KEYSTONE WG1: Activities and Results Overview on Representation of Structured Data Sources

Raquel Trillo-Lado[1(✉)] and Stefan Dietze[2]

[1] Depto. de Informática e Ingeniería de Sistemas (DIIS) e I3A, Universidad de Zaragoza, Zaragoza, Spain
raqueltl@unizar.es
[2] L3S Research Center, Appelstrasse 9a, 30167 Hannover, Germany
dietze@l3s.de

Abstract. The main goal of research in the Keystone Action COST IC1302 is to manage *big amounts of heterogeneous data*, particularly *structured data*, in order to provide users (people or software agents) with the data they require in an effective way with the minimum cost. The processes of managing and organizing data to provide users with them in an efficient way also generate new data that can be recollected and exploited to improve the processes; i.e., data about the processes involved can be used as feedback to improve them.

Keystone is organized in 4 working groups: Representation of Structure Data Sources (WG1), Keyword-based Search (WG2), User Interaction and Keyword Query Interpretation (WG3), and Research Integration, Showcases, Benchmarks and Evaluations (WG4). This chapter is focused on the research related to WG1 focusing on profiling, assessment, representation and discovery of structured datasets. The results of WG1 influence WG2 and WG3, whereas WG4 focuses on the integration of the results of all working groups and how to exploit them.

1 Introduction

There exists an increasing amount of data available. Some data are public (for example, on the Web) and everybody can exploit them, while others are accessible only for particular collectives of people under licenses that constrain their exploitation and exploration (for example, the clinical history of patients in hospitals, industrial patents, etc.). Moreover, in the last decade there has been an increment of the amount of structured data sources available, due to multiple reasons such as:

- the development of Information and Communication Technologies (ICT) which provide an infrastructure to digitalized, process and consume data;
- the popularization of the Linked Data Web and the Internet of Things, which foster a change of behavior on people and many companies and administrations, who decided to publish structured data (some of them coming from different types of sensors) on the Web; and

J. Szymański and Y. Velegrakis (Eds.): IKC 2017, LNCS 10546, pp. 196–214, 2018.
https://doi.org/10.1007/978-3-319-74497-1_20

– new politics fostered by different organizations and governments (for example, the Organization for Economic Cooperation and Development -OECD- declared that all publicly funded data should be available for everybody).

The main goal of research in the Keystone Action COST IC1302 is to manage *big amounts of heterogeneous data*, especially *structured data*, in order to provide users (people or software agents) with the data they require in an effective way with the minimum cost. Keystone is organized in 4 working groups: Representation of Structure Data Sources (WG1), Keyword-based Search (WG2), User Interaction and Keyword Query Interpretation (WG3), and Research Integration, Showcases, Benchmarks and Evaluations (WG4) as shown in Fig. 1. This chapter is focused on the research related to WG1, whose results influence WG2 and WG3, whereas WG4 focuses on the integration of the results of all working groups and how to exploit them.

Independently of the kind of data considered, the consumption or reuse of structured data (taking into account its licenses and legislation) is limited yet. Thus, in the Web of Linked Data, most users only consume and reuse well-known general-purpose reference data sources, such as WikiData and DBpedia, despite the fact that there exist other domain-specific data sources that could be more appropriate for their purposes. We consider that the causes of this behavior are the difficulties to locate, identify and exploit suitable data sources due to:

– *Noise and inconsistencies appear in the generation, transfer and transformation of data.* Thus, the longer the transfer chain and the more transformations, the more noise and inconsistencies are introduced.
– *A big amount of technologies to store and index structured data sources have appear recently.* There exist multiple technologies, such as MongoDB, Cassandra, traditional relational databases, graph databases and different RDF stores, to store and index structured data sources. So, there is a big amount

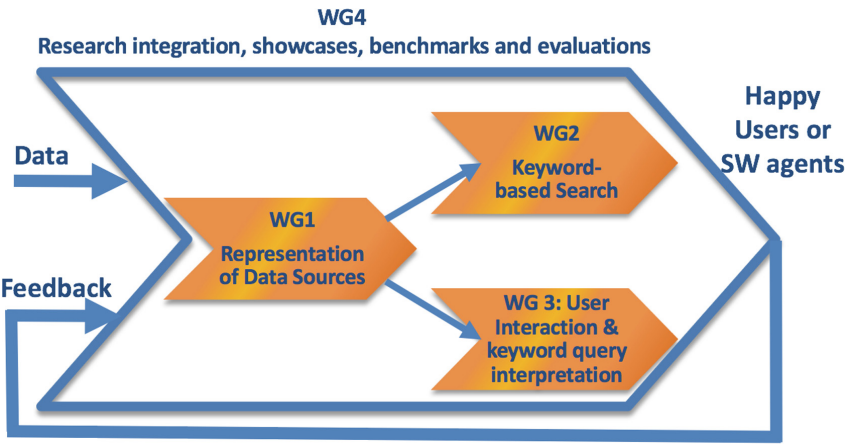

Fig. 1. Working groups of the Keystone Action COST IC-1302.

Fig. 2. Pipeline considered to publish or select a specific data source.

of interesting alternatives and it is difficult to decide which one is the most appropriate in a specific context, as there are not consensus guidelines and/or standards to allow users to select the most appropriate technologies for particular contexts.
- *The lack of knowledge and tools with reliable information about the nature of distributed third-party datasets.* For instance with respect to their quality, dynamics, temporal coverage or the addressed domains.
- *The lack of knowledge and tools to locate the data sources that are interesting for users with an specific purpose in an efficient way*; ideally, in an totally automatic way even when the users do not ask for it explicitly (i.e., based on a pull-based approach).

So, in this chapter, the research of different research groups and authors involved in Keystone WG1 is organized in the areas or categories of the pipeline (or data source value chain) in Fig. 2: Generation of Structured Data (WG1.A); Storing and Indexing of Structured Data (WG1.B); Characterization, Integration and Federation of Data Sources (WG1.C); and Selection and Retrieval of Data Sources (WG1.D).

Moreover, information about the different research groups that contribute to the research on which Keystone WG1 is focused, is also provided in Sect. 6 and the list of the authors that contributed to the elaboration of this survey chapter is provided in Sect. 7.

2 Generation of Structured Data (WG1.A)

The first question that arose when discussions about the generation of structured data took place in Keystone Working Group 1 was: *where do structured data come from?*, i.e., *who or what generates structured data?* The received feedback was organized in four overlapping groups: (1) from unstructured or semi-structured data sources (such as documents written in natural language and traditional HTML web pages), (2) from human users in a collaborative way, (3) from sensors and Internet of Things (IoT) devices, and (4) from other structured data sources. Moreover, discussions about how to publish or generate structured data are also relevant at this point (see Sect. 2.5).

2.1 From Unstructured or Semi-structured Data Sources

Since its creation in 2001, Wikipedia has become one of the most important sources of reliable information on the Web. Thus, currently, there are more than 280 active versions of Wikipedia in different languages. Wikipedia articles are typically split into two parts: (1) a body of unstructured text with details on the article subject and (2) an optional semistructured box usually called *infobox*. A considerable number of projects have exploited infoboxes in order to create structured data, such as Google's Knowledge Graph, Microsoft's Satori, and DBpedia [4] (the most famous one nowadays). More recently, in 2012, Wikimedia Deutschland proposed the Wikidata project [55], whose main goal is providing high-quality structured data acquired and maintained collaboratively to be directly used by Wikipedia to enrich its contents. DBpedia and Wikidata have become two important structured data sources in the current Linked Data Web. Thus, according to [42], DBpedia is the second node with more incoming links on the Linked Data Web, whereas Wikidata has been continuously increasing its popularity since its creation [56]. So, a great amount of data sources refer to them. A comparison between DBpedia and Wikidata and an analysis of their evolution are also done [23] by considering different criteria, metrics and frameworks focused on the quality of structured data sources [25,40,58]. There also exist recent works based on a wide set of techniques (text mining and ontology alignment, entity linking [21], etc.) such as [31], whose main purpose is the discovery of relationships among the elements of wikis written in different natural languages.

A great effort has been also made to develop techniques to extract structured data from texts and documents. So, techniques, tools and frameworks to perform extraction of data from the Web and texts have been developed. Some examples of this tools are: DBpediaSpotlight[1], Babelfy[2], different Temporal Taggers (e.g., SUTime[3] and HeidelTime[4]), Stanford Name Entity Recognition (NER)[5] and *Part of speech (POS)* tagging systems. On the other hand, there also exist other digital resources with a great potential from which structured data can also be extracted, such as images, video, multidimensional arrays containing for example environmental data, data streams coming from sensors with a certain frequency, etc. Some works around these topics are [20,54]. The former work presents a tool to create structured data sources about meteorological issues, while the latter one is focused on revealing new information about a virus by using *Information Extraction (IE)* techniques combined with existing genome sequenced data. However, there are not widely standard methods or techniques to answer the following questions: what features should be considered for images, video, and streaming data from sensors?, how a unique identifier for all data of these kinds of data sources should be built?, should this kind of data be associated to a

[1] https://github.com/dbpedia-spotlight/dbpedia-spotlight.
[2] http://babelfy.org/.
[3] https://nlp.stanford.edu/software/sutime.shtml.
[4] http://heideltime.ifi.uni-heidelberg.de/heideltime.
[5] https://stanfordnlp.github.io/CoreNLP/.

geographical position?, which granularity should be considered (country, GPS coordinates, region, ...)?, etc.

2.2 From Human Users in a Collaborative Way

In the context of Knowledge Representation and Artificial Intelligence, structured data sources are usually called *Knowledge Bases*. Thus, a *Knowledge Base* is considered a store of information or data that is available to draw on, or the underlying set of facts and rules of a certain domain stored in a specific format. Therefore, creating/generating a structured data source is quite similar to creating a Knowledge Base in these contexts.

Some authors considered that a Knowledge Base (KB) is composed of two main elements [37]: (1) a set of ontologies that establish the model of the data that the KB contains (this set of ontologies is also known as TBox or Terminological Box), and (2) data or instances that represent facts of the domain modeled by the TBox (this set of data or instances is also known as ABox or Assertional Box). There is a certain agreement of the definition of ontology; thus, the most popular definition of ontology is "explicit specification of a conceptualization" [35]. However, there exist different approaches to create ontologies. Some groups and tools use bottom-up approaches starting from folksonomies [57], while other methodologies such as NeON [53] follow a top-down approach that considers as starting point the knowledge of domain experts. On the other hand, up to our knowledge, there is no widely-adopted tool or technique to populate a KB, i.e., there does not exist a well-known technique or tool to create data of the ABox component. Nevertheless, some Extraction-Transformation and Load (ETL) systems have been adapted to populate a specific KB in certain domains. Besides, there are some emerging tools oriented to non-technical users that suggest attributes/properties and values to be filled by users to populate a KB in a collaborative way [49]. Moreover, adapting recommender system techniques, such as collaborative recommender techniques, content-based recommender techniques, knowledge-based recommender techniques, context aware recommender techniques and so on [51], could be also a possibility to be explored. The main challenge of using Recommender Systems in this context is how to evaluate their performance, as there are not standard benchmarks or datasets. So, a recent framework to generate synthetic data for the evaluation of context-aware recommendations systems was created [16]. On the other hand, recent studies focused on analyzing how the data sources evolve along the time, in particular on how the evolution of the editions of data performed collaboratively is [50].

Despite the fact that there is neither a widely-adopted methodology to create the TBox of a KB nor a widely-adopted system to populate the KB (to insert/update and delete data in the ABox component), there exists a widely adopted set of standard languages to create KB and ontologies. The most popular languages to create ontologies are: RDFS [11] and OWL [34], standardized by the World Wide Web Consortium (W3C), while the most popular language to populate ontologies is RDF [32].

2.3 From Sensors and IoT Devices

With the popularization of the Internet of Thing, multiple devices provide digital data coming from different types of sensors (temperature, location, humidity, etc.) for different purposes: remote control, automatic control, monitoring areas, etc. Independently of the type of sensor and the purpose of obtaining those data, different issues have arisen:

- Which devices should process the data? where the data process should be performed? in the sensors themselves?, in the infrastructure used to create the sensor network?, in servers where all data are collected and grouped? Different research groups have focused on solving these questions and new trends such as *Fog Computing* or *Edge Computing* extend the Cloud Computing paradigm to the edge of the network [52].
- When data should be transmitted from the sensors to the network? each certain time or period or when a relevant change happens (changes of values bigger than a certain threshold). The proposal in [38] considers only the transmission of certain values and a function to predict the new values, and when the trend of the sequence of values changes, a new transmission of values and prediction functions is performed. Other works follow a pull approach, i.e., consulting the current values of the sensors when they are required instead of generating and transmitting data from the sensors all the time (push approach).
- Which type of data should be provided to the consumers: raw data or smart data? The purpose of smart data is avoiding overload of raw data that are difficult to process and digest by final users; on the other hand when smart data are provided, usually several filters to simplified the output are applied and some filters could remove relevant raw data for the final user.

Another important issue that arises in the discussions about the totally connected world is how to exploit structured data by taking into account privacy and security issues. Moreover, industries demand methods to anonymize personal data in order to exploit them while guaranteeing the protection of their clients and workers and dealing with the right to forgetfulness in an appropriate way. So, guidelines, techniques and tools that deal with these issues from an interdisciplinary point of view are required. Frameworks and directives about security and/or safety issues usually consider the following dimensions of security represented as a triangle (as when priority to one of them is given, the other ones usually become weaker points): *Availability*, *Confidentiality* and *Integrity* of data. Besides, some frameworks consider other sub-dimensions of *Integrity* such as *Authenticity* and *Traceability* (or the provenance of data, where recent relevant works have been published [39,45]). Finally, we would like to remark that works related to the recent *block-chains* are also emerging to take into account security issues in environments with sensors [41].

Finally, notice that nowadays a great amount of data are also generated by the execution of different processes. This data are usually stored in logs of different type with a specific structure. Therefore, they could by analyze to

extract knowledge about the processes executed and evaluate their performance. These issues are study on a recent research area called Process Mining [1].

2.4 From Other Structured Data Sources

Nowadays, there exist a great number of *secondary data sources*, that obtain their contents from one or several different data sources, in contrast to *primary data sources*, that create their own contents from scratch. For example, the content of *Data-ware houses* for analytical purposes is usually generated from transactional systems, along the time. In this context, tools to facilitate: (a) the extraction of the data from the original systems, (b) the transformation of the data to integrate and clean it, and (c) loading the transformed data in the destiny systems have been popularized. Some examples of popular *Extraction Transformation and Load (ETL)* systems are Rapid Miner[6], Talend[7] and Pentaho[8]. All these tools provide mechanisms to deal with data curation. Nevertheless, this is a topic where there exist some open research issues such as *Entity De-duplication*, *Entity Disambiguation* and dealing with *multilingual aspects*.

Standards to translated relational data bases into semantic data sources based on RDF such as Direct Mapping [3] and R2RML [22] have also become popular in recent years. However, their use have not been widely adopted yet. When this translation is performed, there are two alternatives options: (1) materializing the RDF data by storing the same data in two different storage (one representation based on a relational approach, and another based on an RDF approach) and (2) using wrappers to create an virtual RDF model maintaining the original relational data base. The first option requires to deal with the problem of data redundancy (the same data stored by using two different models and structures), while the second option usually requires more processing time. A more complex scenario arise when dealing with heterogeneous data sources, that use different models with different semantics, is required.

2.5 Methodologies, Standards and Good Practices to Publish and Consume Structured Data

The main principles to publish Linked Data were established by Tim Berners-Lee et al. in 2006[9]. This proposal is based on: (1) use URIs as names for things/resources, (2) use HTTP dereference URIs so that people can look up data about the resources that they represent by means of a browser, (3) include links to other URIs in order to discover more things. These principles were refined later by Bizer et al. [8]. Besides, the releasing of DBpedia caused the creation of new RDF-based data sources on the Web, as it showed the required steps to develop and implement a *linked data source*. After that, a standard formal language to query that kind of data sources was required. So, in 2008, the

[6] https://rapidminer.com.
[7] https://www.talend.com.
[8] www.pentaho.com.
[9] https://www.w3.org/DesignIssues/LinkedData.html.

Protocol and RDF Query Language (SPARQL) was released as W3C a Recommendation [28]. Later, in 2013, this standard was updated [33]. Moreover, web services, *called SPARQL endpoints*, that allow to submit SPARQL queries to RDF data sources were also standardized. Unfortunately, a high percentage of users is not able to express their information needs in a SPARQL query as it requires to know the following elements: (1) the syntax of SPARQL in order to build a syntactically correct query, and (2) the underlying data structure of the source, i.e., how the data are organized (its schema or intension) and its semantics, in order to build a semantically correct query and to express the information need properly.

In order to easy the automatic interpretation and process of the available content of the web pages, i.e., to made them understandable for machines and not only for humans (that can read them), semantic annotations were created. A semantic annotation is an annotation embedded in the HTML source of a web pages that makes explicit the semantic meaning of a certain content (for example, a sequence of characters) for a machine. The standard language to make semantic annotations is the W3C Recommendation RDFa (RDF annotation). This language was release as Recommendation in 2012 [43], and later updated in 2015 [44]. Despite the fact that there exist a great amount of semantically annotated and linked data on the Web, most of its current content of the Web is not annotated. On the other hand, during the last decade, initiatives promoted by the main search engine companies (Google, Yandex, Bing, etc.) have created standard vocabularies (ontologies) to annotate the web content (Schema.org [36] has been one of these successful initiatives). Moreover, these companies promote the use of annotation by ranked annotated web pages on the first positions of the pages of results of the searches performed by their users.

In conclusion, currently, there exist two main ways of consuming Linked Data Web: (1) crawling web pages with semantic annotations (such as RDFa annotations) periodically in order to discover new data, and (2) querying SPARQL endpoints either to find out their structure or to obtain specific data.

3 Storing and Indexing of Structured Data (WG1.B)

In 1970, Edgar F. Codd defined the foundations of the relational database model to structure data within a database. This model has been widely used and considered so far, and its implementations satisfy the properties ACID (Atomicity, Consistency, Isolation and Durability). At the same time as the foundations of the relational model were being defined, Donald D. Chamberlin and Raymond F. Boyce, developed a language called Specifying Queries As Relational Expressions (SQUARE) to query databases based on that model. The evolution of SQUARE was later, in 1974, called Structured English Query Language (SEQUEL). SEQUEL was oriented to non-experts users because it specified "what" data to retrieve instead of "how" to retrieve them (i.e., it was a declarative language instead of a procedurallanguage). SEQUEL was renamed and standardized as the widely adopted Structured Query Language (SQL) in the

middle of the 80's and became the most used language to query data sources in the 80's and the 90's decade. Thus, although other types of models (e.g., deductive, pure object oriented, etc.) were proposed, they did not achieve commercial success.

With the explosion of the Web in the middle of 90's decade, the use of databases oriented to manage text documents increased. Moreover, currently there exist a wide range of different types of data sources with different purposes based on different models which are generally classify NoSQL databases or Not only SQL databases. The most popular ones are the following ones:

1. graph-oriented databases (where the Triple Stores for managing RDF fit),
2. multivalued databases,
3. object-oriented databases,
4. columnar databases,
5. key-value databases, and
6. multi-model databases.

Most of these new types of models focused on satisfying a different set of properties from ACID properties. Thus, the NoSQL databases focus on the Basically Availability, Soft-State, Eventually Consistent (BASE) properties [10] emerged when *Consistency, Availability and Partition-Tolerance (CAP)* theorem became popular around 2000 [9]. Moreover, notice that the storage of these data sources can be distributed on a network. On the other hand, federation of independent data sources is commonly required to tackled complex problems; for example, in order to create pollutant dispersions models for an city is required to obtain data from meteorological models, traffic models, geographical information systems, and the geometry of the buildings of the city.

Regarding the structures and indexes to store the structured data sources, the most popular ones are the following ones:

1. balance trees and B+ trees for relational databases,
2. inverted indexes for databases oriented to store documents, and
3. different formats based on text files for RDF such as RDF Turtle [24], RDF N-Triple [5], RDF/XML, RDF/JSON, etc.

Moreover, several Keystone members has proposed different structures to index and store RDF in a binary way. The most popular approaches proposed are: Head Dictionary Triple (HDT) [29], HDT-MR (based on Map-Reduce) [30], RDFCSA (a compact RDF store based on compressed Suffix Arrays, a well-known self-index) [13] and K2-Triples (a compressed vertical partitioning for RDF) [2]. Moreover, works about versioning RDF, i.e., the evolution of an RDF data source along time have also been proposed. Some relevant works are: compressed kd-tree for temporal-graphs [18], compressed suffix-array from temporal-graphs [12] and RDF-Archive [19].

Finally, notice that indexes or structures to improve the access or storage of structure data sources can be classified by considering the following categories: (1) *In Memory Structures* vs *In Disk Structures*; (2) *Compact Structures* vs

Structures Over Plain Data (generally text); and (3) *Self-indexing Structures* where the index and the data are kept in a unique in-memory data structure that allows indexed searches and to recover the original data.

4 Characterization, Integration and Federation of Data Sources (WG1.C)

At this point the main questions about characterization, integration and federation of data sources discussed in the context of the Semantic Web by Keystone Working Group 1 were: *which meta-data should be considered to describe a (RDF) data source?*, *how to evaluate the quality of a data source?* and *how to integrate/federate heterogeneous data sources?*.

With respect to the first question, notice that, currently, there exist multiple standards languages and initiatives to describe the content of a data source, such as: RDFS [11], OWL [34], VoID[10] and DCAT[11]. Nevertheless, recently, some Keystone members have been working on a survey to provide a comprehensive overview of the RDF dataset profiling features, methods, tools and vocabularies [7]. With respect to the second question, a great amount of work have been done recently. Some works focus on defining methodologies and metrics grouped in dimensions to study the quality of the data sources such as [58]; while others focused of creating methods and tools to perform that evaluation efficiently. Some recent tools are: qSKOS[12], Skosify[13], Luzzu [25] and PoolParty[14]. Finally, with respect to the third question, some systems developed by Keystone members in order to integrate/federate heterogeneous data sources are briefly described in the following:

- *MOMIS* [14]. It is an open data tool, developed by the University of Modena and Reggio Emilia and the Enterprise DataRiver, to perform data integration from heterogeneous static data sources.
- *SOS-SM* [47,48]. It is a framework, developed by the University of Santiago de Compostela, whose aim is the semantic mediation between environmental observation datasets through OGC Sensor Observations Service interfaces. The framework combines a mediator/wrapper architecture with a Local As View approach for data integration, supported by a global model based on the Semantic Sensor Network ontology proposed by the W3C. General purpose wrappers were also developed to incorporate vector-based datasets recorded in spatial relational databases and raster-based datasets accessed through UNIDATA NETCDF Subset services.

There also exist multiple initiatives and projects to exploit open data in the context of smart cities [46]. However, up to our knowledge, most of them are

[10] https://www.w3.org/TR/void/.
[11] https://www.w3.org/TR/vocab-dcat.
[12] https://github.com/cmader/qSKOS.
[13] https://github.com/NatLibFi/Skosify.
[14] https://www.poolparty.biz/.

focused on specific domains such as transportation, pollution, energy, point of interests for tourists, etc. On the other hand, KnowledgeManagement4City [6] is an ontology oriented to model smart city services. This ontology provides a unified view that facilitates the creation of any service for the city, as all services are managed in an uniform way.

5 Selection and Retrieval of Data Sources (WG1.D)

At this point two main questions were discussed by Keystone Working Group 1 were: *how to discover or recommend structure data sources?* and *how to discover equivalent concepts, properties and instances from two different data sources?*. With respect to the first question, different research groups involved in the WG1 of Keystone Action COST IC1302 have published recent works about recommendation. Some representatives examples are: [15,17]. On the other hand, with respect to the second question, some relevant research papers have been also recently published [26,27].

6 Composition of Working Group 1

According to the information in the website of the Keystone Action COST IC1302 (http://www.keystone-cost.eu/), the Working Group 1 is composed of 162 members (41 females and 121 males) belonging to 28 research groups (see Tables 1 and 2). For more detail about the host countries of the different researchers involved in the working group see Table 3. Most of members of the working group are currently active in research areas related to the Working Group 1 of Keystone (Representation of Structured Data Sources). In more detail, the distribution of people involved who have provided feedback for this chapter, by considering their host countries, is shown in Table 4.

When the papers recollected to analyze the research results of the Keystone WG 1 are clustered by considering the research groups to which their authors belong, clusters showing collaborations on topics related to WG1 among the research groups participating in this package are created (see Figs. 3 and 4). Notice that the research groups that have published more joint papers with authors from other research groups are those groups whose researchers have been involved the leadership of the WG 1 (the leaders of the WG1 belong to the research groups represented by DE1 and ES4) or the network (the chair of the Action belongs to the research group IT1, while the scientific coordinator of the action belongs to the research group represented by IT2).

Finally, research groups were also categorized by considering the research topics of the papers that they provided and the steps of the data value chain defined in this chapter (WG1.A, WG1.B, WG1.C and WG1.D). Moreover, the category other was also considered (Fig. 5).

Table 1. Research groups in Keystone Working Group 1 per country (part 1 of 2).

Country	Name of the research group and number of researchers
Albania (AL)	Computer Engineering Department Epoka University in Tirane (1 researcher)
Austria (AT)	Vienna University of Economics and Business (1 researcher)
	Information and Software Engineering Group (IFS) Institute of Software Technology and Interactive Systems (ISIS) TU Wien (1 researcher)
Belgium (BE)	None research group
Bulgaria (BG)	Bulgarian Academy of Sciences in Sofia (1 researcher)
Croatia (HR)	None research group
Cyprus (CY)	None research group
Estonia (EE)	School of Information Technologies. Dept. of Software Science. Tallinn University of Technologies (1 researcher)
Finland (FI)	School of Information Technologies. Dept. of Software Science of Tallinn University of Technologies (1 researcher)
France (FR)	University Claude Bernard Lyon (1 researcher)
	CNRS - Centre national de la recherche scientifique (1 researcher)
Germany (DE)	L3S Research Center of the Leibniz University Hannover (2 researchers)
	Hannover University of Applied Sciences and Arts (2 researchers)
Greece (GR)	Department of Computer Science and Biomedical Informatics School of Sciences, University of Thessaly (1 researcher)
	Computer Science Department, University of Crete (1 researcher)
	Software Technology and Network Applications Laboratory Department of Electronic and Computer Engineering Technical University of Crete (1 researcher)
	Institute for the Management of Information Systems Research and Innovation Center ATHENA in Athens (1 researcher)
	Knowledge and Uncertainty Research Laboratory (RAB Lab) Department of Informatics and Telecommunications University of Peloponnese (1 researcher)
Ireland (FI)	Insight Center for Data Analytics National University of Ireland -NUIGalway- (1 researcher)
Italy (IT)	Databases (DBGroup), University of Modena and Reggio Emilia (2 researchers)
	Data Management Group Dept. of Information Engineering and Computer Science University of Trento (1 researcher)
	Process and Data Intelligence (PDI), Information Technology Center Fondazione Bruno Kessler (1 researcher)
	Department of Computer Science and Engineering University of Bologna (1 researcher)
Macedonia (MK)	None research group
Malta (MT)	None research group
Netherlands (NL)	None research group
New Zealand (TK)	None research group
Norway (NO)	None research group
Poland (PL)	None research group
Portugal (PT)	None research group
Romania (RO)	Faculty of Automatic Control and Computers Computer Science Department University Politehnica of Bucharest (1 researcher)
Serbia (RS)	None research group
Slovenia (SI)	None research group
Slovakia (SK)	None research group

Table 2. Research groups in Keystone Working Group 1 per country (part 2 of 2).

Country	Name of the research group and number of researchers
Spain (SP)	DataWeb Group, Depto. of Computer Science University of Valladolid, Segovia (1 researcher)
	Databases Laboratory (LBD). Computer Science and Technology Faculty University of A Coruña (5 researchers)
	Databases Laboratory (LBD) Computer Graphics and Data Engineering (COGRADE) Singular Information Technologies Research Center (CiTIUS) University of Santiago de Compostela (1 researcher)
	Computer Science and Software Engineering Department (DIIS) University of Zaragoza (5 researchers)
	Aragon Institute of Engineering Research (I3A) University of Zaragoza (2 researchers)
	Khaos Research, Depto. of Computer Languages and Computing Sciences University of Malaga (3 researchers)
	ETSE Telecomunicación, University of Vigo (1 researcher)
	Barcelona Supercomputing Center (1 researcher)
Sweden (SE)	None research group
Switzerland (CH)	University of Geneva, Faculty of economics and social sciences Department Hautes études commerciales (1 researcher)
Ukraine (UA)	None research group
United Kingdom (UK)	None research group

Table 3. Number of researchers in Keystone Working Group 1 per country.

Country	N^o people	Country	N^o people	Country	N^o people
Albania (AL)	1	Austria (AT)	2	Belgium (BE)	3
Bulgaria (BG)	3	Croatia (HR)	4	Cyprus (CY)	1
Estonia (EE)	1	Finland (FI)	2	France (FR)	10
Germany (DE)	12	Greece (GR)	8	Ireland (FI)	5
Italy (IT)	13	Macedonia (MK)	1	Malta (MT)	4
Netherlands (NL)	5	New Zealand (TK)	2	Norway (NO)	None
Poland (PL)	2	Portugal (PT)	5	Romania (RO)	23
Serbia (RS)	7	Slovenia (SI)	2	Slovakia (SK)	3
Spain (SP)	27	Sweden (SE)	7	Switzerland (CH)	2
Ukraine (UA)	2	United Kingdom (UK)	5		

Table 4. Number of researchers per country who provided feedback to create this chapter.

Country	N^o people	Country	N^o people	Country	N^o people
Albania (AL)	None	Austria (AT)	2	Belgium (BE)	None
Bulgaria (BG)	1	Croatia (HR)	None	Cyprus (CY)	None
Estonia (EE)	1	Finland (FI)	1	France (FR)	2
Germany (DE)	2	Greece (GR)	5	Ireland (FI)	1
Italy (IT)	5	Macedonia (MK)	None	Malta (MT)	None
Netherlands (NL)	None	New Zealand (TK)	None	Norway (NO)	None
Poland (PL)	None	Portugal (PT)	None	Romania (RO)	1
Serbia (RS)	None	Slovenia (SI)	None	Slovakia (SK)	None
Spain (SP)	12	Sweden (SE)	None	Switzerland (CH)	2
Ukraine (UA)	None	United Kingdom (UK)	None		

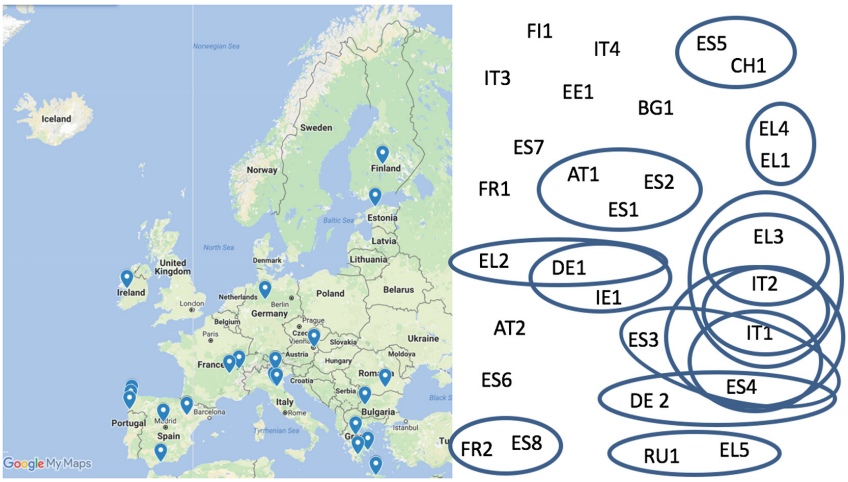

Fig. 3. Research groups clustered by considering joint papers related to WG1 in the last 4 year.

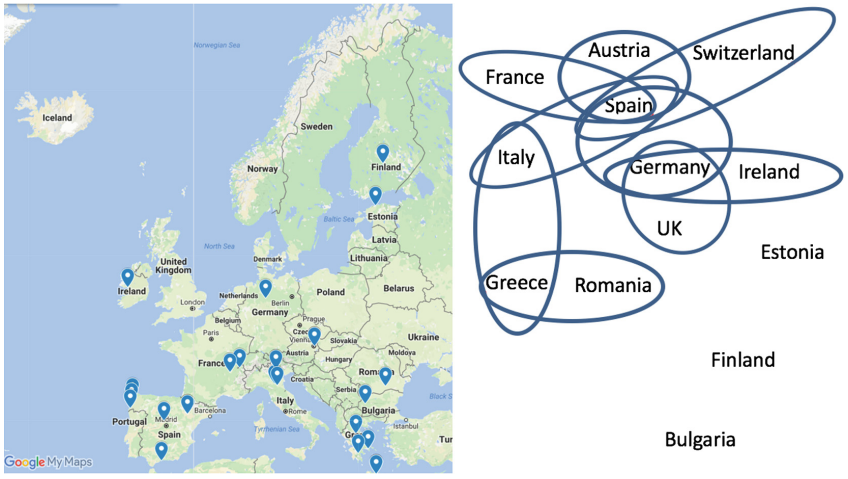

Fig. 4. Countries of the researchers from WG1 clustered by considering joint papers related to topics of WG1.

7 Researchers Contributing to This Survey

We sincerely thank every member of the working group for the work done along the last four years. We specially thank those members that help us to analyze the state of the art of research related to Keystone WG1 and provided us with references to their research papers. These researchers are the following ones (in alphabetic order by surname): Prof. José F. Aldana, Prof. Nieves R. Brisaboa, Dr. Ioannis Anagnostopoulos, Dr. Ilaria Bartolini, Dr. Fernando Bobillo,

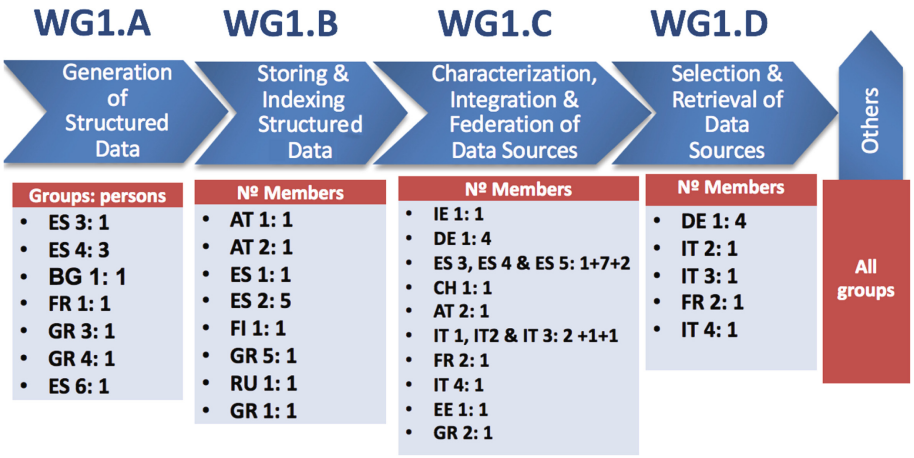

Fig. 5. Research groups of Keystone Action COST IC-1302 categorized according to the data value chain defined in this chapter.

Dr. John Breslin, Dr. Ana Cerdeira Pena, Dr. Elena Demidova, Dr. Stefan Dietze, Dr. Mauro Dragoni, D. Dudić, Dr. Pablo Fafalios, Prof. Gilles Falquet, Prof. Antonio Fariña Martínez, Dr. Javier D. Fernández, Catarina Ferreira da Silva, Dr. Francesco Guerra, Dr. Ramón Hermoso, Dr. Claudia Ifrim, Dr. Sergio Ilarri, Dr. Ekaterini Ioannou, Dr. Javier Lacasta, Dr. Susana Ladra, Dr. Martín López Nores, Dr. Mihai Lupu, Dr. Miguel A. Martínez, Dr. Javier Nogueras, Dr. Enn Õunapuu, V. Pajić, Dr. José Ramón Paramá Gabía, Dr. Laura Po, Prof. José Ramón Ríos Viqueira, Dr. Ma del Mar Roldán, Dr. Tarcísio Souza, Dr. Yannis Stavrakas, Dr. Velislava Stoykova, Prof. Vagan Terziyan, Dr. Raquel Trillo Lado, Dr. Genoveva Vargas, Prof. Yannis Velegrakis, and Dr. Manolis Wallace. Thus, feedback to create this chapter has been received from 12 different countries (see more details in Table 4).

References

1. van der Aalst, W.M.P.: Process Mining - Data Science in Action, 2nd edn. Springer, Heidelberg (2016). https://doi.org/10.1007/978-3-662-49851-4
2. Álvarez-García, S., Brisaboa, N.R., Fernández, J.D., Martínez-Prieto, M.A., Navarro, G.: Compressed vertical partitioning for efficient RDF manag-breakement. Knowl. Inf. Syst. **44**(2), 439–474 (2015). https://doi.org/10.1007/s10115-014-0770-y
3. Arenas, M., Bertails, A., Prud'hommeaux, E., Sequeda, J.: A direct mapping of relational data to RDF (2012). http://www.w3.org/TR/2012/REC-rdb-direct-mapping-20120927/
4. Auer, S., Bizer, C., Kobilarov, G., Lehmann, J., Cyganiak, R., Ives, Z.: DBpedia: a nucleus for a web of open data. In: Aberer, K., et al. (eds.) ASWC/ISWC 2007. LNCS, vol. 4825, pp. 722–735. Springer, Heidelberg (2007). https://doi.org/10.1007/978-3-540-76298-0_52. http://dl.acm.org/citation.cfm?id=1785162.1785216

5. Beckett, D.: RDF 1.1 n-triples, a line-based syntax for an RDF graph, W3C recommendation, 25 February 2014
6. Bellini, P., Benigni, M., Billero, R., Nesi, P., Rauch, N.: Km4city ontology building vs data harvesting and cleaning for smart-city services. J. Vis. Lang. Comput. **25**(6), 827–839 (2014)
7. Ben Ellefi, M., Bellahsene, Z., John, B., Demidova, E., Dietze, S., Szymanski, J., Todorov, K.: RDF dataset profiling - a survey of features, methods, vocabularies and applications. Semantic Web J. (2017). http://www.semantic-web-journal.net/content/rdf-dataset-profiling-survey-features-methods-vocabularies-and-applications. Accepted in August 2017 (to appear)
8. Bizer, C., Heath, T., Berners-Lee, T.: Linked data - the story so far. Int. J. Semantic Web Inf. Syst. **5**(3), 1–22 (2009)
9. Brewer, E.: Invited keynote on 19th ACM Symposium on Principles of Distributed Computing (PODC) (2000)
10. Brewer, E.: Pushing the cap: strategies for consistency and availability. Comput. **45**(2), 23–29 (2012). https://doi.org/10.1109/MC.2012.37
11. Brickley, D., Guha, R.: RDF schema 1.1, W3C recommendation, 25 February 2014. https://www.w3.org/TR/rdf-schema/
12. Brisaboa, N.R., Caro, D., Fariña, A., Rodríguez, M.A.: A compressed suffix-array strategy for temporal-graph indexing. In: de Moura, E., Crochemore, M. (eds.) SPIRE 2014. LNCS, vol. 8799, pp. 77–88. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-11918-2_8
13. Brisaboa, N.R., Cerdeira-Pena, A., Fariña, A., Navarro, G.: A compact RDF store using suffix arrays. In: Iliopoulos, C., Puglisi, S., Yilmaz, E. (eds.) SPIRE 2015. LNCS, vol. 9309, pp. 103–115. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-23826-5_11
14. Cabri, G., Guerra, F., Vincini, M., Bergamaschi, S., Leonardi, L., Zambonelli, F.: Momis: exploiting agents to support information integration. Int. J. Cooperative Inf. Syst. **11**(3), 293–314 (2002)
15. Cadegnani, S., Guerra, F., Ilarri, S., del Carmen Rodríguez-Hernández, M., Lado, R.T., Velegrakis, Y., Amaro, R.: Exploiting linguistic analysis on URLs for recommending web pages: a comparative study. Trans. Comput. Collect. Intell. **26**, 26–45 (2017)
16. del Carmen Rodríguez-Hernández, M., Ilarri, S., Hermoso, R., Lado, R.T.: DataGenCARS: a generator of synthetic data for the evaluation of context-aware recommendation systems. Pervasive Mob. Comput. **38**, 516–541 (2017). https://doi.org/10.1016/j.pmcj.2016.09.020
17. del Carmen Rodríguez-Hernández, M., Ilarri, S., Lado, R.T., Guerra, F.: Towards keyword-based pull recommendation systems. In: ICEIS, vol. 1, pp. 207–214. SciTePress (2016)
18. Caro, D., Rodríguez, M.A., Brisaboa, N.R., Fariña, A.: Compressed k^d-tree for temporal graphs. Knowl. Inf. Syst. **49**(2), 553–595 (2016). https://doi.org/10.1007/s10115-015-0908-6
19. Cerdeira-Pena, A., Fariña, A., Fernández, J.D., Martínez-Prieto, M.A.: Self-indexing RDF archives. In: Bilgin, A., Marcellin, M.W., Serra-Sagristà, J., Storer, J.A. (eds.) 2016 Data Compression Conference, DCC 2016, Snowbird, UT, USA, 30 March–1 April 2016, pp. 526–535. IEEE (2016). https://doi.org/10.1109/DCC.2016.40
20. Dudic, D., Zlatanovic, I., Gligorević, K., Urosevic, T.: Solar: a software tool for meteorological data processing. Agri. Eng. **39**(4), 51–61 (2014). ISSN 0554-5587

21. Dai, H.-J., Wu, C.-Y., Tzong-Han, R., Hsu, T.W.-L.: From entity recognition to entity linking: a survey of advanced entity linking techniques (2013)
22. Das, S., Sundara, S., Cyganiak, R.: R2RML: RDB to RDF mapping language (2012). http://www.w3.org/TR/2012/REC-rdb-direct-mapping-20120927/
23. Abián, D., Guerra, F., Guerra, F., Martinez-Romanos, J., Trillo-Lado, R.: Wikidata and DBpedia: a comparative study. In: Proceedings of the 3rd International Keystone Conference (2017)
24. Beckett, D., Berners-Lee, T., Prud'hommeaux, E., Carothers, G., Machina, L.: RDF 1.1 turtle, terse RDF triple language, W3C recommendation, 25 February 2014
25. Debattista, J., Auer, S., Lange, C.: Luzzu: a methodology and framework for linked data quality assessment. J. Data Inf. Qual. **8**(1), 4:1–4:32 (2016). https://doi.org/10.1145/2992786
26. Ben Ellefi, M., Bellahsene, Z., Dietze, S., Todorov, K.: Beyond established knowledge graphs-recommending web datasets for data linking. In: Bozzon, A., Cudre-Maroux, P., Pautasso, C. (eds.) ICWE 2016. LNCS, vol. 9671, pp. 262–279. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-38791-8_15
27. Ben Ellefi, M., Bellahsene, Z., Dietze, S., Todorov, K.: Dataset recommendation for data linking: an intensional approach. In: Sack, H., Blomqvist, E., d'Aquin, M., Ghidini, C., Ponzetto, S.P., Lange, C. (eds.) ESWC 2016. LNCS, vol. 9678, pp. 36–51. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-34129-3_3
28. Prud'hommeaux, E., Seaborne, A.: SPARQL Query Language for RDF, W3C Recommendation, 15 January 2008. https://www.w3.org/TR/rdf-sparql-query/
29. Fernández, J.D., Martínez-Prieto, M.A., Gutiérrez, C., Polleres, A., Arias, M.: Binary RDF representation for publication and exchange (HDT). J. Web Sem. **19**, 22–41 (2013). https://doi.org/10.1016/j.websem.2013.01.002
30. Giménez-García, J.M., Fernández, J.D., Martínez-Prieto, M.A.: HDT-MR: a scalable solution for RDF compression with HDT and MapReduce. In: Gandon, F., Sabou, M., Sack, H., d'Amato, C., Cudré-Mauroux, P., Zimmermann, A. (eds.) ESWC 2015. LNCS, vol. 9088, pp. 253–268. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-18818-8_16
31. Gottschalk, S., Demidova, E.: Multiwiki: interlingual text passage alignment in Wikipedia. ACM Trans. Web **11**(1), 6:1–6:30 (2017)
32. Klyne, G., Carroll, J.J., McBride, B.: RDF 1.1 concepts and abstract syntax, W3C recommendation, 25 February 2014. https://www.w3.org/TR/rdf11-concepts/
33. The W3C SPARQL Working Group: SPARQL 1.1 W3C recommendation, 21 March 2013. https://www.w3.org/TR/sparql11-overview/
34. W3C OWL Working Group: OWL 2 web ontology language document overview, 2nd edn., W3C Recommendation, 11 December 2012. https://www.w3.org/TR/owl2-overview/
35. Gruber, T.R.: A translation approach to portable ontology specifications. Knowl. Acquis. **5**(2), 199–220 (1993). https://doi.org/10.1006/knac.1993.1008
36. Guha, R.V., Brickley, D., MacBeth, S.: Schema.org: evolution of structured data on the web. Queue **13**(9), 1010–1037 (2015). http://doi.acm.org/10.1145/2857274.2857276
37. Hernández, I.R.: Development of a system to populate Knowledge Bases on the Web of Data, Final Project for the Computer Science Degree. University of Zaragoza (2016)

38. Ilarri, S., Wolfson, O., Mena, E., Illarramendi, A., Sistla, A.P.: A query processor for prediction-based monitoring of data streams. In: Kersten, M.L., Novikov, B., Teubner, J., Polutin, V., Manegold, S. (eds.) Proceedings of the 12th International Conference on Extending Database Technology, EDBT 2009, Saint Petersburg, Russia, 24–26 March 2009, International Conference Proceeding Series, vol. 360, pp. 415–426. ACM (2009). https://doi.org/10.1145/1516360.1516409
39. Karsai, L., Fekete, A., Kay, J., Missier, P.: Clustering provenance facilitating provenance exploration through data abstraction. In: Binnig, C., Fekete, A., Nandi, A. (eds.) Proceedings of the Workshop on Human-In-the-Loop Data Analytics, HILDA@SIGMOD 2016, San Francisco, CA, USA, 26 June–1 July 2016, p. 6. ACM (2016). https://doi.org/10.1145/2939502.2939508
40. Kontokostas, D., Westphal, P., Auer, S., Hellmann, S., Lehmann, J., Cornelissen, R., Zaveri, A.: Test-driven evaluation of linked data quality. In: Proceedings of the 23rd International Conference on World Wide Web, WWW 2014, pp. 747–758. International World Wide Web Conferences (2014). http://svn.aksw.org/papers/2014/WWW_Databugger/public.pdf
41. Kosba, A.E., Miller, A., Shi, E., Wen, Z., Papamanthou, C.: Hawk: the blockchain model of cryptography and privacy-preserving smart contracts. IACR Cryptology ePrint Archive 2015, 675 (2015). http://dblp.uni-trier.de/db/journals/iacr/iacr2015.html#KosbaMSWP15
42. Lehmann, J., Isele, R., Jakob, M., Jentzsch, A., Kontokostas, D., Mendes, P.N., Hellmann, S., Morsey, M., van Kleef, P., Auer, S., Bizer, C.: DBpedia - a large-scale, multilingual knowledge base extracted from Wikipedia. Sem. Web J. **6**(2), 167–195 (2015). http://jens-lehmann.org/files/2015/swj_dbpedia.pdf
43. Sporny, M., Digital Bazaar, Inc.: RDFa lite 1.1, W3C recommendation 7 June 2012. https://www.w3.org/TR/2012/REC-rdfa-lite-20120607/
44. Sporny, M., Digital Bazaar, Inc.: RDFa lite 1.1, 2nd edn., W3C recommendation, 17 March 2015. https://www.w3.org/TR/2015/REC-rdfa-core-20150317/
45. Oliveira, W., Missier, P., Ocaña, K., de Oliveira, D., Braganholo, V.: Analyzing provenance across heterogeneous provenance graphs. In: Mattoso, M., Glavic, B. (eds.) IPAW 2016. LNCS, vol. 9672, pp. 57–70. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-40593-3_5
46. Nesi, P., Po, L., Viqueira, J.R.R., Trillo-Lado, R.: An integrated smart city platform. In: Proceedings of the 3rd International Keystone Conference (2017)
47. Regueiro, M.A., Viqueira, J.R.R., Stasch, C., Taboada, J.A.: Sensor observation service semantic mediation: generic wrappers for in-situ and remote devices. In: Comyn-Wattiau, I., Tanaka, K., Song, I.-Y., Yamamoto, S., Saeki, M. (eds.) ER 2016. LNCS, vol. 9974, pp. 269–276. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46397-1_21
48. Regueiro, M.A., Viqueira, J.R.R., Stasch, C., Taboada, J.A.: Semantic mediation of observation datasets through sensor observation services. Future Gener. Comp. Syst. **67**, 47–56 (2017)
49. Rodriguez-Hernandez, I., Trillo-Lado, R., Yus, R.: WikInfoboxer: a tool to create Wikipedia infoboxes using DBpedia. In: XXI Jornadas de Ingeniería del Software y Bases de Datos (JISBD 2016), Demo track, Salamanca (Spain), 4 p., September 2016
50. Sarasua, C., Checco, A., Demartini, G., Difallah, D.E., Feldman, M., Pintscher, L.: Editing behavior over time power vs. Standard Wikidata editors at Wikidatacon (2017). https://www.slideshare.net/cristinasarasua/editing-behavior-over-time-power-vs-standard-wikidata-editors-81276124

51. Smith, B., Linden, G.: Two decades of recommender systems at Amazon.com. IEEE Internet Comput. **21**(3), 12–18 (2017)
52. Stojmenovic, I., Wen, S.: The fog computing paradigm: scenarios and security issues. In: Proceedings of the 2014 Federated Conference on Computer Science and Information Systems, Warsaw, Poland, 7–10 September 2014, pp. 1–8 (2014). https://doi.org/10.15439/2014F503
53. Suárez-Figueroa, M.C., Gómez-Pérez, A., Fernández-López, M.: The NeOn methodology for ontology engineering. In: Suárez-Figueroa, M., Gómez-Pérez, A., Motta, E., Gangemi, A. (eds.) Ontology Engineering in a Networked World. Springer, Heidelberg (2012). http://oa.upm.es/21469/
54. Pajic, V., Banovic, M.B.B., Dudic, D.: Mining PMMoV genotype-pathotype association rules from public databases. In: Proceedings of International Conference Belgrade Bioinformatics (BelBI), Belgrade, Serbia (2016)
55. Vrandecic, D.: Wikidata: a new platform for collaborative data collection. In: Mille, A., Gandon, F.L., Misselis, J., Rabinovich, M., Staab, S. (eds.) WWW (Companion Volume), pp. 1063–1064. ACM (2012)
56. Vrandečić, D., Krötzsch, M.: Wikidata: a free collaborative knowledgebase. Commun. ACM **57**(10), 78–85 (2014). http://doi.acm.org/10.1145/2629489
57. Wal, T.V.: Folksonomy coinage and definition (2007). http://vanderwal.net/folksonomy.html
58. Zaveri, A., Rula, A., Maurino, A., Pietrobon, R., Lehmann, J., Auer, S.: Quality assessment for Linked Data: a survey. Semant. Web J. (2015). http://www.semantic-web-journal.net/content/quality-assessment-linked-data-survey