From Research to Business: The Web of Linked Data

Irene Celino[1], Emanuele Della Valle[1,2], and Dario Cerizza[1]

[1] CEFRIEL – ICT Institute, Politecnico of Milano,
Via Fucini 2, 20133 Milano, Italy
name.surname@cefriel.it
[2] Dipartimento di Elettronica e Informazione, Politecnico di Milano,
Piazza Leonardo da Vinci 32, 20133 Milano, Italy
emanuele.dellavalle@polimi.it

Abstract. The last decade of research in the Web field gave a great importance to the studies about the Semantic Web. The idea of a Web of Data is now becoming more and more popular also outside of the pure scientific community. The idea of *linked data* is thus gaining ground and demonstrating its advantages and its opportunities in the business world. Still a lot of research is there to come.

In this paper, we discuss the need for linked data technologies, we illustrate two case studies from European research projects and we examine the opportunities from the business and the technological point of view.

Keywords: linked data, Semantic Web, Web of Data, Service Web, Reasoning Web, Urban Computing.

1 Introduction

In the area of information management, the market is constantly asked for more and better solutions to solve the *problem of integration*. Why is industry so keen in finding new answers to information integration? Today organizations and enterprises have to face at least three different challenges:

- they have a problem of *scale*: they must manage very large amounts of data, which grow and evolve continuously;
- they have a problem of *data heterogeneity*: the data they produce and consume every day belong to numerous and different typologies (documents, media, email, Web results, contacts, etc.);
- they have a problem of *system heterogeneity*: those data are managed by numerous and different information systems (DB, legacy systems, ERP, etc.).

But why is integration so significant? Because integration always gives an added value: in getting a global high-level view over different and independent systems; in sharing knowledge between groups and partners; in unleashing business opportunities which are enabled only by unifying and combining efforts; in answering the questions of decisions makers, as it happens in the Business Intelligence area. Integration – using an effective maths metaphor – is when $1 + 1 > 2$.

W. Abramowicz and D. Flejter (Eds.): BIS 2009 Workshop, LNBIP 37, pp. 141–152, 2009.

Therefore, solving the integration problem seems to be the first point in everybody innovation agenda. In order to understand the possible ways out, it is interesting to have a look at those who were able to ride and exploit the integration challenge to their own advantage. Some lessons learned come from the so called Web 2.0.

As James Surowiecki effectively explained in his famous bestseller "Wisdom of Crowds"[1], a strong driver of the integration in the Web 2.0 is constituted by the *participation politics*: the collaboration of people makes a large task smaller, turns a big problem into a manageable one.

Moving from the attitudinal and social aspects to the technological point of view, the Web 2.0 revealed the great success of *mash-ups*. A mash-up is an application made up of the light integration of artifacts provided by third parties (like API or REST services). Beside being a way to implement Web applications, mash-ups are also a new *integration paradigm* to software development, in which loose agreements – instead of tight and complex interface specifications – are needed to get to a useful result. Last but not least, the Web proved to be indispensable for data and information *publication and access*: we are more and more accustomed to store our information on the Web and we also access it more and more frequently on the Web (e.g. by retrieving it via search engines).

From all those considerations, we can easily come to a question: is it possible to reach *integration on the Web*? Can we use the Web as a *platform* for integration? How can we leverage the Web *prosumers* (producers and consumers) to get to data integration? The scientific community shows us the road to the Web integration, i.e. moving from the current Web of Documents – made by and for people – to the Web of Data – where machines can play a crucial role in knowledge management, e.g. by advancing from a pure information retrieval (from a request to some documents that could contain useful information) to a smarter *data retrieval* (from a question to its answer).

In order to reach this objective of the Web of Data, it is clear that we are all invited to take part in this Web evolution, by exposing our data on the Web. This appears to be quite easy and natural for individuals, since it already happens with the so called "user-generated contents", which are more and more frequently annotated with metadata like the tags that can be used by machines for their processing. On the other hand, enterprises and organizations must find a straightforward way to expose data on the Web. This operation can be realized with two different approaches:

- by the *conversion* of the data source, which is translated into a suitable format for its publication on the Web; however, this solution is not always feasible, e.g. when the data source is frequently updated or its scale is so large that the translation process takes too much (in terms of time or costs) with respect to the use of the data;
- by the *wrapping* of the data source, i.e. by inserting a virtualization layer on top of the source which translates the queries and their responses from and to a Web-compliant format; several tools exist now, coming from both the scientific community and the industry [2,3,4,5,6,7,8,9].

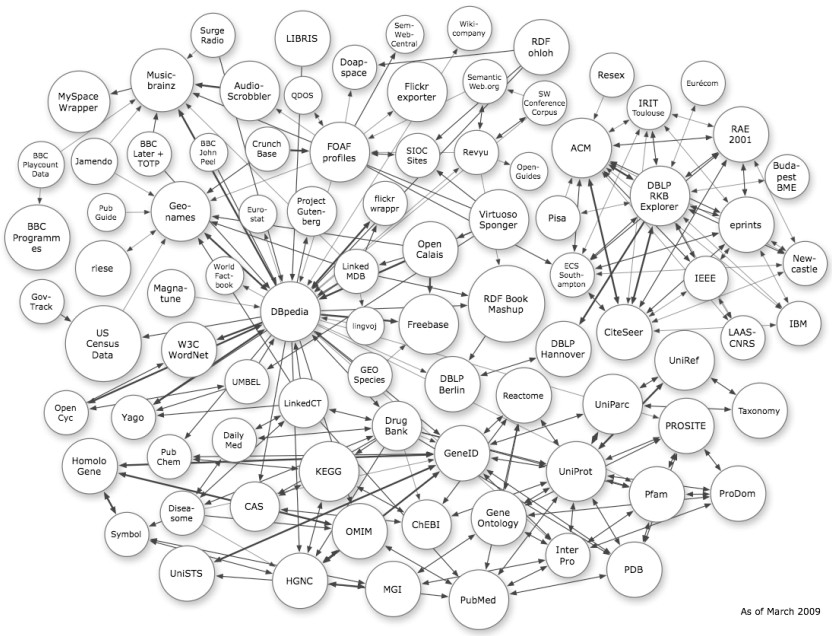

Fig. 1. The LOD cloud, as of March 2009 (source: [11])

In both cases, by publishing the data on the Web or by providing an access point to them, we are concretizing the idea of the Linked Data, as described by Tim Berners-Lee [10]. The community already recognized the value of this idea of data linked, connected to other pieces of data, aimed at forming what is sometimes indicated as Giant Global Graph (GGG); a specific initiative [11] was started to collect available data sources and to link them together as much as possible, so that the contained information can be seamlessly navigated regardless to the sources' boundaries. The so called Linked Open Dataset (LOD), depicted in Figure 1, now encompasses numerous sources of different kind and topic and reached the size of more than 4.5 billion triples.

The LOD "cloud" and its size makes immediately think about the problem of managing that scale of data: is current cloud computing technology up to the task of processing and handling the Web of Data? The scientific and industrial community must find a solution to this problem, by joining the efforts about scalable systems and Web technologies, since it is less and less a computational or storage issue but the challenge lays in data and knowledge management.

Finally, the trend behind the current popularity of linked data and the increasing availability of tools and techniques to deal with them do not mean that the research agenda of Semantic Web technology is over. Instead, several interesting challenges are still to come; among them we list the following ones:

– *Automatic linked data creation and linkage*: the automatic generation of linked data and smart mechanisms to identify "contact points" between different data sources and to seamlessly link them;

- *Distributed querying*: querying distributed data over different Web sources regardless the "physical position" of data and getting aggregated results;
- *Distributed reasoning*: applying inference techniques to distributed data, preserving consistency and correctness of the reasoning.

In the following we present two running research projects which try to address the aforementioned challenges and we foresee the future of the linked data, both from a business and a technological point of view.

2 Production of Linked Data: Service-Finder

The first question to answer when talking about the linked data is: how can we produce them? The easy generation and maintenance of machine-readable data published and accessible on the Web is the first step to take. In this chapter, we provide an example of how to derive linked data from the current Web.

2.1 Concept and Architecture of Service-Finder

The Service-Finder project [12] is addressing the problem of utilizing the Web Service technology for a wider audience by realizing a portal[1] for Web Service discovery by making Web Services available to potential consumers similarly to how current search engines do for content pages.

An essential, but mostly unaddressed problem in the area of Service Oriented Architectures (SOA) is the creation of such semantic descriptions of Web Services. Service-Finder aims to offer automatic creation of service descriptions for a different range of Services (all the publicly available services) and to enable service consumers, not just service providers, to enrich the semantic service descriptions, following a typical contribution-based approach in a Web 2.0 fashion.

The Service-Finder project delivered a Search Engine that enables users to find up-to-date information on available Web Services. By employing automated crawling and analysis techniques, the Service-Finder approach is able to scale with the increasing number of services and does not rely on a central editorial team. Consequently, Service-Finder can adapt quickly to changes in terms of the available services. The Search Engine leverages available information exposed by current technologies and extends this information with semantic annotations to allow for a more accurate retrieval. Thus, Service-Finder approaches the discovery problem by developing novel means of obtaining the underlying semantic models for discovery, by analyzing available Web content and leveraging direct and indirect user feedback on the extracted data.

The Service-Finder Portal is built on top of some core technologies and expertises brought by the project partners that act as the strategic *ingredients* for the project. Figure 2 shows the main six ingredients:

- *Web Services* are the basic paradigm and technology to implement a Service Oriented Architecture;

[1] The Service-Finder Portal is available at `http://demo.service-finder.eu/`

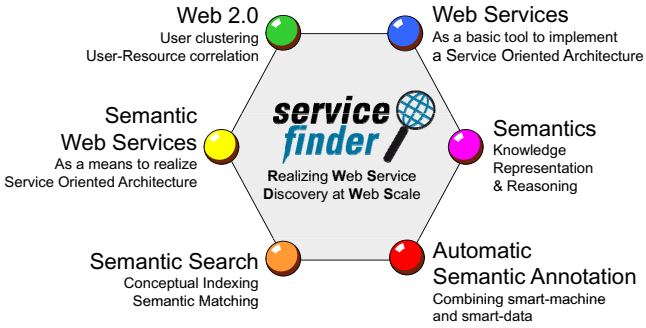

Fig. 2. Service-Finder Core Technologies

- *Semantics* provides the methodologies and tools to represent knowledge and to reason over it;
- *Automatic Semantic Annotation* is a way to enrich the gathered Web Service descriptions with semantic annotations;
- *Semantic Search* improves recall and precision by indexing Web Services related data at the conceptual level and by enabling semantic matching between Web Service descriptions;
- *Semantic Web Services* extend Web Services descriptions to easily realize Service Oriented Architecture;
- *Web 2.0* as a paradigm to involve users in the process of improving their experience with the portal.

Service-Finder combines those ingredients into one coherent architecture, depicted in Figure 3 and based on five internal components:

- The *Service Crawler* (SC) obtains the available services and related information by crawling HTML and PDF documents from the Web.
- The *Automatic Annotator* (AA) receives the crawled data and enriches it with annotations according to the Service-Finder ontology and the Service Category ontology.
- The *Conceptual Indexer and Matcher* (CIM) receives and integrates all the information into a coherent semantic model based on the ontologies and provides reasoning and querying capabilities.
- The *Service-Finder Portal Interface* (SFP) provides the user interface for searching and browsing the data managed by the CIM. It also enables users to contribute information in a Web 2.0 fashion by providing tags, categorizations, ratings, comments and wiki contributions.
- The *Cluster Engine* (CE) analyzes the users' behavior in interacting with the SFP in order to provide them with recommendations and statistics.

2.2 From Service-Finder to the Web of Data

Even if the intended result of the Service-Finder project is the realization of a Web portal for searching for Web Services, a sort of "collateral effect" directly

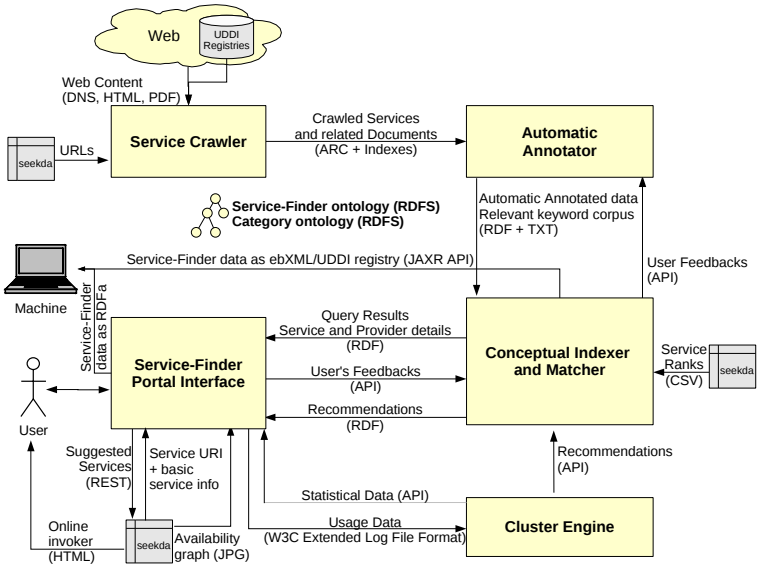

Fig. 3. The logical architecture of the Service-Finder Portal

contributes to the Web of Data. In fact, Service-Finder takes information about services from the Web, translates it into structured information describing services with regards to two domain-specific ontologies, and finally gives this information back to the community that can further enrich it.

Is this linked data? Not completely, since it is not yet in a "linked" and "Webby" format. But also this objective is close, since the Service-Finder project will provide linked data for machines in the following ways (cf. also Figure 3):

- the Service-Finder Portal will soon include *RDFa annotations* [13] in all service pages, so that GRDDL processors [14] will be able to extract service descriptions from the portal pages;
- Service-Finder will also provide *free services* (under the form of API or REST services) to query the knowledge base and get the service information directly via code; in this way, new applications will be allowed to access and exploit the "content" of the Service-Finder system;
- finally, it is possible to envisage a publication or release of a complete "*dump*" of the *Service-Finder knowledge base*; in this way, the content of the Service-Finder system could be easily published on the Web as linked data.

More information and news about Service-Finder are available at the project Web site: `http://www.service-finder.eu`.

3 Management of Linked Data: LarKC

Once the Web of Data grows and more and more data sources are turned into linked data published on the Web, the problem arises of how to manage those

data and how to exploit, in an intelligent and scalable way, the knowledge available on the Web. LarKC[2] is a running research project aimed at realizing a platform for reasoning at Web scale. In this chapter, we present a possible use of linked data in one of the project use case scenarios.

3.1 Urban Computing in the LarKC Project

Urban settings range from our own cars, while we drive them in town, to public spaces such as streets and squares including semipublic ones like cafés and tourist attractions. Urban lifestyles are even broader and include people living, working, visiting and having fun in those settings. Not surprisingly, people constantly enter and leave urban spaces, occupying them with highly variable densities and even changing their usage patterns between day and night [15].

Some years ago, due to the lack of data, solving Urban Computing problems with ICT looked like a Sci-Fi idea. Nowadays, a large amount of the required information can be made available on the Internet at almost no cost: computerized systems contain maps with the commercial activities and meeting places (e.g., Google Earth), events scheduled in the city and their locations, positions and speed information of public transportation vehicles and of mobile phone users, parking availabilities in specific parking areas, and so on.

However, current ICT technologies are not up to the challenge of solving Urban Computing problems: this requires the combination of a huge amount of static knowledge about the city (i.e., urbanistic, social and cultural knowledge) with an even larger set of dynamic data (originating in real time from heterogeneous and noisy data sources) and reasoning above the resulting time-varying knowledge. A new generation of reasoners is clearly needed. This is the purpose of the Urban Computing use case in the LarKC project.

Taking into consideration the peculiarities of urban environments, the LarKC project derived requirements [16] to be addressed by the reasoning community:

- *Coping with Heterogeneity*: data heterogeneity is a common problem for semantic technologies; we distinguish between the following heterogeneity cases:
  - *Representational Heterogeneity*, which means that semantic data are represented by using different specification languages.
  - *Reasoning Heterogeneity*, which means that the systems allow for multiple paradigms of reasoners, like temporal, spatial or causal reasoning; moreover, sometimes precise and consistent inference is needed, but in other cases approximate reasoning or imperfect estimations can be better.
  - *Default Heterogeneity*, which means that systems support for various specification defaults of semantic data; for example, closed world assumption vs. open world assumption, or unique name assumption vs. non-unique name assumption.

[2] `http://www.larkc.eu`

- *Coping with Scale*: although we encounter large scale data which are not manageable, this does not necessarily mean that all of the data must be dealt with simultaneously.
- *Coping with time-dependency*: knowledge and data can change over the time; for instance, in Urban Computing names of streets, landmarks, kind of events, etc. change very slowly, whereas the number of cars that go through a traffic detector in five minutes changes very quickly. This means that the system must have the notion of "observation period", defined as the period when the system is subject to querying.
- *Coping with Noisy, Uncertain and Inconsistent Data*: data about a urban environment can often be noisy (when a part of data is useless or semantically meaningless), inconsistent (when parts of data are in logical contradiction with each another, or are semantically impossible) or uncertain (when the data semantics is partial or incomplete).

This set of requirements clearly shows that linking data and publishing them on the Web is just the first step: a smarter and scalable processing solution is needed.

3.2 A Urban Linked Data Mash-Up in LarKC

In Figure 4, the first Urban Computing application developed on top of the LarKC platform is represented. This is not a fully-fledged Urban Computing system, but it is the first prototype running over the LarKC platform.

The scenario describes a user which is in a city (e.g. Milano) and wants to organize his Saturday night; for this reason, he wants to know what interesting places he can visit (e.g., if he is a tourist, he would like to know what monuments are open at night and are easy reachable from his place), or he would like to

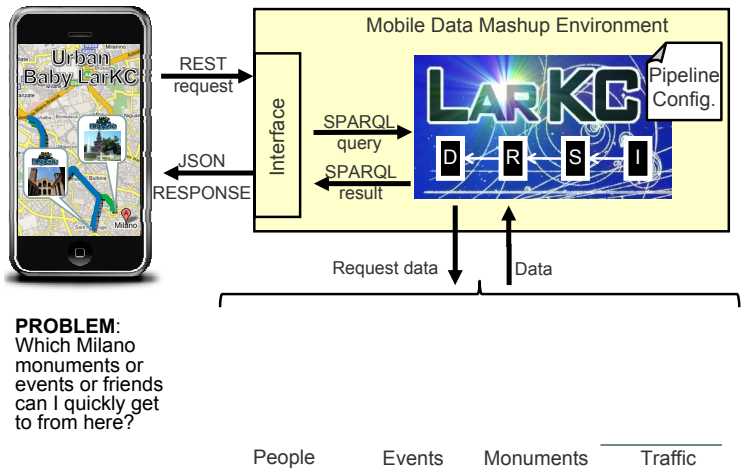

Fig. 4. A graphical representation of the scenario

attend some cultural or music event, or he wishes to meet some friends that happen to be in the same city. Moreover, since traffic in large urban environments is often a mess, he would like to know the most desirable path to his destination and maybe also where to park his car. And, last but not least, he would like to use a single application that fulfill all his needs, without having to manually make sense of the results of different services and applications.

The hypothetic user can turn to a demonstrative application – named Urban Baby LarKC to stress its experimental state – which, by using the LarKC platform, is able to access multiple data sources on the Web and to retrieve the most suitable information to give an answer to the user request. LarKC, in fact, allows for a smarter way to identify relevant data sources, to select a meaningful and useful subset of data and to apply the proper reasoning strategy.

In the scenario case, LarKC identifies the interesting data sources (e.g. an encyclopedia to get information about relevant places, a social networking site to find interesting events or the position of friends, municipality sources to get the real-time situation of traffic, and so on); then LarKC selects the relevant subset of data (e.g., it filters the encyclopedia to get information only about Milano, or events happening on a specific date, or streets only in the area surrounding the current position of the user); finally, LarKC reasons and processes the data to find the most suitable path(s) to suggest to the user.

It is clear that, by increasing the number and typology of data sources, the design of a single application that solves the complete problem becomes more difficult, and scalability, heterogeneity, real-time and noise-tolerance problems become harder to address. LarKC is designed as an extensible platform; it is designed to offer support for building "pipelines" (composite work-flows to realize applications) which invoke several "plug-ins" (reusable pieces of software dedicated to implement a specific function); it allows for data distribution, computation parallelization and so on. For those reasons, LarKC is a suitable architecture to be employed to develop Urban Computing applications that leverage the large amount of linked data that describe the various facets of a urban environment.

4 Conclusions and Outlook

The advent of the Web of Data is capable of radically changing the way we look at our data and their management. The possibilities and opportunities enabled by the linked data approach are emerging and various examples can be found not only in the research field, but also in the industry.

A notable case is *Freebase*[3]. Freebase is "an open, shared database of the world's information"; it is a collaborative effort to build an on-line collection of structured data harvested from many sources, including individual contributions. Freebase releases its content under a Creative Commons "attribution" license, and also offers to programmers an API, an RDF endpoint and a database dump. In this way, Freebase allows people (and machines) to access common information more effectively.

[3] `http://www.freebase.com`

While at the beginning it was an initiative completely disconnected from the Semantic Web community, the recent release of its RDF version[4] made Freebase a relevant part of the Web of Data.

Another interesting example is constituted by *OpenCalais*[5]. The renowned company Thomsom-Reuters released a free service called Calais, which can be used to derive linked data from documents. Unstructured text is processed by Calais, which extracts entities, facts and events; those "metadata" are returned enriched with "keys" that provide access to the Calais Linked Data cloud; in turn, this cloud provides information and other Linked Data pointers to a range of open and partner Linked Data assets, among which Wikipedia, Metaweb and Cnet, and possibly in the future also the data of Thomson-Reuter itself.

The fact that the service is free of charge and that it is in line with the linked data principles makes OpenCalais a good example of how enterprises can leverage Semantic Web technologies and both profit from them and contributing to the community vantage.

The aforementioned examples demonstrate that the linked data vision is far from being a pure academic question. On the contrary, it can represent a valid means to get to a new generation of improved solutions for the information management and for the problem of integration. Our opinion is that the Web of Data should be better explored both from the business and the technological point of view.

From a business perspective, it is incontrovertible that, being storage a commodity, organizations today are used to produce lots of data; this implies that they more and more frequently experience the problem of managing and making sense of all their data. As a consequence, they often ask for Business Intelligence solutions or turn to similar or related technologies to "understand" their data wealth and take informed decisions. However, it also happens that, when strategic decisions are needed, the data within the organization are not enough and they should be integrated or enhanced with external knowledge.

Clearly, this is a case where linked data technologies can play an important role. One obstacle that can hinder the application of linked data approaches is the claim for privacy and security of data; even if there is undoubtedly the need for better and robust solutions to preserve data confidentiality, ownership and protection, enterprises often use the security "excuse" in order to prevent others to access their data. In this way, however, they block innovation and new business opportunities. As Tim Berners-Lee effectively stated in his speech at TED 2009 [17], organizations should "stop hugging their data" and unleash the full power of linked data.

It is worth noting that, also from a pure scientific and technological point of view, new challenges wait for linked data to find a solution. For example, taking into consideration the case cited above, linked data approaches can mine Business Intelligence techniques' basic assumptions: data and data sources can dynamically change, not only because of data streaming, but also because of the

[4] `http://rdf.freebase.com/`
[5] `http://opencalais.com`

unreliability and uncertainty of archives available remotely on the Web (the so called 404 problem); moreover, the Web is intrinsically inconsistent, since you can find everything and the opposite or the negation of everything; the information on the Web can be partial, because of implicit or common knowledge or because of the relations to unavailable sources; finally, on the Web more information than expected or than needed can be found, thus the need for scalable systems or sampling/filtering approaches.

How long is the way to a large scale adoption of linked data in business environments? It is difficult to give a precise answer, but our opinion is that the goal is not far from being accomplished. Best practices and success cases are now more helpful and required than fully-fledged solutions to convince business decision makers to invest in linked data technology; accurate business plans and concrete cost/risk assessments are also needed. But the experience of the Web itself, born in small research labs and gradually spread all over the world, is an outstanding sign of the success ready and waiting for linked data.

Acknowledgments

This research has been partially supported by the *Service-Finder* (FP7-IST-215876) and the LarKC (FP7-IST-215535) EU co-funded projects. The speech entitled "From research to business: the Web of linked data" was given in Poznan on April 29th, 2009 at the joint Enterprise X.0 and Econom Workshops, co-located with the 12th Business Information Systems conference (BIS 2009). The slides are available on the Web at `http://www.slideshare.net/iricelino`.

References

1. Surowiecki, J.: The Wisdom of Crowds. Anchor Books, New York (2005)
2. Bizer, C., Cyganiak, R.: D2RQ: Lessons Learned. In: W3C Workshop on RDF Access to Relational Databases
3. OpenLink Software: Exposing SQL Data as RDF (2007)
4. Barrasa, J., Corcho, O., Gómez-Pérez, A.: R_2O, an Extensible and Semantically Based Database-to-ontology Mapping Language. In: Second International Workshop on Semantic Web and Databases (2004)
5. Cullot, N., Ghawi, R., Yétongnon, K.: DB2OWL: A Tool for Automatic Database-to-Ontology Mapping. Université de Bourgogne (2007)
6. Seleng, M., Laclavik, M., Balogh, Z., Hluchý, L.: RDB2Onto: Approach for creating semantic metadata from relational database data. In: Informatics 2007, the ninth international conference on informatics (2007)
7. de Laborda, C.P., Conrad, S.: Relational.OWL - A Data and Schema Representation Format Based on OWL. In: Second Asia-Pacific Conference on Conceptual Modelling (APCCM 2005) (2005)
8. SquirrelRDF (2007), `http://jena.sourceforge.net/SquirrelRDF/`
9. Corno, W., Corcoglioniti, F., Celino, I., Della Valle, E.: Exposing heterogeneous data sources as SPARQL endpoints through an object-oriented abstraction. In: Domingue, J., Anutariya, C. (eds.) ASWC 2008. LNCS, vol. 5367, pp. 434–448. Springer, Heidelberg (2008)

10. Berners Lee, T.: Linked Data (2006), http://www.w3.org/DesignIssues/LinkedData.html
11. Linking Open Data Initiative (2008), http://www.linkeddata.org
12. Della Valle, E., Cerizza, D., Celino, I., Turati, A., Lausen, H., Steinmetz, N., Erdmann, M., Schoch, W., Funk, A.: Realizing Service-Finder – Web Service Discovery at Web Scale. In: Proceedings of the 2nd European Semantic Technology Conference (ESTC 2008) (2008)
13. Adida, B., Birbeck, M., McCarron, S., Pemberton, S.: RDFa in XHTML: Syntax and Processing – A collection of attributes and processing rules for extending XHTML to support RDF, W3C Recommendation (2008), http://www.w3.org/TR/rdfa-syntax/
14. Connolly, D.: Gleaning resource descriptions from dialects of languages (grddl) W3C Recommendation (2007), http://www.w3.org/TR/grddl/
15. Reades, J., Calabrese, F., Sevtsuk, A., Ratti, C.: Cellular census: Explorations in urban data collection. IEEE Pervasive Computing 6(3), 30–38 (2007)
16. Della Valle, E., Celino, I., Dell'Aglio, D., Kim, K., Huang, Z., Tresp, V., Hauptmann, W., Huang, Y., Grothmann, R.: Urban computing: a challenging problem for semantic technologies. In: 2nd International Workshop on New Forms of Reasoning for the Semantic Web (NEFORS 2008) co-located with the 3rd Asian Semantic Web Conference (ASWC 2008) (2008)
17. Berners Lee, T.: The next Web of open, linked data, Speech at TED 2009 (2009), http://www.ted.com/index.php/talks/tim_berners_lee_on_the_next_web.html