Federating Distributed Social Data to Build an Interlinked Online Information Society

Alexandre Passant[1], Matthias Samwald[1,2], John G. Breslin[1,3], Stefan Decker[1]

[1] Digital Enterprise Research Institute, National University of Ireland, Galway
firstname.lastname@deri.org
[2] Konrad Lorenz Institute for Evolution and Cognition Research, Austria
samwald@gmx.at
[3] Department of Electronic Engineering, National University of Ireland, Galway
john.breslin@nuigalway.ie

Abstract. While research on the relationship between the Semantic Web and social media was originally motivated by the lack of semantics in mainstream Web 2.0 services, this vision can go much further, impacting society at large in terms of how information is shared, interlinked and managed on the Web. In this paper, we will demonstrate how semantic technologies can be applied to social media, thereby creating a Web where data is socially created and maintained through end-user interactions, but is also machine-readable and therefore open towards sophisticated queries and large-scale information integration. In particular, we will emphasise the impact of what we term "Social Semantic Information Spaces" on the creation of an Interlinked Online Information Society, where any social data is a component in a worldwide collective intelligence ecosystem.

Key words: Social Semantic Web, Linked Data, SIOC, Collective Intelligence

1 Introduction

In *Weaving the Web*, Tim Berners-Lee mentions how the Web can lead to *social machines*, where computers help us to achieve our goals [1], a proposal that follows earlier visions such as Vannevar Bush's Memex or Doug Engelbart's work. Some time has passed since then, and we believe we now have the components to make this vision a reality. On the one hand, the Web 2.0 meme introduced new ways to let people share and build data collectively. On the other hand, the Semantic Web [3] provides the means to represent data in an interoperable and machine-readable way. In this paper, we emphasise how these two fields, which were often mistakenly considered as disjoint, can be linked together to lead the Web towards a medium where any data regarding a particular topic becomes an atom of knowledge that can be instantaneously queried, reused and combined with other pieces of knowledge to increase its global value.

First, we will introduce "Social Semantic Information Spaces" (SSIS) and the requirements for their successful implementation, in particular how *lightweight* semantics are important for their realisation. Then we will describe two of our current research areas where SSISs have been efficiently deployed: (1) for adding and leveraging semantics in Enterprise 2.0 environments; and (2) for Health Care and Life Sciences knowledge exchange between researchers. We will go on to describe how other data sources can be taken into consideration in these SSISs as well as how various SSISs can be linked to build an Interlinked Online Information Society of people, machines and knowledge. Finally, we will show how our work fits within the Web Science agenda and how it can help to solve some of its relevant issues.

2 Social Semantic Information Spaces

2.1 "A little semantics goes a long way"

As defined in [4], Social Semantic Information Spaces (SSIS) bridge the gap between social connectivity and semantic technologies (Fig. 1). Therefore, to be successful they require (1) people sharing and building data collectively, thanks to well-known services and tools such as blogs, wikis, bulletin boards or social networks, and (2) a layer of semantics to model both user activities and user-generated content in an interoperable way. The success of the first may depend

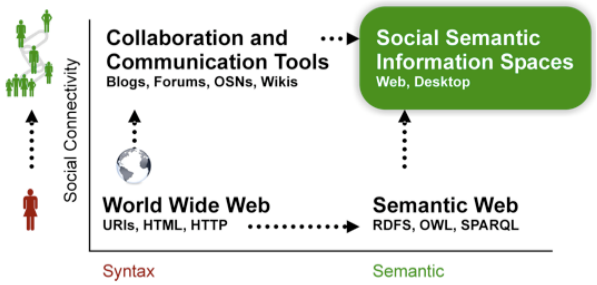

Fig. 1. Social Semantic Information Spaces

on various factors such as how object-centered sociality[4] is taken into account in these services. Regarding the semantics, two combined levels are needed to efficiently build these SSISs. First, semantics regarding the structure of the communities and the content resulting from their social interactions (*e.g.* blog posts). Ontologies such as FOAF and SIOC [5] are clearly appropriate as they provide lightweight but sufficiently powerful semantics to model theses communities: people and acquaintance networks are represented using FOAF, and content and community interactions using SIOC. Second, semantics regarding the data itself is required, *i.e.* the facts contained in such user-generated content. This can be achieved thanks to domain ontologies (that can be modeled using RDF(S)/OWL) or taxonomies (for example, using SKOS) as well as related knowledge bases, but also thanks to background knowledge in data provided by the Linking Open Data initiative[5].

Since tools dedicated to data exchange and collaboration (blogs, wikis, etc.) require minimal effort from end users, the semantic layer provided by these applications must also be deployed with as little effort as possible. As [8] mentioned, semantic blogging services should provide semantically-enhanced data without any additional input. To build these SSISs, we extended this requirement to any socially-aware service by developing various exporters that automatically produce SIOC data from major services[6]. Regarding the content itself, efforts such as semantic wikis can be efficiently used, as we will see later, since they offer ways to collaboratively build and maintain knowledge bases.

Indeed, to completely understand the potential impact of these SSISs on society, more than their technical aspects, we must keep in mind how social media has changed information sharing principles. While its most visible aspect resides in mainstream and leisure-oriented Web 2.0 services such as data sharing or online social networking, it has introduced new paradigms in enterprise information management as well as in scientific data exchange and publishing (both of which we will now focus on).

2.2 Social Semantics for Enteprise 2.0

Enterprise 2.0 [9] is considered as *"the use of emergent social software platforms within companies, or between companies and their partners or customers"*[7]. Many companies have moved from classical top-down architectures to Enterprise 2.0 information systems, using blogs, wikis, etc. While it can ease the process of publishing information, retrieving it can be a costly task. Introducing such tools to an organisation can lead to information fragmentation issues similar to those on the Web, *i.e.*, content about a particular object (such as a project or a partner) can be spread across different blogs, wikis, RSS feeds. This makes it difficult to get a global view of this object. In addition, the use of tagging leads to issues when retrieving information, as experts use keywords that are sometimes difficult to identify by non-experts, because of different "basic levels" of knowledge, depending on the cognitive background of each person [14].

In a recent use-case in which we faced these issues, aditional semantics were used for the enhancement and integration of Enterprise 2.0 components[8]. We researched and deployed a complete but lightweight social semantic stack for Enterprise 2.0 [10] consist-

[4] `http://www.zengestrom.com/blog/2005/04/why_some_social.html`
[5] `http://linkeddata.org`
[6] `http://sioc-project.org/applications`
[7] `http://andrewmcafee.org/blog/?p=76`
[8] `http://www.w3.org/2001/sw/sweo/public/UseCases/EDF/`

ing of: (1) SIOC (the Semantically-Interlinked Online Communities ontology as referenced earlier), (2) semantic wikis (in this case, a particular prototype extending the system already in use and using structured forms mapped to lightweight ontologies), and (3) MOAT[12] (a process that allows to link tags to ontology instances for semantic indexing purposes).

Using this additional stack, more than 300 instances of domain ontologies were created and maintained through these wikis, along with more than 17000 instances of `sioc:Post` (and related subclasses) linked to the previous instances, the whole ecosystem being used as an interoperability layer on the top of existing tools, weaving SSIS into corporate environments. Thanks to this combination of semantics for (1) creating and maintaining instances of domain ontologies and (2) uniform-representation of user-generated content, people were able to find items related to very specific topics (*e.g.*, *Thin-film solar cell*) by searching for a broader one (*e.g.*, *Solar energies*) wherever this content may come from (blogs, wikis, RSS feeds ...). Another important focus of this use-case is how external data has been reused to build low-cost semantic geolocation mash-ups (Fig 2). By reusing in one infrastructure machine-readable data created by other people, it clearly shows how an universal access to open data sources – provided thanks to Semantic Web technologies – can be used to enhance information and put it into context, hence being more valuable for the end-users.

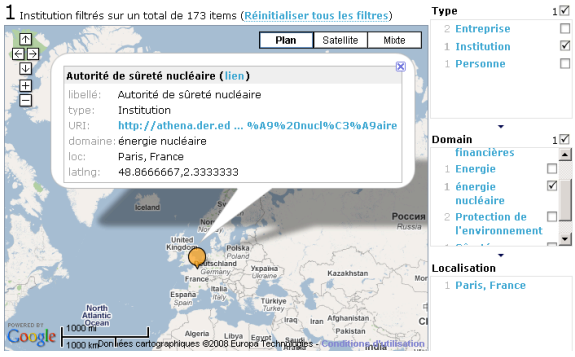

Fig. 2. A semantic mash-up combining internal and external RDF data

2.3 Scientific Knowledge Management

Biomedical research is one of the first knowledge domains where application of SSIS in real-world use cases has started to emerge. It is a good environment for the emergence of SSIS, for several reasons. First of all, biomedical research has a significant demand for information technology to help integrate the massive amounts of data from many different subdisciplines and globally distributed research groups. The universe of discourse is very large, and contains a plethora of named entities (proteins, genes, organisms et cetera). Second, the biomedical domain is distinguished from most other domains in that a large collection of well-structured ontologies and terminologies are readily available to create SSIS with rich semantics. Many of these ontologies are available in OWL format in the Open Biomedical Ontologies repository [13].

One example of a SSIS in the field of biomedical research is the Alzheimer's Research Forum[9]. It contains a collection of hypotheses about Alzheimer's disease, formulated by various participants from the research community. The hypotheses are captured with the SWAN discourse ontology [6]. Through a newly created mapping of SWAN to the SIOC vocabulary[10], the contents of the SWAN-based information space can be interwoven with other SIOC-enabled information spaces, such as scientific blogs.

However, as in Enterprise 2.0 ecosystems, the barriers to the uptake of SSIS-enabled systems are not only technical, but also institutional and social – the proficiency of a scientist being measured by the number and quality of his publications in classical papers. In our Enterprise 2.0 use-case, we noticed that combining top-down and bottom-up strategies were sucessful to increase the adoption of novel services. In the domain of scientific research, uptake of SSIS by mainstream publishers and integration into the scientific publication process is crucial for their widespread success among scientists. The increasing uptake of social software (such as Nature Connotea[11]) and growing

[9] `http://hypothesis.alzforum.org`
[10] `http://esw.w3.org/topic/HCLSIG/SWANSIOC`
[11] `http://www.connotea.org/`

efforts to capture the semantics in biomedical publications (such as the Structured Digital Abstracts project[12] and the SWAN/SIOC effort mentioned earlier) are promising current developments.

3 Extending and Interlinking SSISs

3.1 Integrating with Other Sources of Data

While SSIS currently focus on Web-based data, it is important to consider that semantically-enriched and socially-aware data can also be produced by other means. For instance, desktop data can be integrated, thanks to projects such as Nepomuk[13]. Moreover, we will consider integrating more dynamically-created data, for instance data from popular services such as microblogging, hence emphasizing on a more ubiquitous and multi-devices aspect of SSIS, as well as other sensor data from mobile phones, GPS devices and other kinds of sensors that people want to share. Hence, the Web becomes an hub of ubiquitous sociality, while scale and dynamics would then be further challenges that will have to be taken into account.

3.2 Connecting the Dots

While the two previous examples are based on specific use-cases, they use a similar methodology that can be applied in any project involving social interactions; SSIS can thus be instantiated in any community that shares and collectively builds information. Yet, one important aspect is how these SSIS can be linked together, in order to provide not only isolated semantic ecosystems, but linked information to achive the goal of an Interlinked Online Information Society.Three main strategies can be used to improve linkage between various SSIS. (1) *Using social relationships*: by representing distributed social networks and user-interactions in an interoperable way thanks to the lightweight vocabularies described before, one's identities and networks can be spread among different SSIS instead of living in closed data-silos. (2) *Using content types*: by representing data following shared, lightweight schemas, notably using SIOC wich is now widely adopted, as showed by its recent integration in Yahoo! SearchMonkey, the same representation format apply to any SSIS, hence being linked by these unified content types. (3) *Using topics*: by reusing common identifiers (*e.g.*, DBpedia URIs) for defining topics within SSIS thanks to semantic enrichments frameworks for tagging systems, hence enabling topic-based interlinking.

Connecting such isolated Information Spaces also enriches the values of the global network, as Metcalfe's law states[14], by interlinking more people, machines and data. By connecting the dots, data from one SSIS can also be efficiently used in another one, as we exemplified with the mash-up example. Yet, while it provides interlinking at Web scale, new question will arise, especially regarding how to usefully exploit this amount of data.

4 Relationships with the Web Science agenda

In [2], Web Science is described as *"a science that seeks to develop, deploy, and understand distributed information systems, systems of humans and machines, operating on a global scale"*. As we explained in this paper, our approach of SSIS aims to develop and deploy such systems, in which the machine aims to help people to better collaboratively build knowledge and efficiently reuse it. More generally, we also believe that SSIS – and the Semantic Web as a whole, for which SSIS aims to solve the *chicken and egg* problem – are a way to make the process of the study and understanding of these systems easier with standard representation formats.

Moreover, one of the various challenge of Web Science identified in [7] is defined as follows: *"How can we extend the current Web infrastructure to provide mechanisms that make the social properties of information-sharing explicit and guarantee that the use of this information conforms to relevant social-policy expectations?"* We believe that the proposed approach is an interesting solution to that issue, since

[12] `http://www.febsletters.org/content/sda_summary`
[13] `http://nepomuk.semanticdesktop.org/`
[14] `http://en.wikipedia.org/wiki/Metcalfe's_law`

it does not imply changes regarding the Web architecture [15], but – as we detailed in this paper – require a lightweight layer of semantics to make these social interaction explicit and machine-understandable. Furthermore, as we recently discussed [11] we believe that this amount of interlinked data will not be an issue for privacy but on the contrary will help to provide advanced social policies for trust and access control on the Web.

Conclusion

In this paper, we explained how Social Semantic Information Spaces can be used to build a network of people and computers, aiming to achieve the vision of *social machines*. We showed how they can be deployed in various environments and how various ecosystems of semantically-enriched social data could be linked together to provide an Interlinked Information Society. While some may call it Web 3.0 or n.0, its goal is actually close to the initial vision of the Web, *i.e.* social, open and machine-readable, bringing it to its full potential.

Acknowledgements

The work presented in this paper has been funded in part by Science Foundation Ireland under Grant No. SFI/08/CE/I1380 (Lion-2).

References

1. Tim Berners-Lee and Mark Fischetti. *Weaving the Web: The Original Design and Ultimate Destiny of the World Wide Web by its Inventor*. Harper Collins Publishers, New York, 1999.
2. Tim Berners-Lee, Wendy Hall, and Nigel Shadbolt. The Semantic Web Revisited. *IEEE Intellligent Systems*, 21(3):96–101, May/June 2004.
3. Tim Berners-Lee, James A. Hendler, and Ora Lassila. The Semantic Web. *Scientific American*, 284(5):34–43, 2001.
4. John G. Breslin and Stefan Decker. Semantic Web 2.0: Creating Social Semantic Information Spaces, 2006. Tutorial at the 15th International World Wide Web Conference (WWW2006).
5. John G. Breslin, Andreas Harth, Uldis Bojārs, and Stefan Decker. Towards Semantically-Interlinked Online Communities. In *Second European Semantic Web Conference, ESWC 2005*, 2005.
6. Paolo Ciccarese, Elizabeth Wu, Gwen Wong, Marco Ocana, June Kinoshita, Alan Ruttenberg, and Tim Clark. The SWAN biomedical discourse ontology. *J. of Biomedical Informatics*, 41(5):739–751, 2008.
7. James Hendler, Nigel Shadboldt, Wendy Hall, Tim Berners-Lee, and Danny Weitzner. Web Science: An interdisciplinary approach to understanding the World Wide Web. *Communications of the ACM*, 2008.
8. David R. Karger and Dennis Quan. What Would It Mean to Blog on the Semantic Web? In *ISWC 2004: Third International Semantic Web Conference*, 2004.
9. Andrew P. Mcafee. Enterprise 2.0: The Dawn of Emergent Collaboration. *MIT Sloan Management Review*, 47(3):21–28, 2006.
10. Alexandre Passant. *Semantic Web technologies for Enterprise 2.0*. PhD thesis, 2009. To appear.
11. Alexandre Passant, Philipp Kärger, Michael Hausenblas, Daniel Olmedilla, Axel Polleres, and Stefan Decker. Enabling Trust and Privacy on the Social Web. In *W3C Workshop on the Future of Social Networking*, 2009.
12. Alexandre Passant and Philippe Laublet. Meaning Of A Tag: A collaborative approach to bridge the gap between tagging and Linked Data. In *Proceedings of the WWW2008 Workshop Linked Data on the Web (LDOW2008)*, 2008.
13. Barry Smith, Michael Ashburner, Cornelius Rosse, Jonathan Bard, William Bug, Werner Ceusters, Louis J. Goldberg, Karen Eilbeck, Amelia Ireland, Christopher J. Mungall, Neocles Leontis, Philippe R. Serra, Alan Ruttenberg, Susanna A. Sansone, Richard H. Scheuermann, Nigam Shah, Patricia L. Whetzel, and Suzanna Lewis. The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration. *Nature Biotechnology*, 25(11), 2007.
14. James W. Tanaka and Marjorie Taylor. Object categories and expertise: Is the basic level in the eye of the beholder? *Cognitve Psychology*, 23(3):457–482, 1991.
15. W3C Technical Architecture Group. Architecture of the World Wide Web, Volume One. W3C Recommendation 15 December 2004, World Wide Web Consortium, 2004. `http://www.w3.org/TR/webarch/`.