JLIS.it
Italian Journal of Library and Information Science
Rivista italiana di biblioteconomia, archivistica e scienza dell'informazione

Linked (open) data at web scale: research, social and engineering challenges in the digital humanities

Michele Barbera

In the last decade, the amount of data available in the linked data cloud has grown enormously in several domains, including cultrual heritage and digital humanities. However creative reuse of data both within the scholarly community and within the cultural industry is still very limited. The limited creative reuse of data does not only depend on the limitations of existing technologies, but also on several social and cultural habits whose consequences need to be fully addressed and further researched. If linked data is to be exploited at its full potential, a profound cultural shift needs to occur in the way data is produced, managed and disseminated. This is especially true in the cultural heritage and digital humanities domains, where a strong tradition of two-dimensional, paper-like thinking is still predominant. The first section of this paper briefly presents the most pressing technological and engineering challenges to be addressed within the linked data sector. In the second section it is argued that the full exploitation of the linked data sector does not only depends on technological advancement but also on the possibilities enshrined in a radical cultural change in thinking about the semantic web and linked open data visions. The potential effect

JLIS.it. Vol. 4, n. 1 (Gennaio/January 2013).
DOI: 10.4403/jlis.it-6333

on the cultural industry and on the scholarly community is also explored. The importance of nurturing a lively business ecosystem and the role of public and open data published by GLAM[1] organizations is the main condition enabling the linked data vision to take off. Despite it is here impossible to provide a comprehensive analysis of potential of the semantic web, its enabling conditions and implications, this paper nonetheless aims at offering a stimulating insight into one of the possible ways of thinking about it.

Technical and engineering challenges

The web of data is characterized by its universal nature, its virtually infinite size, and by the heterogeneity of data. It comes at no surprise that these and many other features have influenced the way in which data producing and data consuming applications are -– or ought to be -– designed. First of all, due to the rapidly growing amount of data available in the linked open data cloud and in enterprise linked data repositories, it is not possible to centralize and compute all the data in a single local repository. The largest existing public repository, Sindice.com[2] holds today around 80 billion triples, which is just a fraction of the LOD Cloud. Sindice.com is based on a large map-reduce cluster (implemented on Apache Hadoop) whose TCO is still beyond the possibilities of most small and medium enterprises (SMEs) and research organizations.

One of the most significant features of the linked data vision is its capacity to find novel ways to exploit unexpected information and links to discover new insights from data. A way of *slicing* large datasets and reduce them down to a manageable size – possibly on demand – is necessary in order not to sacrifice this desirable feature.

[1] Galleries, Libraries, Archives and Museums.

[2] http://sindice.com.

This approach is not entirely new and it is sometimes refereed as dataspaces in the relational database community. Linked dataspaces are essentially a transient materialization (or a *view*) computed on demand of a slice of the originating data graph. Despite some implementations of linked dataspaces do already exist, neither standard specifications nor reference implementations have been defined yet.

Streaming linked data

Streaming linked data has been only partially addressed by the research community[3] (Barbieri and Della Valle; Le-Phuoc et al.; Sequeda and Corcho) and almost entirely neglected by production-grade industrial systems. At a first sight, this may appear marginally relevant but it is instead of paramount importance considering the growing amount of live streaming data produced by sensor networks and sensors embedded in personal mobile devices.

Versioning

From the capacity of versioning RDF graphs comes the possibility of identifying evolutionary atomic changes and to roll them back, in order to revert the graph to a previous state. Although some approaches have already been explored in research, efficient and production-ready industrial implementations in commercially or open source graph stores[4] are still under-developed. In the web of data, data sources are highly heterogeneous. The capacity of slicing and mixing different sources, that have various degree of trust (e.g. think about crowsdourced data vs. authoritative national library

[3] LarKC: The Large Knowledge Collider, http://www.larkc.eu.

[4] The availability of Open Source software of high quality is a very important element in the growth of this field of study.

data) and frequent updates a core feature. Hence, the lack of solid versioning systems is an important gap to fill as soon as possible.

Despite the importance of overcoming such limitations, technological innovation is a necessary but not sufficient condition for exploiting the potential of the semantic web. As the next section shows, this must come along with a critical understanding of the cultural shifts which are inner to the Semantic vision itself.

Social challenges

In a famous blog post, written in 2007, ("Giant Global Graph") introduced the concept of "thinking in the graph" and the notion of a "giant global graph" as opposed to the existing web of documents. The most striking changes whose effects are not yet being fully internalized in the digital publishing sector, are well summarised by Berners-Lee who writes: "The less inviting side of sharing is losing some control. Indeed, at each layer – Net, Web, or Graph – we have ceded some control for greater benefits", and "It is about getting excited about connections, rather than nervous". In order to understand the importance of these changes, in the following, three issue are considered: the AAA principle, the graph nature of the giant global graph and the Open World Assumption (OWA).

One of the pillars underpinning the vision of the semantic web and linked data – which also holds true for the web itself – is the principle known as AAA, which stands for *Anyone can say Anything, Anywhere*. This implies a profound shift in the paradigm dominating the publishing and media industry. In the era of the web of documents – and prior to the advent of the web – the dominant conceptual framework rested upon individuals acting as information producers, publishers (gatekeepers) or distributors. Information consumers (e.g. readers), acting as passive actors, were neither involved in the production nor in the distribution of information. As a result of the cultural and technological revolution enshrined in the web 2.0, readers have become active producers of information. Yet, the dissemination and often the production of information, as well as the gatekeeper role of filtering, was in the hands of publishers or aggregators of information. The web of data– which is here used as a synonym for semantic web and linked data – and the prominent role of the AAA principle in shaping publishing and consumption models,including query federation across multiple repositories, imply shifting the core activity of the information lifecycle from the production to the mesh-up of several heterogeneous and *personalized* data sources. In this novel scenario, individuals and organizations play at the same time the roles of information producers, gatekeepers, and consumers of information in an ever-reconfiguring ecosystem. In the traditional publishing world, both in the public and in the private sector, on the web and in other media, information (and data) are modeled in order to maximize the accessibility and especially the usability for consumers. This is achieved by anticipating scenarios in which information is consumed. By design, we cannot know in advance how data will be used, combined, enriched and repurposed to produce information goods. This is one of the premises that makes linked data so powerful. However, habits are difficult to change especially when they affect not only the production of data but also its consumption (e.g. think about the need for data consuming applications to be able to deal with unexpected data). Additionally, without knowing in advance how data will be consumed, data have to be represented in the most universal way possible. Such a universal representation does not allow any optimization in the information design phase and leaves room for optimizing data usage only in the consumption phase.

Secondly, the graph nature of the giant global graph has profound social implications in the way in which information is produced and consumed. Thanks to the great success of relational database technologies and spreadsheets, people mentally model data in tabular structures. IT developers and programmers, as well as information and data managers, scientists, marketers, educators and other actors involved in the data lifecycle think about data in a purpose-specific context and from an individual point of view.

Thinking in the graph as Berners-Lee puts it, is not an easy task when data is modeled, produced, aggregated or consumed. As my colleague Gradmann often remarks, this is even more difficult in the cultrual heritage and digital humanities communities, where there is a strong tradition of two-dimensional thinking derived from the paper-world. The two-dimensional paper-world approach has often been mimicked rather than revolutionized in the digital world. The problem is once again aggravated by technological constraints. First of all, most of the linked data nowadays published in the Cultural LOD Cloud is semi-automatically generated from legacy tabular data repositories. Secondly, many computer science and information design courses in universities are still mostly based on tabular, relational and tree-like data structures. Another source of misunderstandings is the use of URIs or IRIs to identify at the same time informational and non-informational resources, that is seldom accepted and understood, even within the experts community (see for example the recurring discussion about http-range-14). The issue known as the *HTTP Range 14 problem*, is about what mechanisms should be used to distinguish between statements about web pages and statements about the real world item or concepts the web page *talks about*. Along with some technical and engineering implications, what is interesting is that the distinction between information and non-information resources is not always clear for data publishers and data consumers. As a result, the web of data and many data-consuming applications present inconsistencies which derive from the misinterpretation of this concept. This is not a critical problem *per-se*, but it becomes a serious problem when reasoning is applied to these inconsistencies (e.g. same-as reasoning). Finally, the web of data is built upon the Open World Assumption (OWA)[5], according to which ”the truth-value of a statement is independent of whether or not it is known by any single observer or agent to be true. In other words, a statement cannot be considered false just because there is nothing explicitly stating that it is true. This is another important shift from the relational database world, which is based on the Closed World Assumption. The logical and the technical implications of the OWA are beyond the scope of this presentation, however it is worth spelling out some of its social consequences. The choice of operating under OWA is justified by the fact that open worlds are particularly well suited to deal with incomplete information and exceptions. OWA's features are desirable within a universal systems such as the web. However, they also pose some serious challenges. For instance, some problems are inherently related to closed worlds and most importantly many of the IT tools are designed to work in closed worlds. Furthermore, people are more familiar with thinking in closed worlds rather than in open ones. Once it has been shown which are the intrinsic cultural implications of the vision of the semantic web and linked data, the next section introduces another *problematique* which is essential for advancing a revolutionary twist in the semantic web: the importance of nurturing a dynamic linked data economy.

[5] Open World Assumption, in Wikipedia, http://en.wikipedia.org/wiki/Open_world_assumption.

Nurturing a linked data economy

In the last decade, Europe has made large investments in research over semantic technologies. This has generated brilliant ideas, core scientific knowledge and many prototypal implementations. Unfortunately, the research community has not yet been able to leverage this potential within the industry to build production-ready tools easily usable by end-users. There is not yet a Microsoft Excel, or an Apple ITunes for linked data. Similarly, there is not yet MySPARQL or any Apache HTTPD that can serve streamed linked data. A lively data economy, with a rich ecosystem, is not yet in place. The time has come to invest in innovation in order to be able to transform the enormous knowledge accumulated through research and the large amount of data recently produced/liberated into a virtuous circle able to generate a self-sustaining and evolving ecosystem. Recently a number of game-changing announcements has been made which can be considered as potentially contributing to create a linked data economy: first of all, Google Knowledge Graph, a sort of Closed Enterprise linked data cloud as well as the acquisition by the big G of one of the most important nodes of the LOD cloud, Freebase; secondly, the coalition between the largest search engines to introduce schema.org, a combination of a technology and a set of incentives for web publishers to annotate their content with semantic markup. Finally, large private organizations are approaching the web of data, by evolving their business models or by modifying their production processes to comply with the openess of the linked open data cloud, or by building closed enterprise linked data clouds (e.g. many large pharma are bolding their own enterprise linked data). On the one hand some of these announcements may raise some socio-economic issues related to the risk of endangering the public good and to monopolistic threats (see for example the interesting, if a bit outdated, analysis of the risks related to schema.org[6] published by (Tennison) on her blog.

In Europe, there are numerous small-medium Enterprises which are the major driver for innovation and economic growth. A careful strategy to protect our common knowledge-heritage and the (linked!) public good that is at the same time able to offer the right economic incentives to SMEs, is key in paving the way to a vibrant and sustainable linked data economy. On the other hand, this is clearly a huge leap forward for the web of data, whose economic value may start to unlock thanks to the critical mass (of users, investments, technology, media visibility and demand) mobilized around the leading web companies. In a recent post (Dodds) suggests that the media-hype created by Google's KnowledGraph, – that is still mostly fed by public domain and open knowledge – may represent an opportunity for SMEs which can leverage the same public goods to meet the increasing demand of vertical and custom enterprise linked data clouds.

Conclusions

After having presented some of the technological challenges for a full exploitation of the linked open data web, this paper has argued that such innovation must come along with a new understanding of the cultural changes inner to the web of data. This is, however, not enough. The web of data needs a lively economic environment where to flourish and further develop around its potential. Yes, but how? The policy advices suggested by this paper are well expressed by the European Union in the following quote which concludes this contribution:

[6] http://schema.org.

> The volume of data being digitally stored and exchanged is growing exponentially. [...] Obviously, these data generate the potential for many new types of products and services. The accessibility of public services can be improved for open and linked data, smart traffic and cities can improve mobility, products can report their life cycles, monitoring their provenance and quality, social trends can be recognized and turned into services, and products can come closer to meeting consumers' needs. We foresee a whole new industry implementing services on top of large data streams. The impact of this emerging economic sector - the data economy - may soon outrange the current importance of the software industry. The gist of the matter is to turn large streams of data into added value for the public and private sector. This industry can help to increase the efficiency of processes working with these data, it can provide transparency, support well-informed decision making, and enable new services not possible today (e.g., smart cities, interactive trend analysis or seamless data flows along value creation chains). Clearly, research, engineering, policy making for the Data Economy and the exploitation of the unprecedented wealth of data have become keys to the Future of Europe.[7]

References

Barbieri, Davide and Emanuele Della Valle. "A Proposal for Publishing Data Streams as Linked Data - A Position Paper". *Proceedings of the Linked Data on the Web (LDOW2010) Workshop, co-located with WWW2010*. 2010. (Cit. on p. 93).

Berners-Lee, Tim. "Giant Global Graph". http://dig.csail.mit.edu/breadcrumbs/node/215.

Dodds, Leigh. "Welcome to the Knowledge Graph". http://talis-systems.com/2012/05/welcome-to-the-knowledge-graph/.

[7] http://2012.data-forum.eu/about.

Sequeda, Juan F. and Oscar Corcho. “Linked Stream Data: A Position Paper”. *Proceedings of the 2nd International Workshop on Semantic Sensor Networks (SSN09)*. Washington DC, USA, 2009. (Cit. on p. 93).

Tennison, Jeni. “Schema.org and the Responsibility of Monopoly”. http://www.jenitennison.com/blog/node/157.

MICHELE BARBERA, Net7; Spazio Dati.
info@netseven.it

Barbera, M. ”Linked (open) data at web scale: research, social and engineering challenges in the digital humanities”. *JLIS.it*. Vol. 4, n. 1 (Gennaio/January 2013): Art: #6333. DOI: 10.4403/jlis.it-6333. Web.

ABSTRACT: The amount of data available in the linked data cloud has grown enormously in the last years in several domains, including cultural heritage and digital humanities. However creative reuse of data both within the scholarly community and within the cultural industry is still very limited. It depends on a mixture of technical and social problems that needs to be addressed in research and within the industry. The paper will explores some of these challenges with a focus on the digital humanities.

KEYWORDS: Library linked data; Semantic web

Submitted: 2012-06-12
Accepted: 2012-08-31
Published: 2013-01-15