Federating Distributed Social Data to Build an Interlinked Online Information Society

Alexandre Passant, Matthias Samwald, John G. Breslin, and Stefan Decker,
National University of Ireland, Galway

Applying semantic technologies to social media can result in an interlinked online information society where social data becomes part of a worldwide, collective intelligence ecosystem.

In *Weaving the Web,* Tim Berners-Lee suggested that the Web could lead to social machines, to computers that help us achieve our goals—a proposal that followed earlier visions such as Vannevar Bush's memex and Doug Engelbart's work. Some time has passed since Berners-Lee's proposal, 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 provides the means to represent data in an interoperable and machine-readable way. These two fields, often mistakenly considered disjoint, can be linked to lead the Web toward a medium in which any data about 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.

In this article, we introduce *social semantic information spaces* (SSISs) and describe the requirements for their successful implementation—in particular, how lightweight semantics are important for their realization. We have efficiently deployed SSISs in two of our current research areas:

- to add and leverage semantics in Enterprise 2.0 environments, and
- to support healthcare and life sciences knowledge exchange between researchers.

SSISs can take other data sources into consideration, and various SSISs can be linked to build an interlinked online information society of people, machines, and knowledge. We will show how our work fits within the Web science agenda and how it can help to solve some of this agenda's relevant issues.

Social Semantic Information Spaces

As defined in an earlier publication,[1] SSISs bridge the gap between social connectivity and semantic technologies (see Figure 1). To be successful, they require two elements:

- people sharing and building data collectively, using well-known services and tools such as blogs, wikis, bulletin boards, and social networks; and

1541-1672/09/$26.00 © 2009 IEEE
Published by the IEEE Computer Society

- a layer of semantics to model both user activities and user-generated content in an interoperable way.

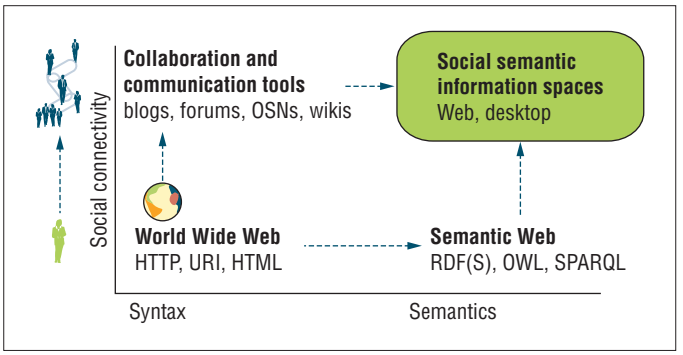

Figure 1. Social semantic information spaces (SSISs). SSISs bridge the gap between social connectivity and semantics by combining tools and paradigms from the former and languages and technologies from the latter.

The success of the first element can depend on various factors, such as how these services take object-centered sociality into account (www.zengestrom.com/blog/2005/04/why_some_social.html). Regarding the semantics, we need two combined levels to build SSISs efficiently. First, we need semantics regarding the structure of the communities and the content resulting from their social interactions (such as blog posts). Ontologies such as FOAF (Friend of a Friend) and SIOC (Semantically Interlinked Online Communities) are clearly appropriate because they provide lightweight but sufficiently powerful semantics to model these communities:[2] FOAF can represent people and acquaintance networks, and SIOC can represent content and community interactions. Second, we need semantics regarding the data itself—that is, the facts contained in such user-generated content. We can achieve this using

- domain ontologies—modeled using RDFS (Resource Description Framework Schema) and OWL (Web Ontology Language);
- taxonomies—using, for example, SKOS (Simple Knowledge Organization System);
- related knowledge bases; and
- background knowledge in data provided by the Linking Open Data initiative (http://linkeddata.org).

Because tools dedicated to data exchange and collaboration (blogs, wikis, and so on) require minimal effort from end users, the semantic layer that these applications provide must also be deployed with as little effort as possible. As David Karger and Dennis Quan mentioned, semantic blogging services should provide semantically enhanced data without any additional input.[3] To build SSISs, we extended this requirement to any socially aware service by developing various exporters that automatically produce SIOC data from major services (http://sioc-project.org/applications). Regarding the content itself, efforts such as semantic wikis can be efficiently used, as we will see later, because they offer ways to collaboratively build and maintain knowledge bases.

Indeed, to completely understand the potential impact of SSISs on society, more than their technical aspects, we must keep in mind how social media has changed information-sharing principles. Although social media's most visible aspect resides in mainstream and leisure-oriented Web 2.0 services such as data sharing and online social networking, it has introduced new paradigms in fields such as enterprise information management as well as in scientific data exchange and publishing.

Social Semantics for Enterprise 2.0

Enterprise 2.0 is "the use of emergent social software platforms within companies, or between companies and their partners or customers."[4] Many companies have moved from classical top-down architectures to Enterprise 2.0 information systems, using blogs, wikis, and so on. Although Enterprise 2.0 can ease the process of publishing information, retrieving it can be costly. Introducing such tools to an organization can lead to information fragmentation issues similar to those on the Web—that is, content about a particular object (such as a project or a partner) can be spread across numerous blogs, wikis, and RSS feeds. This makes it difficult to get a global view of the object. In addition, tagging can lead to problems for information retrieval, because experts use keywords that are sometimes difficult to identify by nonexperts because of different basic levels of knowledge, depending on each person's background.[5]

In a recent use case in which we faced these issues, we used additional semantics to enhance and integrate Enterprise 2.0 components (www.w3.org/2001/sw/sweo/public/UseCases/EDF). We researched and deployed a complete but lightweight social-semantic stack for Enterprise 2.0, consisting of these elements:

- SIOC—the ontology mentioned earlier;
- semantic wikis—in this case, a particular prototype extending the system already in use and using structured forms mapped to lightweight ontologies; and
- Meaning of a Tag (MOAT; http://moat-project.org)—a process that allows the linking of tags to ontology instances for semantic indexing.[6]

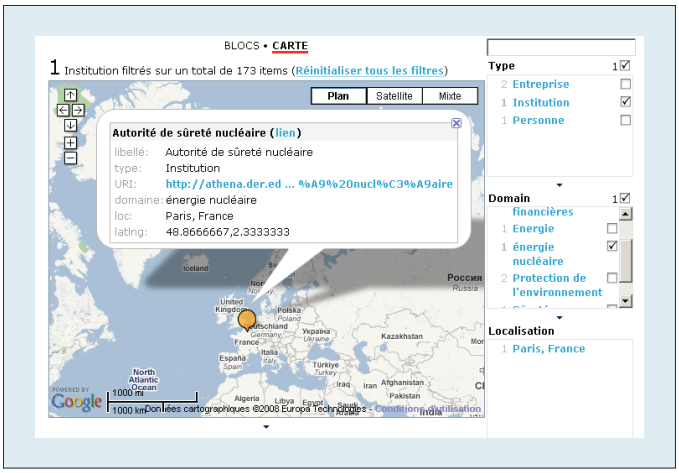

Figure 2. A semantic mashup combining internal and external data in Resource Description Framework (RDF) format. It integrates data created via the corporate semantic wiki with geolocation information from Geonames.org to display wiki instances in a Google map. This map displays a single instance, thanks to the additional faceted browsing capabilities.

Using this additional stack, end users created and maintained more than 300 instances of domain ontologies through these wikis, along with more than 17,000 instances of sioc:Post (and related subclasses) linked to the previous instances. The whole ecosystem became an interoperability layer on top of existing tools, thereby weaving an SSIS into a corporate environment.

Thanks to this combination of semantics for creating and maintaining instances of domain ontologies and a uniform representation of user-generated content, people could find items related to very specific topics (such as thin, 0.1-m solar cells) by searching for broader ones (such as solar energy), wherever this content might come from (blogs, wikis, RSS feeds, and so on).

Another important focus of this use case is how it reuses external data to build semantic mashups, such as the geolocation mashup in Figure 2. By reusing in one infrastructure machine-readable data created by other people, the mashup clearly shows how universal access to open data sources—provided thanks to Semantic Web technologies, especially within the Linking Open Data initiative—can enhance information and put it into context, hence making it more valuable to end users.

Scientific Knowledge Management

Biomedical research is one of the first knowledge domains where real-world use of SSISs has started to emerge. It is a good environment for SSISs for several reasons. First, 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, and so on).

Second, the biomedical domain is distinguished from most other domains in that it already has a large collection of well-structured ontologies and terminologies readily available for the creation of SSISs with rich semantics. Many of these ontologies are available in OWL format in the Open Biomedical Ontologies repository.[7]

One example of an SSIS in the field of biomedical research is the Alzheimer Knowledge Base (http://hypothesis.alzforum.org). It contains a collection of hypotheses about Alzheimer's disease, formulated by various participants from the research community. The hypotheses are captured with the Semantic Web Applications in Neuromedicine (SWAN) discourse ontology.[8] Through a newly created mapping of SWAN to the SIOC vocabulary (www.w3.org/TR/hcls-swansioc), 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: a scientist's proficiency has traditionally been 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 increased the adoption of novel services. In the scientific research domain, uptake of SSISs by mainstream publishers and integration into the scientific publication process is crucial for their widespread success among scientists. Promising current developments include the increasing uptake of social software, such as Nature Connotea (www.connotea.org), and growing efforts to capture the semantics of biomedical publications, as in the Structured Digital Abstracts project (www.febsletters.org/content/sda_summary) and the SWAN-SIOC effort mentioned earlier.

Extending and Interlinking SSISs

Now that we have detailed two domain-specific use cases for SSISs, let's consider how the use of this technology can be extended, both by integrating different data sources in an existing SSIS and by interlinking multiple SSISs.

Integrating with Other Sources of Data

Although SSISs currently focus on Web-based data, it's important to

consider that semantically enriched and socially aware data can also be produced by other means. For instance, desktop data can be integrated in combination with Web data, thanks to projects such as the Nepomuk Semantic Desktop (http://nepomuk.semanticdesktop.org) and vocabularies like the Personal Information Model (PIMO) for representing personal information. Moreover, we can consider integrating more dynamically created data from online communities, drawing on the ubiquitous, real-time, and multidevice capabilities of SSISs. In addition, SSISs could integrate data from microblogging services as well as other streamed data from mobile phones, GPS devices, and other kinds of sensors that people want to share. These types of data are also starting to be semantically enhanced—for instance, within the ConServ project (http://conserv.deri.ie) or the overall Semantic Sensor Web area (http://semsensweb.di.uoa.gr). The Web is on its way to becoming a hub of ubiquitous sociality; further challenges would then be scale and dynamics, as well as trust and privacy of personal data.

Connecting the Dots

Although our two earlier examples are based on particular use cases of domain-specific SSISs, they use a similar methodology that can be applied in any project involving social interactions; a society-wide SSIS can thus be instantiated in any online community that shares and collectively builds information on the Web. Yet, one important question is how we might link these SSISs so that we achieve not merely isolated semantic ecosystems but networked ones to realize the vision of an interlinked online information society (see Figure 3). Three main strategies can improve linkage between various SSISs:

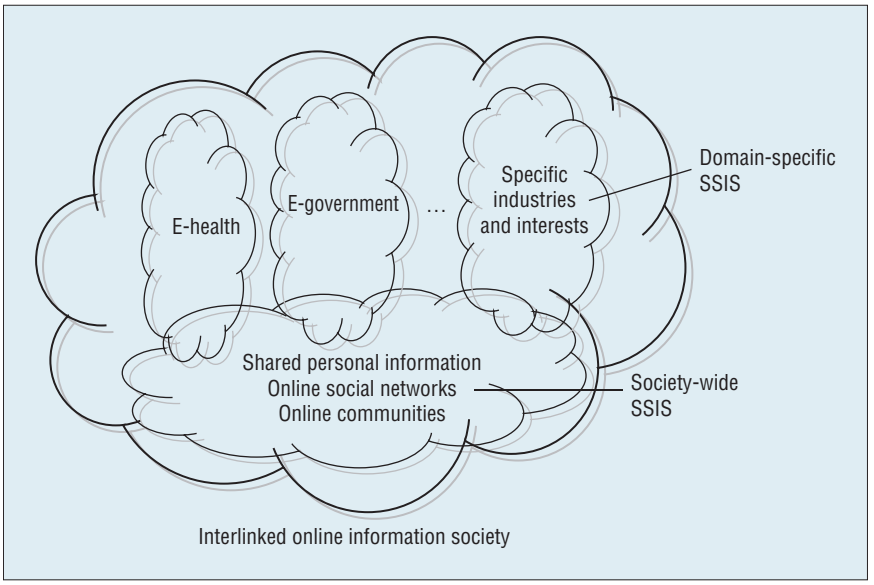

Figure 3. Interlinking social semantic information spaces on a global scale. Domain-specific SSISs are being linked together through a society-wide SSIS to provide an interlinked online information society.

- *Social relationships*. By representing distributed online social networks and user interactions in an interoperable way (using the lightweight vocabularies described earlier), identities and networks can be spread among multiple SSISs instead of remaining in closed data silos.
- *Content types*. When data produced by online communities and personal information management systems is represented using shared, lightweight schemas (notably SIOC, now in wide use as demonstrated by its recent integration into Yahoo SearchMonkey), the same representation format applies to any SSIS. The SSISs are thus linked by unified content types.
- *Topics*. Reusing common identifiers—especially items from the linked data cloud such as DBpedia URIs (uniform resource identifiers) to define topics within SSISs enables topic-based interlinking. Semantic enrichment frameworks for tagging systems make this possible.

By interlinking more people, machines, and data, connecting isolated information spaces also enriches the value of the global network that is the Web, in terms of Metcalfe's law (http://en.wikipedia.org/wiki/Metcalfe's_law). By connecting the dots, we can also efficiently use data from one SSIS in another one, as in our mashup example. Yet, interlinking at Web scale raises new questions, especially about how to usefully exploit this amount of data.

Relationships with the Web Science Agenda

In an earlier *IEEE Intelligent Systems* article, Tim Berners-Lee, Wendy Hall, and Nigel Shadbolt described Web science as "a science that seeks to develop, deploy, and understand distributed information systems, systems of humans and machines, operating on a global scale."[9] As we've explained, our SSIS approach aims to develop and deploy such systems, in which the machine helps people build collaborative knowledge and efficiently reuse it. More generally, we also believe that SSISs—and the Semantic Web as a whole, for which SSISs aim to solve the chicken-and-egg problem by providing, at the same time, structured data and

THE AUTHORS

Alexandre Passant is a postdoctoral researcher in the Digital Enterprise Research Institute at the National University of Ireland, Galway. His research interests focus on the Semantic Web and social software, especially how these fields can interact with and benefit from each other to provide a socially enabled, machine-readable Web. He has a PhD in computer science from Université Paris-IV Sorbonne, France. He is a member of the ACM and the IEEE. Contact him at alexandre.passant@deri.org.

Matthias Samwald is a postdoctoral researcher in the Digital Enterprise Research Institute at the National University of Ireland, Galway, and at the Konrad Lorenz Institute for Evolution and Cognition Research, Austria. His research interests focus on using new Web technologies to accelerate research progress in the life sciences. He has a doctoral degree in natural sciences from the University of Vienna. Contact him at samwald@gmx.at.

John G. Breslin is a lecturer in the School of Engineering and Informatics at the National University of Ireland, Galway. He also leads the Social Software Unit at the Digital Enterprise Research Institute at NUI Galway. His research interests include the social Semantic Web and sensor applications. He is the founder of the SIOC project, which aims to semantically interlink online communities. He has a PhD in electronic engineering from NUI Galway. Breslin is a member of the IEEE, the Institution of Engineering and Technology, and the Institution of Engineers of Ireland. Contact him at john.breslin@nuigalway.ie.

Stefan Decker is director of and a professor in the Digital Enterprise Research Institute at the National University of Ireland, Galway. His research interests include the Semantic Web, digital libraries, and the social semantic desktop. He has a PhD in computer science from the University of Karlsruhe. Contact him at stefan.decker@deri.org.

applications to efficiently take advantage of this data—are a way to make the process of studying and understanding these systems easier with standard representation formats.

Moreover, James A. Hendler and coauthors defined one of the various challenges of Web Science 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?"[10] We believe that the approach we propose is an interesting solution to that issue, because it does not imply changes to the Web architecture (www.w3.org/TR/webarch) but requires a lightweight layer of semantics to make social interactions explicit and machine-understandable. Furthermore, as we recently discussed elsewhere, we believe that this amount of interlinked data will not create a privacy problem, but on the contrary will help provide advanced social policies for trust and access control on the Web.[11]

Social semantic information spaces can be used to build a network of people and computers, aiming to achieve the longstanding vision of social machines. SSISs can be deployed in various environments, and various ecosystems of semantically enriched social data could be linked together to provide an interlinked information society. Although some may call it Web 3.0, or *n*.0, its goal is actually close to the initial vision of the Web: social, open, and machine readable. ■

Acknowledgments

The work presented in this article was funded in part by Science Foundation Ireland under grant no. SFI/08/CE/I1380 (Líon-2).

References

1. J.G. Breslin and S. Decker, "Semantic Web 2.0: Creating Social Semantic Information Spaces," tutorial, 15th Int'l World Wide Web Conf. (WWW 06), 2006, http://www2006.org/tutorials/20060526a.pdf.
2. J.G. Breslin et al., "Towards Semantically-Interlinked Online Communities," *Proc. 2nd European Semantic Web Conf.* (ESWC 05), LNCS 3532, Springer, 2005, pp. 500–514.
3. D.R. Karger and D. Quan, "What Would It Mean to Blog on the Semantic Web?" *Proc. 3rd Int'l Semantic Web Conf.* (ISWC 04), LNCS 3298, Springer, 2004, pp. 214–228.
4. A.P. Mcafee, "Enterprise 2.0: The Dawn of Emergent Collaboration," *MIT Sloan Management Review,* vol. 47, no. 3, 2006, pp. 21–28.
5. J.W. Tanaka and M. Taylor, "Object Categories and Expertise: Is the Basic Level in the Eye of the Beholder?" *Cognitive Psychology,* vol. 23, no. 3, 1991, pp. 457–482.
6. A. Passant et al., "A URI is Worth a Thousand Tags: From Tagging to Linked Data with MOAT," *Int'l J. Semantic Web and Information Systems,* vol. 5, no. 3, 2009, pp. 71–94.
7. B. Smith et al., "The OBO Foundry: Coordinated Evolution of Ontologies to Support Biomedical Data Integration," *Nature Biotechnology,* vol. 25, no. 11, 2007, pp. 1251–1255.
8. P. Ciccarese et al., "The SWAN Biomedical Discourse Ontology," *J. Biomedical Informatics,* vol. 41, no. 5, 2008, pp. 739–751.
9. T. Berners-Lee, W. Hall, and N. Shadbolt, "The Semantic Web Revisited," *IEEE Intelligent Systems,* vol. 21, no. 3, 2006, pp. 96–101.
10. J. Hendler et al., "Web Science: An Inter-disciplinary Approach to Understanding the World Wide Web," *Comm. ACM,* vol. 51, no. 7, 2008, pp. 60–69.
11. A. Passant et al., "Enabling Trust and Privacy on the Social Web," W3C Workshop on the Future of Social Networking, 2009, www.w3.org/2008/09/msnws/papers/trustprivacy.html.