is a relationship wherein researchers and digital tools are collaborators of hybrid networks; they work together and interact to construct knowledge and facts. The second is a relationship wherein researchers and tools coconstitute extended cognitive systems, implying that researchers as knowing subjects have already changed.

### *19.4.1 Collaborators of Hybrid Networks*

From the perspective of posthumanist theories, I oppose considering digital tools as simply tools; such consideration easily defines digital tools, such as all types of software, as passive objects, which are completely controlled by researchers. Consequently, we underestimate the contributions of digital tools in the processes of knowledge production.

To researchers in the digital humanities and social sciences, interpretation and analysis of data are nearly impossible without the help of "machines," because the "data" of the digital humanities and social sciences, regardless of whether they

are "natively digital" or "digitalized" (Rogers 2009), are so extensive that they cannot be processed by humans (or human labors). Hayles (2012) reveals that although humans are regarded as primary actors and analyzers, "if events occur at a magnitude far exceeding individual actors and far surpassing the ability of humans to absorb the relevant information, however, 'machine reading' might be a first pass towards making visible patterns that human reading could then interpret" (p. 47).

This type of reading is called "distant reading," which is contrary to the "close reading" practiced in the traditional humanities and social sciences. Hayles (2012, p. 48) asserts that unlike close reading, which is based on the "unaided human brain–body," distant reading combines humans interpreting and machines reading and, therefore, constitutes a considerably different type of knowledge formation. In other words, distant reading, as a new type of knowing, is "a *synthetic* activity that takes as its raw material the 'readings' of others" (Hayles 2012, p. 46). Therefore, from the perspective of posthumanist theories, machines, as collaborators, literally participate in the reading.

More specifically, these machines comprise the software that operates in the processes of data collection and analysis, such as text analysis tools, data-visualization tools, network analysis tools, and data-mining tools, designed for different social media sites. Berry (2011) asserts, "These devices are delegated particular behaviors and capabilities and become self-actualizing in the sense of realizing their potential by performing or prescribing complex algorithm-based action onto the world and onto us by acting to intervene in our everyday lives" (p. 125). In other words, they are Latour's nonhuman actors. Latour (1992) argues that nonhuman actors can act because of delegated, inscribed, or encoded scripts (p. 232), which constitute a type of program of action. Introna (2011) states that software codes can enact the intentions of designers because encoding extends the "agency."

Therefore, a research in the digital humanities and social sciences can be regarded as an actor-network wherein researchers collaborate with computational devices (Marres 2012). For example, when researchers investigate the posts, comments, shares, and likes of a Facebook event, they work and interact with several digital tools. The application programming interface (API) tool participates in the process of data scraping; subsequently, the word segmentation tool collaborates with the researcher to proceed with text analysis for different purposes; alternatively, the data-visualization tool presents the analyses results. All these tools, as collaborators, are not simply tools manipulated by researchers in the data collection and analysis process; rather, the API tools are delegated the action of "scraping," which entails gathering data from the social media site according to specific rules and conditions. The word segmentation proceeding, as well, is encoded action, thereby extending the agency of its designer; it divides sentences within a text (such as a social media post or digitalized text) into words according to a reference dictionary.

Berry (2011) reveals that the digital tool functions as a type of "double mediation" (p. 16). In other words, it not only translates texts to data but also translates the data to information presented to researchers. Therefore, such mediation not only "makes the user increasingly reliant on the screen image that the computer produces, but also renders them powerless to prevent the introduction of errors and mistakes"

(p. 16). Specifically, such mediation highlights the digital tool's capability to act in transforming facts and knowledge. However, the tool, as a delegated actor, can betray the designer or the user. For example, the API tools may fail to satisfy the user's expectation because of some neglected rules or incorrect codes.

Berry (2011) describes the mediation as a "vicarious relationship," wherein "following the *command* [emphasis added] (order-words), the user transacts with the code to execute the action" (p. 137). However, this relationship must not be understood from the viewpoint of traditional humanism: when the command fails to control, we adopt alternate methods to regain control. For example, the user may believe that if the hidden software codes are identified, the command is more rigid and effective. However, the posthumanist perspective does not regard digital tools as only tools or objects that we can firmly control. Rather, they are actors, similar to humans, in the hybrid network; we must collaborate with them to produce knowledge.

Similarly, Berry (2011) argues that knowing in a vicarious relationship is "taking place within other actors or combinations and networks of actors (which may be human or non-human)" (p. 13). In addition, "these objects are themselves able to exert *agential* [emphasis added] features, either by making calculations and decisions themselves, or by providing communicative support for the user" (Berry 2011, p. 13) Therefore, instead of controlling, we must understand the activity (or inactivity) of digital tools, such an understanding may highlight the limitations to and deficiencies in the knowledge of the digital humanities and social sciences, and possibly provide opportunities for additional improvement.

### *19.4.2 Coconstitutors of Extended Cognitive Systems*

Reflecting on the epistemological transformation of the digital humanities and social sciences elucidates the collaborative relationship between researchers and digital tools and recognizes their coconstitutive relationship. This recognition explains the fundamental changes that occur.

According to the posthumanist perspective, both the researcher and digital tool constitute an "extended cognitive system," whose operation produces knowledge of the digital humanities and social sciences. Coconstituting the extended cognitive system implies that the digital tool inevitably affects the researcher. As previously stated, the "subject," or the self, of the posthuman emerges from the operation of the extended cognitive system. Therefore, the researcher experiences a transformation of self-image (or the image of the subject), when the elements of the system and the operating mode change.

In other words, the activities of digital tools discussed previously, such as scraping of API tools, are not only types of collaborations, but also parts of the operations of the extended cognitive system. Moreover, the subject is constituted in the system, rather than in the researcher's brain. Therefore, during the extension of the circuit of knowing and acting, these tools augment the subject's abilities.

For example, the researchers cannot read, or even access, the numerous and varying posts and comments of a Facebook event without the help of the API tools. However, this augmentation has inevitable consequences. Agreeing with Ihde (1990), I argue that every digital tool has its "magnification–reduction" structure, and during incorporation into the extended cognitive system, digital tools not only augment researchers' abilities but also change how they perceive and act.

Ihde (1979) illustrates how the "magnification–reduction" structure works; he states that the telephone, as an electronic medium, enables long-distance conversations; its magnification is apparent because "whether you are in the next room, the next state, or even in the next country, my hearing is 'extended' to you through the phone" (p. 24). However, the telephone is a "mono-sensory" device that reduces everything to a voice. Thus, the telephone reduces not only the multiple sensory dimensions of a conversation but also the conversation itself. Ihde (1979) asserts, "This is to say that this particular technology is inclined to purveying 'information.' But this, too, is a kind of reduction in the sense that 'information' is only one dimension of the total richness of human experience" (p. 26). Therefore, the telephone's incorporation into the extended cognitive system has changed our sense and practices of conversations and augmented our ability to establish long-distance communication.

Similarly, a certain type of magnification–reduction structure is observed in digital tools. For example, when the API tool operates as a part of the extended cognitive system, it augments the researcher's data-collecting ability. This system, which comprises the API tool and researcher, can scrape and collect all the posts, comments, and the number of likes and shares of an event in a specific period. Therefore, computable dimensions of phenomena are magnified. According to Berry (2012), "A computer requires that everything is transformed from continuous flow of everyday life into a grid of numbers that can be stored as a representation which can then be manipulated using algorithms" (p. 2). Although the API tool can collect the content of a post or comment, the frequency of a word remains the primary concern. This means that the meaning and motivation of a user's activities are relatively reduced.

The content of a post or comment and the likes or shares may have multiple meanings. For example, the implication of an ironic sentence, which is often used to criticize, is the opposite of the literal meaning. We cannot understand this by only counting the number and frequency of words. Moreover, we cannot ensure that the simple acts of likes and shares always represent a user's agreement. In addition, the motivation for a user's activities on a social media site cannot be recorded and counted.

The system comprising the API tool and researcher is concerned with the number of a user's activities and words and accordingly identifying the hidden relationships and modes. Ruppert et al. (2013) state, "Here, the focus of inquiry is not on the individual factors that affect behavior, but on the spatial flows of behaviors and contacts: contagion, pollution, influence, etc." (p. 35). They argue that "to this extent, humanist conceptions of society are being eclipsed" (p. 36).

Therefore, changes in researchers can be understood only by reflecting on this coconstitution relationship. Ihde suggests that the magnification–reduction structure of a tool causes two effects: the dramatic standing out of the magnified dimension and the neglect, or even forgetting of the reduced dimension. Thus, I argue that while the digital tool, as a part of the extended cognitive system, magnifies and reduces the *world*, the researcher, who coconstitutes the same system, experiences corresponding changes.

## 19.5 Conclusion

In this article, I reflect on the fundamental and epistemological changes in the digital humanities and social sciences. By applying posthumanist theories, I argue that the relationship between researchers and digital tools can be examined from two perspectives. First, the researcher and the digital tool are regarded as collaborators of a hybrid network, which produces knowledge. Second, they not only collaborate but also coconstitute each other; thus, the digital tool changes the researcher when it magnifies and reduces the different dimensions of the world.

Evans and Rees (2012) argue, "Perhaps the new digital humanities face the same problem—by being too caught up in the excitement of new analytical tools, are we in danger of losing sight of the bigger picture of the humanities and its role in questioning and revealing the human condition?" (p. 29). Corresponding with their assertions, this article highlights the *danger* in the development of the digital humanities and social sciences. I indicate that the digital tool is not only a passive object manipulated by the researcher; it is also a delegated actor that extends the agency of its user, but is capable of betraying users because of neglected rules or incorrect codes. Moreover, Introna (2011) states, "In our continual pursuit of convenience and efficiency we 'delegate' to digitally encoded actors the most intimate details of our lives, and, in doing so, we conveniently forget and lose track of these encodings" (p. 114). Therefore, we may lose the clear comprehension of the delegation or even *forget* the delegation; the notion of "let the data speak for themselves" may represent a symptom of such loss.

This article reiterates that researchers and their digital tools coconstitute each other and are parts of the extended cognitive system. I explain how the digital tool changes researchers through its magnification–reduction structure. I argue that when the digital tool augments researchers' ability by magnifying one dimension of the world, it also reduces and eclipses the other dimensions. Evans and Rees (2012) also express the similar concern, "the importance of human hermeneutic interpretation potentially diminishes" (p.21).

Ruppert et al. (2013, p. 32) assert, "We need to get our hands dirty and explore their affordances" (p. 32). This article provides one possible approach to using posthumanist theories to understand the changes caused by digital tools. Future studies must examine the actual situations of digital humanities and social science researches and investigate the two types of relationships suggested in this article.

## References


Berry, D. M. (2011). *The philosophy of software: Code and mediation in the digital age*. New York: Palgrave Macmillan.

Berry, D. M. (2012). Introduction: Understanding the digital humanities. In D. M. Berry (Ed.), *Understanding digital humanities* (pp. 1–20). New York: Palgrave Macmillan.

Boyd, D., & Crawford, K. (2012). Critical questions for big data: Provocations for a cultural, technological, and scholarly phenomenon. *Information, Communication & Society, 15*(5), 662–679.

Braidotti, R. (2013). *The Posthuman*. Cambridge: Polity Press.

Burdick, A., Drucker, J., Lunenfeld, P., Presner, T., & Schnapp, J. (2012). *Digital humanites*. Cambridge: MIT Press.

Callon, M. (1986). Some elements of a sociology of translation: Domestication of the scallops and the fishermen of St Brieuc Bay. In J. Law (Ed.), *Power, action and belief: A new sociology of knowledge* (pp. 196–233). London: Routledge.

Clark, A. (2003). *Natural-born cyborgs: Minds, technologies, and the future of human intelligence*. Oxford: Oxford University Press.

Clark, A. (2008). *Supersizing the mind: Embodiment, action, and cognition extension*. Oxford: Oxford University Press.

Edwards, A., Housley, W., Williams, M., Sloan, L., & Williams, M. (2013). Digital social research, social media and the sociological imagination: Surrogacy, augmentation and re-orientation. *International Journal of Social Research Methodology, 16*(3), 245–260.

Evans, L., & Rees, S. (2012). An interpretation of digital humanities. In D. M. Berry (Ed.), *Understanding digital humanities* (pp. 21–41). New York: Palgrave Macmillan.

Frabetti, F. (2012). Have the humanities always been digital? For an understanding of 'digital humanities' in the context of originary technicity. In D. M. Berry (Ed.), *Understanding digital humanities* (pp. 161–171). New York: Palgrave Macmillan.

Gunkel, D. (2000). We are Borg: Cyborgs and the subject of communication. *Communication Theory, 10*(3), 332–357.

Haraway, D. J. (1991). *Simians, cyborgs, and wormen: The reinvention of nature*. New York: Routledge.

Hayles, N. K. (1999). *How we became posthuman: Virtual bodies in cybernetics, literature, and informatics*. Chicago: The University of Chicago Press.

Hayles, N. K. (2012). How we think: Transforming power and digital technologies. In D. M. Berry (Ed.), *Understanding digital humanities* (pp. 42–66). New York: Palgrave Macmillan.

Ihde, D. (1979). *Technics and praxis*. Boston: D. Reidel Pub. Co.

Ihde, D. (1990). *Technology and the lifeworld*. Bloomington: Indiana University Press.

Ihde, D. (2002). *Bodies in technology*. Minneapolis: The University of Minnesota Press.

Introna, L. D. (2011). The enframing of code: Agency, originality and the plagiarist. *Theory, Culture & Society, 28*(6), 113–141.

Kitchin, R. (2014). Bid Data, new epistemologies and paradigm shifts. *Big Data & Society, 1*, 1–12.

Lash, S. (2002). *Critique of information*. London: Sage.

Latour, B. (1992). Where are the missing masses? The sociology of a few mundane artifacts. In W. E. Bijker & J. Law (Eds.), *Shaping technology/building society: Studeies in sociotechnical change* (pp. 225–258). Cambridge: MIT Press.

Latour, B. (1993). *We have never been modern*. New York: Harvard University Press.

Law, J., & Callon, M. (1992). The life and death of an aircraft: A network analysis of technical change. In W. E. Bijker & J. Law (Eds.), *Shaping technology/building society: Studeies in sociotechnical change* (pp. 21–52). Cambridge: MIT Press.

Manovich, L. (2011). Trending: The promises and the challenges of big social data. In M. K. Gold (Ed.), *Debates in the digital humanities*. Minneapolis: The University of Minnesota Press Retrieved from http://dhdebates.gc.cuny.edu/debates/text/15.

Marres, N. (2012). The redistribution of methods: On intervention on digigtal social research, broadly conceived. *The Sociological Review, 60*(S1), 139–165.

Presner, Todd. (2010). Digital humanities 2.0: A report on knowledge. http://cnx.org/contents/J0K7N3xH@6/Digital-Humanities-20-A-Report

Rogers, R. (2009). *The end of the virtual: Digital methods*. Amsterdam: Vossiuspers UvA.

Ruppert, E., Law, J., & Savage, M. (2013). Reassembling social science methods: The challenge of digital devices. *Theory, Culture & Society, 30*(4), 22–46.

Schnapp, Jeffrey, & Presner, Todd. (2009). Digital humanities manifesto 2.0. http://www.humanitiesblast.com/manifesto/Manifesto_V2.pdf

Tsao, C.-R. (2016). The hybrid subject: The "post-human" in philosophy of technology. *A Journal for Philosophy Study of Public Affairs, 57*, 47–93

# Index





S.-H. Chen (ed.), *Big Data in Computational Social Science and Humanities*,
Computational Social Sciences, https://doi.org/10.1007/978-3-319-95465-3

**D**