Towards the Web in Your Pocket: Curated Data as a Service

Stuart Dillon[1], Florian Stahl[2], and Gottfried Vossen[2,1]

[1] University of Waikato Management School, Hamilton, New Zealand
stuart@waikato.ac.nz
[2] ERCIS, University of Münster, Münster, Germany
{florian.stahl,gottfried.vossen}@ercis.de

Abstract. The Web has grown tremendously over the past two decades, as have the information needs of its users. The traditional "interface" between the vast data resources of the Web and its users is the search engine. However, search engines are increasingly challenged in providing the information needed for a particular context or application in a comprehensive, concise, and timely manner. To overcome this, we present a framework that does not just answer queries based on a pre-assembled index, but based on a subject-specific database that is curated by domain experts and dynamically generated based on vast user input.

Keywords: Data curation, digital curation, curation process, Data as a Service, Web in the Pocket, information provisioning.

1 Introduction

Over the past 20 years, the volume of data stored electronically has become superabundant and the Web is said to be the biggest collection of data ever created [3]. Moreover, accessible data on the Web, whether created by computers, by users, or generated within professional organizations, continue to grow at a tremendous pace [19,22,28]. Social networks like Facebook, search engines like Google, or e-commerce sites like Amazon, generate and store new data in the TB range every day. Due to the emerging usage of cloud computing [2], this trend will not only continue, but accelerate over the coming years, as more data is permanently stored online, is linked to other data, and is aggregated in order to form new data [1]. As a consequence, the question arises as to how to best source the precise information required[1] at a given moment or in a particular context, be it for personal or professional use. The concept of a "Web in the Pocket" proposed in this paper is a possible solution.

For the past 15 years, "search" has been the tool of choice for garnering Web information; the underlying data is extracted using a search engine [9,8,24,7]. But while (dynamic, on-demand) search is preferable to (static) directories in

[1] We here use *data* and *information* interchangeably, although we consider information as the result of extracting "usable" data by a user.

N.T. Nguyen et al. (Eds.): *Adv. Methods for Comput. Collective Intelligence*, SCI 457, pp. 25–34.
DOI: 10.1007/978-3-642-34300-1_3

the face of an exploding Web, the point has been reached where search, even though nowadays available in many facets, may no longer be appropriate in many situations. Having data relevant to specific areas or applications selected, quality-checked and integrated ("curated") in advance as well as made available as a configurable, adaptable, and updatable service can provide a competitive advantage in many situations [25,15,33,26]. While recent data marketplaces [32] go a step in this direction at least for specific applications, the idea presented in this paper is to have access to a service that has been configured precisely to a user's needs and budget and that composes and delivers corresponding data in a topic-centric way.

We envision that this data service can even be made available offline on both stationary and mobile devices, the latter resulting in a "Web in your pocket" (WiPo). A simple way to view the WiPo idea is to think of complete context- or application-specific data being "pushed" to the user when required; this data has been collected according to user input, and has then been processed in various ways (explained below) to ensure completeness, accuracy, quality, and up-to-dateness. This is in contrast to the current search model, where information is dynamically "pulled" by the user, often only in rudimentary ways. A further key feature of WiPo is that it will generate data from a range of public and private sources. The way to achieve quality data is to exploit curation, where data curation refers to the long-term selection, cleansing, enrichment, preservation and retention of data and to provide access to them in a variety of forms. It is known, for example, from museums and has already been successfully applied to scientific data (amongst other by the British Digital Curation Centre, see www.dcc.ac.uk). An application built on curated data can help solve the identified information problems, as raw data will become less useful because of its availability in sheer quantities. We consider WiPo to be a data centric application that provides a comprehensive overview and detailed information on a given topic area through a configurable service.

As will be seen, WiPo can be useful to both business (i.e., professional) and non-business (i.e., private) users. Since the information needs of these two groups are likely to be different, we will treat them as separate cases. Further, information may be required at different frequencies and at differing levels of quality. For instance, stock prices are data that are usually needed with a relatively high frequency and quality, whereas other information may be required less frequently or where quality is less important. Quality (reliability) of information is normally increased as the number of data sources employed increases. We term this "broadness" (a term suggested to us by Jim Hendler). While both frequency and broadness values are likely to be continuous (i.e., on a continuum), for simplicity we take a bimodal approach. As shown in Figure 1, there are four possible classifications for each of the business and non-business scenarios, represented as a four-field matrix. In the following, we present sample use cases for all eight possible scenarios resulting from the four-field matrix and the two application areas of *businesses* and *non-business*.

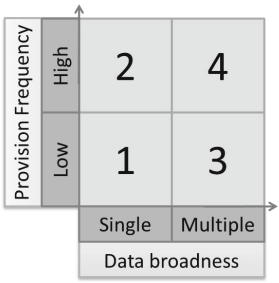

Fig. 1. Classification Matrix

Business Cases

1. Low frequency of data provision (one-off), low data broadness (single source): A company wishes to determine the cost of airline lounge membership for its frequently flying CEO.
2. High frequency of data provision, low data broadness: A manufacturing firm wishes to keep abreast with the cost of steel on the spot market. One (reliable) source is satisfactory; however, the data needs to be provided on a continuous basis.
3. Low frequency of data provision, high data broadness (multiple sources): A company is considering changing its sole provider of office supplies. It has already been through a tendering process, but wishes to source external information about providers.
4. High frequency of data provision, low data broadness: It is essential that firms have a good understanding of the marketplace(s) in which they operate. They need to know what others in the supply chain are doing and what their customers are doing. They need to know about external (e.g., political or legal) events that may change operating conditions. This information might come from a variety of sources including official indexes, market reports, business commentators, as well as internal data sources for comparison and benchmarking, and has to be up to date.

Non-business Cases

1. Low frequency of data provision (one-off), low data broadness (single source): An air traveler is searching for a one-night layover in an arbitrary city. They wish to know "What is the cheapest hotel within 5 km of the international terminal?"
2. High frequency of data provision, low data broadness: A small-time investor has a number of shares and stocks. He/she wants to be able to view the current share price of the shares that they own on demand.
3. Low frequency of data provision, high data broadness (multiple sources): A hospital patient has recently been diagnosed with a genetic condition that others in his or her extended family have also suffered from. The patient has

some information (from the hospital and their family) about the condition, but also wishes to gain a better understanding of alternative treatments and of local support groups.

4. High frequency of data provision, low data broadness: Many people regularly bet on sporting events, e.g., football games. Gamblers can bet on a range of things such as results, first to score, etc. Here, the more up-to-date information one can obtain, the more successful one can be. These users will need extensive statistics over the past few seasons, weather forecasts for games, team and player news (e.g., injuries) to help them improve their betting returns.

We will further explore the case of the hospital patient later as an example of how WiPo is intended to work. The rest of the paper is structured as follows: Section 2 outlines related work. Section 3 elaborates on how to provide curated data, starting with an overview, then highlighting the speciality of user input, and finally discussing the process of curation. Future work is outlined in Section 4.

2 Related Work

The Web in the Pocket is, in a sense comparable, to a materialized data warehouse [20] that is made portable, and to Fusion Cubes as suggested in [1] for situational data. More related to our work is research on search engines. Cafarella et al. [11] examine how search engines can index pages in the deep Web. Building data marts or services on given deep Web sources is described by Baumgartner et al. [4]. Generally speaking, however, retrieval and indexing of documents is no longer the problem it used to be, though far from being solved in its entirety [3].

There are several issues to address in order to fully satisfy current and future information needs. Dopicha [16] found that it was technically impossible to meaningfully answer queries such as "return all pages that contain product evaluations of fridges by European users". A similar conclusion was reached by Ceri [13]. To solve this, Dopichaj advocates the Semantic Web, whose idea is to enhance text on the Web by semantic information to make it machine understandable. Ceri [13], on the other hand, proposed the so called *Search Computing* (SeCo) framework. A detailed description of an architecture for SeCo is described in [5]. Briefly, a query to a SeCo search engine is processed by a query optimizer that determines suitable search services to which it then sends subqueries. Campi et al. [12] describe how service marts (e.g., as suggested in [4]) can be built and registered with the framework. Search service results are then joined and displayed to users. In a subsequent step, users have the opportunity to modify their queries. This is referred to as *liquid* query processing [6]. In order to realize this framework, two new user groups need to be created: providers of data offering data as a services and developers building search services based on data services.

One of the few approaches combining curation and Web Information Retrieval (IR) has been proposed by Sanderson et al. [30] who suggest (focusing on a single domain) a process similar to life science data consolidation as described in

[21]. In life science, for instance, different institutions contribute to a nucleotide sequence databases; to ensure consistency data are exchanged on a nightly bases [21]. Sanderson et al. [30] apply this procedure to content retrieval by sending predefined nightly queries to pre-registered services. Relevant information is harvested and temporarily stored. The information retrieved is then audited by data curators who decide upon its relevance; only relevant data is kept. Moreover, care is taken to prevent harvesting the same data more than once. A similar harvest-and-curate approach was suggested by Lee et al. [23] who outline ideas to enhance their ContextMiner (see `contextminer.org` for a tool offering contextual information to data) by making it scalable. However, beyond these sources there is apparently no in-depth research on combining data curation and IR, and even those do not appear to have the potential to serve information needs on a large scale in a sophisticated way. In any data curation process humans are needed at some point, which is why the education and qualification of data curators is also intensively discussed in the literature [18,31,27].

3 The WiPo Approach

3.1 Overview

The overall process view of the WiPo approach is generic and applicable to many use cases. The concrete application design, however, is use-case dependent, i.e., the individual steps of the process have to be implemented considering the purpose of the application and the specific needs of the user. Figure 2 presents the basic process underlying WiPo in the form of a Petri net [34], in which activities are depicted as rectangles and states are depicted as circles; the important rule is that states and activities must strictly alternate. A state (of the underlying data) can be thought of as objects from a database, a document containing certain information, or of collections of such items (e.g., a list of selected sources after the Source Selection activity). A double lined box indicates that this particular activity has a refinement, i.e., is composed of other activities at a lower level of abstraction. At the high level shown in Figure 2 the overall process consist of steps *Input Specification*, resulting in a list of potential sources and pre-filters; *Source Selection*, resulting in a list of sources; *Data Mining*, resulting in raw data; *Past Filters*, reducing the raw data to "relevant" raw data; and *Curation*. These sub-processes will be explained in more detail in the following natural language description.

The first step is receiving user input. In this step it is specified what the data service should be looking for. An input specification usually is a list of links, keywords, or documents, or a combination thereof. Due to the importance of this step, Section 3.2 elaborates further on input specification and presents a refinement of this step. In the case of the hospital patient, the input could be bookmarks of medical Web sites, documents describing cures and treatments, or clinical trial results. After this pre-processing step the most appropriate sources are determined based on standard algorithms and heuristics (e.g., [3,7]). This explicitly includes discovery of more recent documents of the same class as the

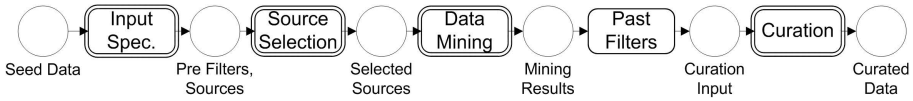

Fig. 2. WiPo process overview as a Petri net

ones provided, e.g., more recent medical publications in our sample case. Web sources can be categorized in two ways: A source may be either structured or unstructured. Concerning access to it, it may be either publicly available, semi-public (i.e., only available to subscribers), or private (i.e. only available to a given user). The task of selecting sources also entails the setting of pre-filters (e.g., filters of time if only recent documents are relevant). Furthermore, it is strongly connected to the action of meta-data management which focuses on storing information about the various sources.

The next step is data mining or information retrieval. Using techniques from these well-established fields (e.g., [36], relevant information is extracted from the predefined sources. Depending on the degree of structure the sources provide, the complexity of this process can range from very simple to extremely complex. Following data mining, a post-filter can be applied in order to reduce the potentially vast amount of mining results to a more manageable size, containing only most relevant information. Considering our patient example, mining could be restricted to a medical journal database as well as a collection of sites such as `uptodate.com`; a pre-filters would be time (since only recent documents are of interest); a post-filter can limit the results to the patient's specific condition.

The mining process is followed by curation, which we define as *making data fit for a purpose*. To ensure the "fitness" of the data, a domain expert will be involved in this step, which can depend on the particular knowledge domain, the professionalism of the data service provider, or even on the price the user is willing to pay for quality information. For the former option, a medical doctor as domain expert could ensure that the content of the received documents is factually true, and design a process to integrate various different medical sources in a timely manner. Curation itself consists of four subtasks: a) imposing data quality by finding trustworthy sources; b) adding value; c) adding data lineage information and d) integration of different data sources, including additional sources and files given by the user, as well as tasks such as data cleansing and data fusion. Fusion also includes joining mined results with the initial information provided by the user. When curation is completed, the data will be written to the curated database from where the data service collects, collates and presents them — appropriately visualized by the WiPo application — to the user. The entire setup just described is presented in Figure 3 starting from User Input. Thin arrows indicate a relationship between two items and block arrows represent the flow of data within the process.

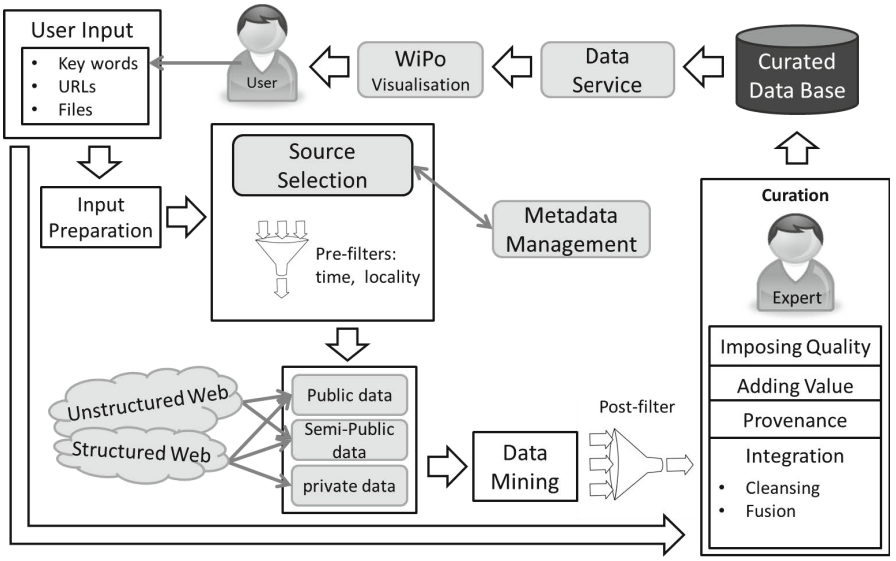

Fig. 3. WiPo Architecture

3.2 Incorporating User Input

In most of the established search engines such as Google or Bing, keywords and search strings, maybe even regular expressions or SQL-like queries, are used to interpret what a user is searching for. Those are then matched against an index of pre-crawled Websites to retrieve relevant documents [35]. Most of them allow the search to be restricted to a specific domain or top-level domain; others can be given a predefined list of URLs and restrict search to that list. The "history" of search engines has seen numerous approaches to capture a user's intuition. However, there is — to the best of our knowledge — no approach incorporating personal, offline data with externally, Web-sourced data.

In practical terms, using a suitable interface, the user can upload documents, supply a list of relevant links and potentially also keywords. WiPo will provide a dedicated language interface for this purpose, for which we envision usage of an XML standard along the lines of DITA (Darwin Information Typing Architecture). Input Analysis will then convert the given documents to a homogeneous file format in order to apply a generic classifier. The given URLs will be crawled and also fed into the classifier. In that way from both the given documents and URLs the essence is extracted in the form of topics or additional keywords. These mined keywords as well as the user-supplied keywords are then used to choose appropriate sources from a list of all available sources (which might also be extended based on user supplied URLs). Furthermore, relevant other dimensions such as time or location are determined to generate a list of sources including pre-filters. This process is shown in Figure 4.

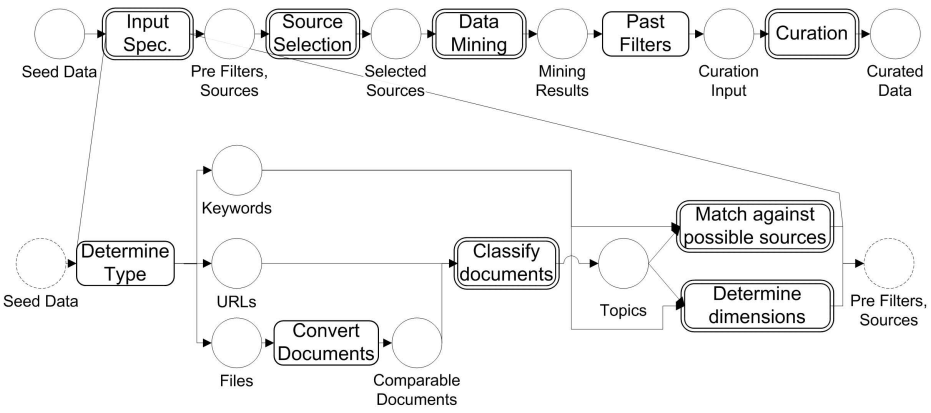

Fig. 4. Detailed view on Input Specification

3.3 Bringing Data Curation to IR

A research stream that thus far has only received little attention in the area of Web IR is data curation. The original idea of "digital" curation was first discussed in the library and information sciences, where it still vastly resides. It is focused on huge scientific data sets of physical, biological, or astronomical data with the aim of preserving data gained through scientific experiments for later usage [14,27]. To achieve this goal the idea is to clean and update data to new technical standards in order to make them accessible for future use [17,25,29]. In library science, electronic repositories storing data and information are referred to as institutional repositories [31,14]. Research conducted in that domain can also be beneficial when developing the central data repository as basis of a new Web search utility. Also relevant in this context is data *provenance* brought into the discussion of data curation by (amongst others) Buneman et al. [10], who are highly involved in the British Digital Curation Centre. Data provenance or lineage refers to the action of making the history of some piece of data available in a way that its origin can be traced and modifications reconstructed. This is in particular important w.r.t. crediting the right people when using their data (compare the action of referencing works by other authors). Also knowing the origin of a source can arguably increase trust and support reproducibility.

4 Conclusions and Future Work

We have introduced a framework that combines established methods from the fields of information retrieval, Web search and library sciences. These combined together offer a unique approach which we term *Web in the Pocket* (WiPo). We believe this offers an opportunity for a revolutionary new approach that will, in time, supersede the current, and somewhat limited, search-based approach for information retrieval. This paper has focused on the presentation of an overall conceptual model for WiPo as well as a basic architecture. Two aspects of

WiPo require further investigation and clarification, namely curation and input specification. In Section 3.2 we showed how we envision input specification; the consequential next step is to work on a concrete implementation. In 3.3 we briefly summarized the concept of curation, and this also needs a more detailed exploration within the context of the WiPo approach.

References

1. Abello, A., et al.: Fusion Cubes: Towards Self-Service Business Intelligence. To Appear in Journal on Data Semantics (2013)
2. Armbrust, M., et al.: A view of cloud computing. CACM 53(4), 50–58 (2010)
3. Baeza-Yates, R., Raghavan, P.: Chapter 2: Next Generation Web Search. In: Ceri, S., Brambilla, M. (eds.) Search Computing. LNCS, vol. 5950, pp. 11–23. Springer, Heidelberg (2010)
4. Baumgartner, R., Campi, A., Gottlob, G., Herzog, M.: Chapter 6: Web Data Extraction for Service Creation. In: Ceri, S., Brambilla, M. (eds.) Search Computing. LNCS, vol. 5950, pp. 94–113. Springer, Heidelberg (2010)
5. Bozzon, A., Brambilla, M., Ceri, S., Corcoglioniti, F., Gatti, N.: Chapter 14: Building Search Computing Applications. In: Ceri, S., Brambilla, M. (eds.) Search Computing. LNCS, vol. 5950, pp. 268–290. Springer, Heidelberg (2010)
6. Bozzon, A., Brambilla, M., Ceri, S., Fraternali, P., Manolescu, I.: Chapter 13: Liquid Queries and Liquid Results in Search Computing. In: Ceri, S., Brambilla, M. (eds.) Search Computing. LNCS, vol. 5950, pp. 244–267. Springer, Heidelberg (2010)
7. Bozzon, A., Brambilla, M., Ceri, S., Fraternali, P., Vadacca, S.: Exploratory search in multi-domain information spaces with liquid query. In: Proc. 20th Int. Conf. on World Wide Web, pp. 189–192. ACM, New York (2011)
8. Braga, D., Corcoglioniti, F., Grossniklaus, M., Vadacca, S.: Panta Rhei: Optimized and Ranked Data Processing over Heterogeneous Sources. In: Maglio, P.P., Weske, M., Yang, J., Fantinato, M. (eds.) ICSOC 2010. LNCS, vol. 6470, pp. 715–716. Springer, Heidelberg (2010)
9. Brin, S., Page, L.: The anatomy of a large-scale hypertextual web search engine. Computer Networks 30, 107–117 (1998)
10. Buneman, P., Chapman, A., Cheney, J., Vansummeren, S.: A Provenance Model for Manually Curated Data. In: Moreau, L., Foster, I. (eds.) IPAW 2006. LNCS, vol. 4145, pp. 162–170. Springer, Heidelberg (2006)
11. Cafarella, M.J., Halevy, A., Madhavan, J.: Structured Data on the Web. CACM 54(2), 72–79 (2011)
12. Campi, A., Ceri, S., Gottlob, G., Maesani, A., Ronchi, S.: Chapter 9: Service Marts. In: Ceri, S., Brambilla, M. (eds.) Search Computing. LNCS, vol. 5950, pp. 163–187. Springer, Heidelberg (2010)
13. Ceri, S.: Chapter 1: Search Computing. In: Ceri, S., Brambilla, M. (eds.) Search Computing. LNCS, vol. 5950, pp. 3–10. Springer, Heidelberg (2010)
14. Choudhury, G.S.: Case Study in Data Curation at Johns Hopkins University. Library Trends 57(2), 211–220 (2008)
15. Doorn, P., Tjalsma, H.: Introduction: archiving research data. Archival Science 7, 1–20 (2007)
16. Dopichaj, P.: Ranking-Verfahren für Web-Suchmaschinen. In: Lewandowski, D. (ed.) Handbuch Internet-Suchmaschinen. Nutzerorientierung in Wissenschaft und Praxis, pp. 101–115. AKA, Akad. Verl.-Ges., Heidelberg (2009)

17. Gray, J., Szalay, A.S., Thakar, A.R., Stoughton, C., van den Berg, J.: Online Scientific Data Curation, Publication, and Archiving. CoRR Computer Science Digital Library cs.DL/0208012 (2002)
18. Heidorn, P.B., Tobbo, H.R., Choudhury, G.S., Greer, C., Marciano, R.: Identifying best practices and skills for workforce development in data curation. Proc. American Society for Information Science and Technology 44(1), 1–3 (2007)
19. Hey, T., Trefethen, A.: The data deluge: An e-science perspective. In: Berman, F., Fox, G.C., Hey, A.J. (eds.) Grid Computing — Making the Global Infrastructure a Reality, pp. 809–824. Wiley (2003)
20. Inmon, W.: Building the Data Warehouse. Wiley Technology Publishing, Wiley (2005)
21. Kulikova, T., et al.: The embl nucleotide sequence database. Nucleic Acids Research 32(suppl. 1), 27–30 (2004)
22. Laudon, K., Traver, C.G.: E-commerce: business, technology, society, 9th edn. Pearson/Prentice Hall (2013)
23. Lee, C.A., Marciano, R., Hou, C.Y., Shah, C.: From harvesting to cultivating: transformation of a web collecting system into a robust curation environment. In: Proc. 9th ACM/IEEE-CS Joint Conference on Digital Libraries, pp. 423–424. ACM, New York (2009)
24. Levene, M.: An Introduction to Search Engines and Web Navigation, 2nd edn. Wiley (2010)
25. Lord, P., Macdonald, A., Lyon, L., Giaretta, D.: From data deluge to data curation. In: Proc. UK e-Science All Hands Meeting, pp. 371–375 (2006)
26. Meliou, A., Gatterbauer, W., Halpern, J.Y., Koch, C., Moore, K.F., Suciu, D.: Causality in databases. IEEE Data Eng. Bull. 33(3), 59–67 (2010)
27. Palmer, C.L., Allard, S., Marlino, M.: Data curation education in research centers. In: Proc. 2011 ACM iConference, pp. 738–740. ACM, New York (2011)
28. Ramírez, M.L.: Whose role is it anyway? a library practitioner's appraisal of the digital data deluge. ASIS&T Bulletin 37(5), 21–23 (2011)
29. Rusbridge, C., et al.: The digital curation centre: a vision for digital curation. In: Proc. 2005 IEEE Int. Symp. on Mass Storage Systems and Technology, pp. 31–41. IEEE Computer Society, Washington, DC (2005)
30. Sanderson, R., Harrison, J., Llewellyn, C.: A curated harvesting approach to establishing a multi-protocol online subject portal. In: Proc. 6th ACM/IEEE-CS Joint Conference on Digital Libraries, pp. 355–355. ACM, New York (2006)
31. Smith, P.L.: Where IR you?: Using "open access" to Extend the Reach and Richness of Faculty Research within a University. OCLC Systems & Services 24(3), 174–184 (2008)
32. Stahl, F., Schomm, F., Vossen, G.: Marketplaces for data: An initial survey. ERCIS Working Paper No. 12, Münster, Germany (2012)
33. Tan, W.C.: Provenance in Databases: Past, current, and future. IEEE Data Eng. Bull. 30(4), 3–12 (2007)
34. Van der Aalst, W., Van Hee, K.: Workflow Management: Models, Methods, and Systems. MIT Press (2004)
35. Vossen, G., Hagemann, S.: Unleashing Web 2.0: From Concepts to Creativity. Morgan Kaufmann Publishers (2007)
36. Witten, I., Frank, E., Hall, M.: Data Mining: Practical Machine Learning Tools and Techniques, 3rd edn. Morgan Kaufmann Publishers (2011)