Available online at www.sciencedirect.com
ScienceDirect
Web Semantics: Science, Services and Agents on the World Wide Web 6 (2008) 70–75
JOURNAL OF Web Semantics
www.elsevier.com/locate/websem

The two cultures: Mashing up Web 2.0 and the Semantic Web☆

Anupriya Ankolekar*, Markus Krötzsch, Thanh Tran, Denny Vrandečić

Institut AIFB, Universität Karlsruhe (TH), Germany

Received 8 June 2007; received in revised form 10 September 2007; accepted 6 November 2007
Available online 19 November 2007

Abstract

A common perception is that there are two competing visions for the future evolution of the Web: the Semantic Web and Web 2.0. A closer look, though, reveals that the core technologies and concerns of these two approaches are complementary and that each field can and must draw from the other's strengths. We believe that future Web applications will retain the Web 2.0 focus on community and usability, while drawing on Semantic Web infrastructure to facilitate mashup-like information sharing. However, there are several open issues that must be addressed before such applications can become commonplace. In this paper, we outline a semantic weblogs scenario that illustrates the potential for combining Web 2.0 and Semantic Web technologies, while highlighting the unresolved issues that impede its realization. Nevertheless, we believe that the scenario can be realized in the short-term. We point to recent progress made in resolving each of the issues as well as future research directions for each of the communities.

Keywords: Web 2.0; Semantic Web; Blog; RDF; Vision

1. Introduction

The Semantic Web vision [3] has inspired a large community of researchers and practitioners, who have achieved several early successes in the past six years. After years of successful progress in semantic technologies and the concurrent emergence of Web 2.0, it is time to re-evaluate the progress made by the Semantic Web community, in particular considering current Web 2.0 applications and tools. Looking back, we can characterise the Semantic Web effort thus far as follows[1]:

- *Closed domains*. In contrast to Web 2.0, most Semantic Web applications have assumed closed domains of manageable size, such as the proteins domain, digital libraries and corporate intranets.
- *Complex and comprehensive modeling*. The Semantic Web community has aimed to model as much of the underlying complexity of a domain as possible while covering the domain comprehensively. This is reflected in RDF and RDF(S) and in the standardised *Web Ontology Language* OWL. In addition, academic research has contributed methodologies for ontology engineering, evolution, debugging, and modularisation, aiming for a thorough understanding of the complexity of common ontology languages.
- *Design for knowledge engineers*. The complexity of the modeling and modelling languages has meant that trained knowledge engineers are required for domain modelling and are implicitly assumed to be involved in the design and maintenance of the ontology.
- *Sophisticated reasoning*. Due to the complex domain modelling, there has been a need for sophisticated inferencing methods and scalable reasoners. This has led to the development of increasingly scalable reasoning solutions.
- *Complex specifications and heavy-weight tools*. The documentation and specifications of Semantic Web languages are well-known to be complex and often inscrutable to the average Web developer. Similarly, in comparison to Web 2.0, the vast majority of tools for the Semantic Web are heavy-weight to the point of being unwieldy. There are some improved modelling tools like *Protégé* or *Swoop*, but those too tend to focus on the knowledge engineer as opposed to a Web developer.

☆ With all due respect to C.P. Snow whose title we reuse. This paper is essentially a revised and shortened version of [1], published at WWW 2007, and incorporates feedback from the conference.

* Corresponding author. Present address: Hewlett-Packard Laboratories, Palo Alto, USA.
E-mail address: anupriya.ankolekar@hp.com (A. Ankolekar).

[1] Note that these are broad characterizations and may not hold for all Semantic Web research, but they are valid enough for purposes of discussion.

1570-8268/$ – see front matter © 2007 Elsevier B.V. All rights reserved.
doi:10.1016/j.websem.2007.11.005

Web 2.0, sometimes positioned as conflicting with the Semantic Web vision, has been outlined in [15] and characterises current state-of-the-art in web engineering, as exemplified by sites such as *Wikipedia*, *flickr* and *HousingMaps*. Web 2.0 technologies augment the Web, allowing for easy distributed collaboration and can be distinguished from the classical Web by the following characteristics:

- *Community*. Web 2.0 pages allow contributors to collaborate and share information easily. The emerging result could not have been achieved by each individual contributor, be it a music database like *freedb*, or an event calendar like *upcoming*.
- *Mashups*. Data from different sites can be pulled together in order to provide new values with the different combinations of the data. This allow for a whole range of handcrafted merges of data sources, from the dynamic embedding of advertisements in *AdSense* to the dynamic visualisation of housing information on *Google Maps*.
- *AJAX*. The technological pillar of the Web 2.0 enables the creation of responsive user interfaces, thus facilitating both other pillars: community pages with slick user interfaces can reach much wider audiences, and mashups that incorporate data from different websites introduce asynchronous communication for more responsive pages.

We believe that the Semantic Web and Web 2.0 are complementary rather than competing—their goals are in harmony and each brings its own strengths into the picture. The Semantic Web community is beginning to realise the value that active communities and AJAX technology can bring [13,16]. However, we believe that the Semantic Web and Web 2.0 can be intertwined much more deeply.

To demonstrate this, we will describe a Web 2.0 scenario using semantic technologies, which we believe can become reality within less than two years. We outline an architecture for the scenario and describe the gaps in achieving the vision. Addressing these gaps is unlikely to require huge engineering efforts or the solving of open research issues, but will inevitably lead to a whole slew of new requirements, helping the research community to focus on those topics that are most relevant for the open Semantic Web.

We formulate our scenario on the basis of the following three hypotheses. They criticise certain assumptions held by part of the Semantic Web community, but we will show how they can help to reconcile the two communities.

1. *The Semantic Web will be a World Wide Web*. The Semantic Web will not be restricted to corporate intranets or singular islands of knowledge. Rather, it will incorporate large portions of the Web, displaying heavy reuse of URIs and high interconnection. Corporate semantic intranets will of course continue to exist, possibly with certain advantages over the world-wide Semantic Web, but in general the latter will easily be the most prominent and demanding use case.
2. *A bottom up, user-centred approach is required for the Semantic Web to take hold*. The Web itself did not come about as a result of a commercial push. It began in research facilities and with private, personal Web sites, with years passing before companies recognised the need for a Web presence. Similarly, we believe that the first popular Semantic Web sites will grow as a result of community-centred efforts such as semantically-enhanced blogs and wikis.
3. "*A little semantics goes a long way*."[2] The first iteration of the Semantic Web will profit enormously from light-weight languages for exchanging information. These will have to go beyond the expressiveness of RDFS, e.g. to allow instance identification and some light-weight mappings, but they may well be below the expressivity offered by OWL Lite.[3]

2. Scenario

In this section, we describe a concrete scenario of how Semantic Web technologies could enhance current Web 2.0 tools and experience. We pick blogging as a typical example of a Web application that is widely used, in particular for posting opinions and links to other content on the Web. This makes it fertile ground to explore the possibilities of extensive data integration and reuse enabled by the Semantic Web.[4]

Let's consider Chrissie, a fairly typical Web blogger, using *Movable Type*, a popular weblog publishing system[5]. She goes to the cinema regularly and blogs afterwards about the movies she watched. Her audience consists mostly of her friends some people who accidentally stumble upon her movie reviews. She follows a straightforward workflow when writing reviews: she creates a new blog entry, enters a title, writes the text, and maybe tags it with one or more tags like the genre of the movie [12]. The blog publishing system takes care of displaying, syndicating, and archiving the entry.

2.1. Reusing data from the Web

Now imagine a blog application movie plug-in that uses Semantic Web technologies and allows people to add information about movies to their blog entries. Chrissie chances upon this plug-in-let's call it *Smoov*—and installs it. Her workflow for writing movie reviews now changes slightly: she first has to explicitly state that she is writing a movie review. This causes a number of extra fields to appear in her blogging application, that allow her to identify (e.g. via its *IMDb* page), rate the movie, etc. Now Smoov is able to pull in some data about the movie and create a movie sidebar, as shown in Fig. 1. Chrissie configured the sidebar once to show specific information about movies, such as the director, the major actors, a link to the official Web site of the movie, and more. She now checks to see whether the sidebar looks good, and chooses a picture to display. Using RDF licence information accompanying pictures from the movie [8],

[2] Jim Hendler, Opening the International Semantic Web Conference in 2003.
[3] Tractable fragments of OWL 1.1, for instance, could become very relevant http://owl1_1.cs.manchester.ac.uk/tractable.html.
[4] The idea of extending blogging with semantics is not new, and has been described previously, e.g. in [11,6].
[5] http://www.movabletype.org/

Fig. 1. A screenshot of the movie plug-in used in a blog entry. The plug-in adds a sidebar to the entry containing the picture, data about the movie (running time, director, actors, etc.), and the screening times dynamically acquired from external sites.

Smoov guides Chrissie in choosing a picture that conforms to legal requirements.

The movie data pulled in by Chrissie's blog is available on a central space in a machine-readable format. This could be a semantically enhanced Wikipedia [17], a screenscraping service (such like the various scrapers available at SIMILE[6]) that extracts the information from the IMDb movie page, or *freebase*, a collaboratively-edited database of cross-linked data.[7] The movie information displayed can range from static data (like the director) to highly dynamic data, like the movie's chart position. Based on the nature of the data, different caching and retrieval mechanisms need to be applied in order to ensure an acceptable response time of Chrissie's blog.

2.2. Dynamic data sources

Configured with a (URL) list of Chrissie's favourite cinemas, Smoov could locate additional dynamic information, such as the playtime of the movies at a local cinema. Such a service could be offered by city guide sites that collect such information anyway or by the cinemas themselves. Once the movie stops running in the cinemas, Smoov would simply stop displaying the movie showtimes. Once the DVD of the movie is out, as reported by IMDb, the plug-in could link to Chrissie's favourite movie stores and online rental services, as configured by her, and display the prices of the movie.

Why is this scenario of dynamic data sources realistic? Cinemas have several benefits when providing information about current movies and their showtimes in RDF. Based on XML, RDF is an universal model for data representation and at the same time, simple enough for many processing tasks like the combination of disparate data, e.g. automated mashups. Moreover, ontologies associated with RDF data deliver the semantics that facilitate machine-based interpretation and processing. Most importantly, there are already RDF stores and reasoners available for the exploitation of these merits. These technologies enable greater interoperability, control, correctness and consistency of the data that can be transferred over the Web. Thus, cinemas can reach larger user groups and propagate changes in their programmes more efficiently in a standardised and uniform way through the Web. Offering such information also requires fairly low effort. Many cinemas already maintain that information in a database and thus, only need to attach a SPARQL endpoint to their database, or write a simple RDF exporter besides an existing HTML exporter.

2.3. Personalisation of Web sites

There are also some interesting personalisation possibilities in this scenario. Readers of Chrissie's blog, who do not live geographically close to Chrissie, would be more interested if Chrissie's blog could display movie showtimes for *their* favourite cinemas, instead of those cinemas configured by Chrissie. There are several ways to realize such a scenario:

- Smoov could try to guess the location of the reader, based on her IP. Web advertisements often use this form of personalisation, but the major drawback is that this is only helpful in identifying the user's location—but no further information.
- By offering login and accounts, readers could set up their own preferences. This requires the blog application to handle user

[6] http://simile.mit.edu/wiki/Category:Javascript_screen_scraper.
[7] http://www.freebase.com.

accounts and users to create and remember login credentials as well as potentially replicate the same information on several websites. It also prevents serendipitous usage of data, since readers always have to register before getting the advantage of context-aware data reuse.
- Web surfers could offer information about themselves in Semantic Web formats like FOAF [5] and then point Chrissie's blog to such a resource. This could be either done with connecting an identity system like OpenID command in order to send a reference to the user's FOAF file [2].

Without imposing further efforts on Chrissie, she and her readers reap immediate benefits by providing a highly personalised Web experience.

2.4. Giving back to the Web

Chrissie and her blog readers clearly benefit from Smoov's Web data integration, reuse, and personalisation capabilities. But does the Web itself benefit from Chrissie's Semantic Web site? It can! Smoov could export Chrissie's movie ratings in computer understandable format – be it encoded in RDF or hReview – so that crawlers and agents can collect and understand that data.

The Semantic Web is built on a decentralised and open infrastructure that can facilitate data interoperability. There is a great potential for having all sides participate in an open data Web, and having intelligent services present and adapt data to the users—such as Smoov. Web sites can then benefit from collecting the review data from many different, heterogeneous sources like Chrissie's blog. They can display aggregated reviews, and look out for trends (the blogosphere typically has more and quicker reviews than the reviews on most online stores). Machine-understandable ratings make it much easier to put up pages like Google's Movie Ratings page.[8] This would provide new services like *FilmTrust* with enough data to immediately produce meaningful movie recommendations, turning data into a commodity rather than an asset.

In order to aggregate the data from different sources – to display the average rating, for example – a system would not apply logical reasoning, but statistics. Widely used features like tag clouds and recommendations demonstrate that data mining and clustering can handle the scale of the web, whereas reasoning in description logics system usually will not be able to deal with a massive amount of instances. It is a common misconception that the Semantic Web requires reasoning on a big scale: whereas ontologies will provide the necessary background knowledge for data integration and cleansing, data mining can be applied to the data after integration, thus combining and leveraging the strengths of well-known technologies where they work best.

3. Infrastructure

The scenario just presented is certainly not a pure Semantic Web application, but involves a number of related Web technologies, and – maybe most importantly – significant human contribution. We argue that this paradigm shift from an overly machine-centred AI view of the Semantic Web is necessary and healthy both for the involved research communities and for the Web as a whole. This claim, however, also provokes two kinds of critical reactions:

1. "The scenario is not realistic, since it assumes significant background infrastructure that is not available today – the Semantic Web still lacks some crucial technologies to make this possible."
2. "The scenario is not a Semantic Web scenario, since it does not really challenge semantic technologies – you could just as well use XML to transfer data in the described way."

In the remainder of this section, we will argue against these positions by an elaboration on a basic Semantic Web infrastructure that can support our scenario, and can be built with existing semantic technologies. Imagining the Semantic Web to be an ecosystem of entities creating, sharing and reusing (meta)data on the Web, the following (non-exclusive) tasks need to be solved: *creation* (what are the sources of semantic data?), *exchange* (how to distribute, gather and combine semantic data?), and *reuse* (how can semantic data be put to practical use?). While these tasks could be supported by hard-coded procedures built on XML-based technologies alone, we will argue that semantic technologies can provide more flexibility and in particular, can facilitate the exchange and reuse of data.

3.1. Creation

The Semantic Web uses a large number of machine-readable data formats that are the basis for semantic technologies. But where should this data that humans can hardly read, not to mention author, actually come from? An early attempt to answer this was made by the FOAF project, the idea being that a large number of people author small amounts of semantic data. Tools like FOAF-a-matic[9] simplify the creation of FOAF files. The SIOC[10] has a similar goal in that it aims to generate semantic data from online communities' discussion channels and posts. In spite the relative success, FOAF in particular, it is hard to claim that such approaches alone can really solve the problem of data creation.

But many web applications are already based on well-structured data – often maintained in an internal database in an application-specific format –, and semantic data formats are suggestive for publishing such pre-existing data. Encoding such data may need some work, but there are hardly any technical problems. The approach already works in specific domains. For instance, *flickr* embeds RDF into HTML pages for publishing available license information, and all major blogging engines provide (RDF-based) RSS feeds. Much more existing data, e.g. the millions of available library catalogue records, could be pub-

[8] http://www.google.com/movies.

[9] http://www.ldodds.com/foaf/foaf-a-matic.html.

[10] http://sioc-project.org/vision.

lished in a similar way. On the other hand, there are also efforts to simplify the direct authoring of semantic data. Examples of this include Semantic MediaWiki [17], where semantic data is edited in a wiki, and the recent "machine tags" in *flickr*, that allow (RDF) namespaces within tags. Incorporating the creation of semantic data into the interfaces of existing applications, most kinds of blogs, forums, online directories, etc. can easily become semantic data sources as well.

3.2. Exchange

Exchanging existing data is a straightforward task in classical Web scenarios. On the Semantic Web, however, data must also be transformed, merged, and collected to enable later reuse. The most prominent related task is *mapping* available data to a common terminology/format that can be further processed. Languages used on the Semantic Web ease the exchange of structural information, but they do not encode the intended meaning of such structures. Yet using the data also requires understanding its informal semantics and handling it in an application-specific way.

One existing solution to this problem is to refer to established ontologies. Applications that are aware of a given ontology can easily interpret respective data sets, as in our blogging scenario. In addition, further pre-processing steps might be required to reconcile data. For instance, the *Planet* blog reader aggregates machine-readable feeds from many blogs, merges the collected news items by date, and supports various additional filtering functions. Another fully customisable online tool for processing various kinds of data feeds is *Yahoo!pipes*,which produces data in multiple machine-readable formats. Similar *aggregators* will play an important role in the emerging Semantic Web, especially as ontologies become more numerous and filtering methods become more complex. We believe aggregators are where most of the research challenges and future commercial interests lie.

3.3. Reuse

Creation, publication, and exchange of data are only useful if there are ways of exploiting this information. A large number of tools currently is exploiting semantic data in one or the other way, but many of them are used only within a very limited academic context. There are various tools that process FOAF or RSS data, which we do not attempt to list here, but at the moment only RSS readers have really made the leap to user desktops [18].

Examples of large scale web applications include semantic search engines, such as the Creative Commons Search engine,[11] or *Swoogle* [10]. These applications are especially interesting since they provide services beyond mere display of data, and employ technical solutions for more complex processing tasks. Another important use of semantic data is the recombination of data sources on the Web, creating what is typically known as *mashup*. Mashups have already been realised using XML feeds, JavaScript APIs or REST, an API containing simple operations against entities identified by URIs. However, these implementations require significant programming effort, are very sensitive to changes on the source sites, and rely on proprietary APIs. Semantic technologies advertise the use of common data formats that are universal across application domains, and hence greatly facilitate the construction of mashups. The aforementioned aggregators *Planet* and *Yahoo! pipes* also provide online interfaces that are good examples of successful semantic mashups. It is not obvious how a tool as versatile as *Yahoo! pipes* could be build without the use of machine-readable formats that enable seamless data exchange.

Besides the data available in standardised Semantic Web formats, there is plenty of data available on the web in well-specified semantic formats, like iCalendar [9], Atom [14], hReview [7]. Such standards, especially the set of Microformats[12], can usually be transformed easily into the RDF data model and thus allow to be integrated into the Semantic Web vision, just as the vast amount of data found in databases do [4].

4. Conclusions

The Semantic Web and the Web 2.0 are often presented as competing visions for the future of the Web. However, there is growing realisation that the two ideas complement each other, and that in fact both communities need elements from the other's technologies to overcome their own limitations.

As we have shown, existing Web application scenarios, such as blogging, can be worthwhile goals for spurring the further development of semantic technologies. We advocate a paradigm shift from an *overly* machine-centred AI view of the Semantic Web towards a more user—and community-centred approach that draws from the insights of Web 2.0. Of course, research on foundational topics, such as expressive ontology languages and the associated technologies and methodologies, will be required, but the need of the hour is to focus on more simple Web application scenarios. Semantic technologies, in turn, bear a great potential of providing a robust and extensible basis for emerging Web 2.0 applications. Interchange, distribution, and reuse of data can be greatly facilitated by the infrastructures that the Semantic Web offers. Jointly exploiting each other's achievements and insights, the two communities can realise their respective visions of the web—because there's only one Web, after all.

Acknowledgements

We want to thank everybody who has engaged in fruitful discussions over the ideas described in this paper, which includes basically everybody from the Knowledge Management groups of AIFB and FZI. We want to thank especially Valentin Zacharias, Tom Heath, and Peter Haase for their valuable and extensive comments.

[11] http://search.creativecommons.org/.

[12] http://microformats.org.

References

[1] A. Ankolekar, M. Krötzsch, D.T. Tran, D. Vrandečić, The two cultures: Mashing up web 2.0 and the semantic web, in: Proceedings of the 16th Conference on the World Wide Web (WWW), Banff, Canada, 2007.

[2] A. Ankolekar, D. Vrandečić, Personalizing Web surfing with semantically enriched personal profiles., in: M. Bouzid, N. Henze (Eds.), Proceeding of Semantic Web Personalization Workshop, Budva, Montenegro, 2006.

[3] T. Berners-Lee, J. Hendler, O. Lassila, The Semantic Web, Sci. Am. (2001) 5.

[4] C. Bizer, A. Seaborne, D2RQ–treating non-RDF databases as virtual RDF graphs., in: S.A. McIlraith, D. Plexousakis, F. van Harmelen (Eds.), Proceedings of 3rd International Semantic Web Conference (ISWC04), Hiroshima, Japan. Springer, November, 2004.

[5] D. Brickley, L. Miller, FOAF vocabulary specification, revision, 2003. Available at http://xmlns.com/foaf/0.1/.

[6] S. Cayzer, Semantic blogging and decentralized knowledge management, Communications of the ACM 47 (12) (2004) 47–52.

[7] T. Celik. hReview 0.3, 22 February 2006. Available at http://microformats.org/wiki/hreview.

[8] Creative Commons. Some Rights Reserved: Building a Layer of Reasonable Copyright. http://creativecommons.org.

[9] F. Dawson, D. Stenerson. Internet Calendaring and Scheduling Core Object Specification, RFC 2445, IETF, November 1998.

[10] L. Ding, T. Finin, A. Joshi, R. Pan, R.S. Cost, Y. Peng, P. Reddivari, V. Doshi, J. Sachs, Swoogle: a search and metadata engine for the Semantic Web., in: Proceedings of 13th ACM Conference on Information and Knowledge Management, 2004, pp. 58–61.

[11] D.R. Karger, Quan, D., What would it mean to blog on the semantic web? in: S.A. McIlraith, D. Plexousakis, F. van Harmelen (Eds.), Proceedings of 3rd International Semantic Web Conference (ISWC04), Hiroshima, Japan, Springer, November, 2004, pp. 214–228.

[12] C. Marlow, M. Naaman, d. boyd, M. Davis, HT06, tagging paper, taxonomy, flickr, academic article, to read, in: In Proceedings of 17th Conference on Hypertext and Hypermedia (HYPERTEXT'06), 2006, pp. 31–40.

[13] P. Mika, Ontologies are us: a unified model of social networks and semantics., in: Proceedings of 4th International Semantic Web Conferences (ISWC05), 2005, pp. 522–536.

[14] M. Nottingham R. Sayre, The Atom Syndication Format, RFC 4287, Internet Engineering Task Force, December 2005.

[15] T. O'Reilly. What is Web 2.0—Design Patterns and Business Models for the Next Generation of Software, 2005. Available at http://www.oreillynet.com/pub/a/oreilly/tim/news/2005/09/30/what-is-web-20.html.

[16] E. Oren, R. Delbru, S. Decker, Extending faceted navigation for rdf data., in: I. Cruz, S. Decker (Eds.), Proceedings of 5th International Semantic Web Conference (ISWC06), 2006, pp. 559–572.

[17] M. Völkel, M. Krötzsch, D. Vrandečić, H. Haller, R. Studer, Semantic Wikipedia., in: Proceedings of 15th International Conference on World Wide Web (WWW 2006), Edinburgh, Scotland, May, 2006, pp. 23–26.

[18] V. Zacharias, M. Sibler, Semantic announcement sharing., in: Proceedings Fachgruppentreffen Wissensmanagement, 2004.