The Two Cultures*

Mashing up Web 2.0 and the Semantic Web – Position paper –

Anupriya Ankolekar, Markus Krötzsch, Thanh Tran, Denny Vrandecic
Institut AIFB, Universität Karlsruhe (TH), Germany
[ankolekar,kroetzsch,tran,vrandecic]@aifb.uni-karlsruhe.de

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.

Categories and Subject Descriptors

H.3.5 [**Information Storage and Retrieval**]: Online Information Systems; H.4.3 [**Communications applications**]: Information browsers; H.5.3 [**Information Interfaces**]: Group and Organization Interfaces—*Web-based interactions*

General Terms

Human Factors, Languages

Keywords

Web 2.0, Semantic Web, Blog, RDF, Vision

1. INTRODUCTION

There is a common perception that there are two competing visions about the future evolution of the Web: the Semantic Web and Web 2.0. We believe that the technologies and core strengths of these visions are complementary, rather than in competition. In fact, both technologies need each other in order to scale beyond their own strongholds. The Semantic Web can learn from Web 2.0's focus on community and interactivity, while Web 2.0 can draw from the Semantic Web's rich technical infrastructure for exchanging information across application boundaries.

The Semantic Web vision outlined in [6] has inspired a big community of researchers and practitioners, and they have achieved a number of goals: languages like RDF [37] and RDF(S) [12] were revised, the *Web Ontology Language* OWL [47] was standardised. Academic research contributed methodologies for ontology engineering [48, 51], evolution [36], debugging [30], and modularisation [46], and has lead to a thorough understanding of the complexity and decidability of common ontology languages [4]. These insights enabled the implementation of increasingly scalable solutions for inferencing [29], and of improved modelling tools for ontologies [32]. Based on those achievements, major companies like *Oracle* and *IBM* are working on large scale data stores supporting Semantic Web standards, and a growing number of specialised companies such as *Aduna*, *Altova*, *Ontoprise*, and *TopQuadrant* provide industrial strength tool sets facilitating the use of semantic technologies in corporate settings.

The Web 2.0 technologies, as outlined in [42] and exemplified by sites such as *Wikipedia*[1], *flickr*[2] and *HousingMaps*[3], augment the Web and allow for an easier distributed collaboration. They are distinguished from classical web technologies by various characteristic features:

- **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*,[4] or an event calendar like *upcoming*[5]. Each contributor gains more from the system than she puts into it.
- **Mashups.** Certain services from different sites can be pulled together in order to experience the data in a novel and enhanced way. This covers a whole range of handcrafted solutions, ranging from the dynamic embedding of *AdSense*[6] advertisements to the visualisation of *CraigList's*[7] housing information on *Google Map*[8], as done by *HousingMap*[9].

*With all due respect to C.P. Snow whose title we reuse.

WWW 2007, May 8–12, 2007, Banff, Alberta, Canada.
ACM 978-1-59593-654-7/07/0005.

[1] http://wikipedia.org
[2] http://www.flickr.com
[3] http://www.housingmaps.com
[4] http://www.freedb.org
[5] http://upcoming.org
[6] http://www.google.com/adsense
[7] http://www.craigslist.org
[8] http://maps.google.com
[9] http://www.housingmaps.com

- **AJAX.** The technological pillar of the Web 2.0 allows to create responsive user interfaces, and thus facilitated both of the other pillars: community pages with slick user interfaces could reach much wider audiences, and mashups that incorporate data from different websites introduced asynchronous communication for more responsive pages.

It is notable that the term Web 2.0 was actually not introduced to refer to a vision, but to characterise the current state of the art in web engineering [42].

We believe that these two ideas are complementary rather than competing – a view that is gaining acceptance within the Semantic Web community, as shown, e.g., in panel discussions at WWW 2006[10] and ISWC 2006[11]. The goals of the Semantic Web vision and Web 2.0 are aligned, and each brings it owns strengths into the picture.

The Semantic Web vision predates the rise of the Web 2.0, but did not foresee the emergence of the Web 2.0 or take this into proper account. After years of successful progress in semantic technologies, we believe that the time has come for the Semantic Web community to look back at the Web, in particular Web 2.0 applications and tools. The Semantic Web community is realising the potential that communities and AJAX can bring to the Semantic Web, as exemplified in research studying the relationship of folksonomies and ontologies [39], or in the growing number of Semantic Web tools using AJAX technology [43]. On the other hand, it is time to study Web 2.0 mashups, identify their limitations, and leverage existing Semantic Web solutions in order to boldly go beyond these limitations.

In order to demonstrate the complementarity of the two ideas, we will describe a Web 2.0 scenario using semantic technologies. We claim that the presented vision can become reality within less than two years. We outline an architecture and describe the missing parts that are required in order to achieve the vision. None of these missing parts require huge engineering efforts or will be hampered with open research issues. Nevertheless, actually realizing these parts will inevitably lead to a whole slew of new requirements and will help the research community to focus on the topics that emerge as being the most relevant on the open Semantic Web.

We base our work on the following three hypotheses. They criticise certain assumptions that are found in some parts of the Semantic Web community, and they have guided us in the formulation of the presented example scenario. We think that these hypotheses can help to reconcile the two communities, as the remainder of the paper will show.

1. **The Semantic Web will be a World Wide Web.** This means, it will not be restricted to corporate intranets or consist of singular islands of knowledge. It rather will be based on large portions of the web, displaying a heavy reuse of URIs and a high level of interconnection. This does not mean, that corporate semantic intranets will not exist: it is even expected, that in these cases they will have certain advantages over the Worldwide Semantic Web, but in general the latter will be easily the most prominent and most demanding part of the infrastructure.

2. **A bottom up, user-centred approach is required for the Semantic Web to take hold.** The Web itself was not started by major companies: it started in research facilities and with private, personal web sites. Only years later companies recognised the need for a web presence. Indeed we think that among the first popular Semantic Web sites we will find community-centred efforts such as semantically enhanced blogs and wikis. Yet many large scale projects today move the other way.

3. **"A little semantics goes a long way."**[12] The first iteration of the Semantic Web will profit enormously from light-weight languages for exchanging information. They will have to go beyond the expressiveness of RDFS, for example in order to allow instance identification and some light-weight mappings, but they might also be well below the expressivity offered by OWL DL or OWL Lite.[13]

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.[14]

Let's consider Chrissie, who has been blogging for three years. She first started with a Web-based blogging service, but recently moved to a Webspace that runs PHP and MySQL, allowing her to use one of the popular blog publishing systems like Movable Type[15] or WordPress[16].

Chrissie is a fairly typical Web blogger, having some basic skill in HTML and CSS. Her blog offers an RSS feed [5], but this is automatically provided by the blogging application itself. She does not know how an RSS feed is written, although she can subscribe to RSS feeds. She has never heard of RDF, and is thus unaware that her RSS feed is probably based on RDF. She of course is not aware of Semantic Web standards like OWL, SPARQL [45], or XML [10].

Chrissie goes to the cinema regularly and tend to blog afterwards about the movies she watched. Her audience is fairly small, mostly friends and acquaintances, and some people who might accidentally stumble upon her movie reviews. She follows a straightforward workflow when writing reviews on her blog. Just as for any other blog entry, she creates a new entry, enters a title, writes the text, and maybe tags it with one or more tags, e.g. describing the genre of the movie [38]. After pushing the *publish* button, the entry

[10] `http://www2006.org/programme/item.php?id=panelk01`

[11] `http://iswc2006.semanticweb.org/program/webpanel.php`

[12] Jim Hendler, Opening the International Semantic Web Conference in 2003

[13] Tractable fragments of OWL 1.1, for instance, could become very relevant `http://owl1_1.cs.manchester.ac.uk/tractable.html`.

[14] The idea of *semantic blogging* is not new, and has been described previously in [31, 15, 16].

[15] `http://www.movabletype.org/`

[16] `http://wordpress.org/`

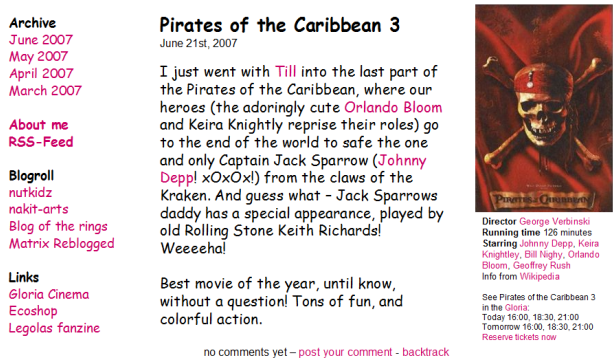

Figure 1: A screenshot of the movie plug-in used in a blog entry. The plug-in adds a sidebar to the entry containing the picture, the data about the movie (running time, director, actors, etc.), and the screening times dynamically acquired from external sites.

is saved in her blog database. The blog publishing system then takes care of displaying the entry on the front page of her blog and archiving the entry appropriately. Furthermore, the RSS feed will be updated with the entry, so that subscribed feed readers can get the new 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. To begin with, Chrissie has to explicitly state that she is writing a movie review. This causes a number of extra fields to appear in her blogging application. The first field asks her to identify the movie. She can specify the exact title of the movie or search for movies by entering the actors, the director etc. Other methods of identifying movies could use, e.g., a few selected authoritative sources such as the IMDb[17] page or the Wikipedia article of the movie and reuse their URIs. If both pages are known to refer to the same movie, e.g. via the Wikipedia's external link to the IMDb movie page, it would not matter which page Chrissie uses as the movie identifier.

Now that Chrissie has identified the movie, Smoov pulls in some data about the movie and creates a movie sidebar, as shown in Figure 1. Chrissie configured the sidebar once to show specific information about movies, such as the director, the major actors, running time, production company, release year, but also the URL of a poster or some distinctive pictures from the movie, and a link to the official Web site of the movie. She now checks to see whether the sidebar looks good, and chooses an appropriate picture to display on the movie sidebar. Using RDF licence information accompanying pictures from the movie [17], Smoov guides Chrissie in choosing a picture that conforms to legal requirements.

[17] http://www.imdb.com/

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 [52], a semantically enhanced IMDb, or simply a screenscraping service (such like the various scrapers available at SIMILE[18]) that extracts the requested information from the IMDb movie page. There are already mature technologies available and it is only a matter of time before such a data source becomes reality. In fact, DBLP for instance, a service for collecting and providing bibliographic metadata[19], has recently been enhanced with a SPARQL endpoint that enables the query and reuse of DBLP data in a simple and standard way.

The movie information that Chrissie draws from may be static data, for instance the director and the actors of the movie. Such information can be pulled into the sidebar once and will basically remain unchanged afterwards, barring of course factual inaccuracies and typos. Other displayed information may be semi-static, such as the awards received by the movie, or dynamic, such as the current chart position of the movie or the availability of tickets for that movie in local cinemas. These different kinds of data will each need to be cached differently. In addition, the sidebar might itself be displayed before all the data has actually been gathered and then updated dynamically, e.g. using AJAX. This is necessary to keep the response time of Chrissie's blog acceptable.

2.2 Dynamic data sources

If Chrissie were to configure Smoov with a (URL) list of her favourite cinemas, the plug-in could locate additional dynamic information, such as the movies currently playing at the cinemas. Such a service could be offered by city guide sites that collect such information anyway, or by the cinemas themselves or perhaps by applications that screenscrape the cinema websites and generate RDF data. 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 the data that can be transferred over the Web. Thus, cinemas can reach a 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. Most cinemas maintain that information in a database anyway and only need to attach a SPARQL endpoint to their database, or write a simple RDF exporter besides an existing HTML exporter.

[18] http://simile.mit.edu/wiki/Category:Javascript_screen_scraper

[19] http://dblp.uni-trier.de/

The interoperability of data exchange of course assumes the existence of an accepted vocabulary for such data. Creating a novel vocabulary, on the other hand, would require considerable engineering effort, which a single cinema cannot and should not provide. However, there are increasingly more ontologies available on the web which can be reused – Swoogle[20] alone has indexed more than 10,000 ontologies.

Other data sources like Amazon[21] also benefit similarly: their data pushed this way to targeted groups is under their control, with dynamic updates to prices and other product information.

Bloggers like Chrissie also benefit with low effort. She only needs to configure which Web services should be used and in return, she get current data on her blog entries and the chance to benefit financially, e.g. through an affiliates program, like the one Amazon offers.

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, may not care too much about information on movie showtimes in Chrissie's favourite cinema. On the other hand, what if Chrissie's blog could display movie showtimes for *their* favourite cinemas? There are several ways to realize such a scenario:

- Smoov could try to guess the location of the reader, based on her IP. Although Web advertisements often use this form of personalisation, it has several drawbacks. Smoov might just guess wrong or determine a location that is too generic to be useful. This also help only in the identification of the user's location – but no other information about her.

- If Chrissie's blog offers user registration, it could allow her to set up preferences like favourite movie genres and location, and either store a cookie or require an explicit login. While this allows for the best service, it requires Chrissie's blog application to handle user accounts, and users to create and remember an account 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.

- The ideal solution would allows Smoov to use information about the reader encoded in open Web standards. For example, [1] describes an infrastructure that uses an extension of the HTTP GET command in order to send a reference to the reader's FOAF file [13]. This FOAF file would include data about the location or even the favourite cinema of the reader and can thus be immediately reused for displaying highly personalised information. Other options include connecting the FOAF data to an OpenID[22] account, or including pointers to a locally running SPARQL endpoint at the reader's machine that would furnish further data, or even personalised answers, to be displayed on the site. With such FOAF information available, Smoov could query an open review system like *revyu*[23] [27] and display further reviews by the reader's friends.

Using these extensions does not impose any further costs on Chrissie, but still reap immediate benefits – for Chrissie and the readers of her blog – leading to a highly personalised Web experience. If Smoov cannot figure out anything about who is reading the blog, it defaults to Chrissie's preferences.

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? What if Chrissie, while giving Smoov metadata about a movie, would also rate the movie on her preferred rating scale? If Smoov would export Chrissie's movie ratings to the Semantic Web, her rating and review text would represent a contribution to the (Semantic) Web.

The Semantic Web is built on a decentralised and open infrastructures that can facilitate data interoperability by means of standardised taxonomies and ontologies. Such common vocabularies make it easier to unlock and share the data between different Web pages. 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 the plug-in here. 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.[24] This would provide novel experimental pages like FilmTrust[25] [26] with enough data to immediately produce meaningful movie recommendations.

3. INFRASTRUCTURE

The scenario of the previous section is certainly not a pure Semantic Web application, but involves a number of related Web technologies, and – maybe most importantly – significant human contributions. 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. But this claim 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 as well use XML to transfer data in the described way."

We will address these two somewhat complementary critiques in this and the following section, our claim being that application scenarios of the described kind are realistic and still bear complex research challenges.

[20] http://swoogle.umbc.edu/
[21] http://www.amazon.com
[22] http://openid.net/
[23] http://revyu.com
[24] http://www.google.com/movies
[25] http://trust.mindswap.org/FilmTrust/

In the remainder of this section, we discuss the basic Semantic Web infrastructure that our scenario requires, and show how it can be built with current semantic technologies. Focussing on the Semantic Web's goal of enabling the sharing and reusing of (meta)data on the Web, the following (non-exclusive) tasks need to be solved:

Creation. What are the sources of semantic data?

Exchange. How can semantic data be distributed, gathered, and combined?

Reuse. How can semantic data be put to practical use?

The Semantic Web requires a complete implementation of the above "food chain," and our imaginary scenario also assumes respective components and service providers.

3.1 Creation

The Semantic Web uses a (growing) number of machine-readable data formats that are the basis for semantic technologies. Any practical (re)use of semantics thus hinges upon the availability of such data. But in contrast to the classical Web, semantic data formats are not mere encodings of human-readable multimedia documents, and so it is usually not obvious how to even *present* semantic data to users. So 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. In spite of the relative success of FOAF, it is hard to claim that such an approach can really solve the problem of data creation, since the barrier of authoring OWL/RDF is too high for most Web users. Tools like FOAF-a-matic[26] simplify the creation of FOAF files, but publication and update of FOAF files often remain tedious manual tasks.

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 published 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 [33], where semantic data is edited in a wiki, and the recent "machine tags" in *flickr*,[27] 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 at first seems to be a simple task, and it often is 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 to understand this informal aspect, and to treat it in an application-specific way.

One existing solution to this problem is to refer to established ontologies. These should consist of a well-specified vocabulary of URIs, and a machine-processable set of axioms that describe their interrelationships, specific constraints, or connections to other ontologies. Applications that are aware of a given ontology can easily interpret respective data sets, and we also made this assumption in our earlier blogging scenario.

Exchanging data also may suggest further pre-processing steps. For instance, the *Planet* blog reader[28] 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*,[29] cf. Figure 2. The result in each case can again be obtained in multiple machine-readable formats. We believe that similar *aggregators* will play an important rôle in the emerging Semantic Web, especially as ontologies become more numerous and filtering methods become more complex.

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 [54].

Examples of large scale web applications include semantic search engines, such as the Creative Commons Search engine,[30] or Swoogle [20]. These applications are especially interesting since they provide services beyond mere display of data, and successfully 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 based on classical HTML data, but each of those implementations requires significant programming effort, is very sensitive to changes on the source sites, or relies on certain 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-

[26] http://www.ldodds.com/foaf/foaf-a-matic.html
[27] http://www.flickr.com/groups/api/discuss/72157594497877875/
[28] http://www.planetplanet.org/
[29] http://pipes.yahoo.com/
[30] http://search.creativecommons.org/

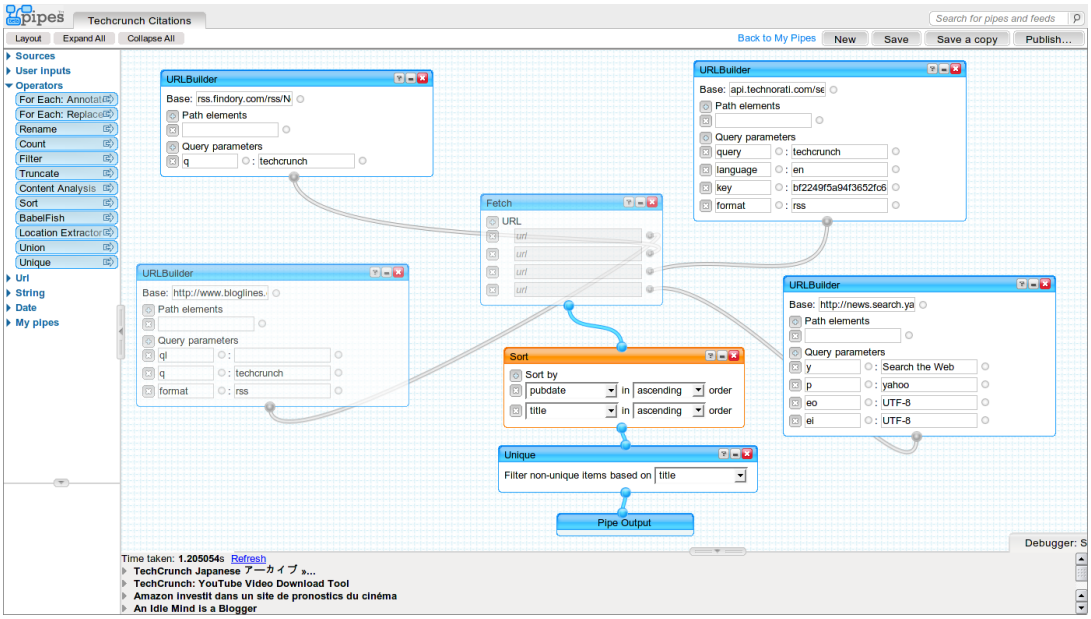

Figure 2: *Yahoo! pipes* (`http://pipes.yahoo.com/`) enables users to create custom mashups, thus successfully combining modern Web 2.0 interfaces with the advantages of machine-readable data feeds. The screenshot shows how multiple RSS feeds are aggregated, sorted, and filtered to produce a novel feed.

specified semantic formats, like iCalendar [19], Atom [41], vCard [18], hReview [22]. Such standards, especially the set of Microformats [31], 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 [8].

4. NEXT STEPS

The previous section may give the impression that the realisation of the Semantic Web is merely a question of implementation and continued user adoption, finally leading to what could be described as an elaborate version of RSS. But we do not intend to reduce the Semantic Web in such a way, and indeed believe that semantic technologies bear many further opportunities.

Our scenario meant to show that some of the technologies can be easily applied – albeit only within a restricted setting where only a few participants are involved in the data exchange. Thereby, we want to promote the adoption of semantic technologies in some specific domains, which we consider as an important first step in realizing the Semantic Web vision [6]. The merits of Semantic Web technologies must be proven and be communicated through real-life application scenarios. But even in this scenario, and even more when moving beyond this, both foundational and applied research faces new challenges, as open research questions turn into key success factors.

In the following, we point out a number of promising technologies that could become relevant on the Web, but also to remaining challenges and open research issues that relate to them. We expect these topics to become highly relevant in the short-term, but do not intend to make any claims of comprehensiveness.

4.1 Expressive ontologies

Light-weight ontology languages like RDF often are easier to handle, both for machines and for humans, than more complex formalisms like OWL. The impressive benefits that even simple machine-readable data can bring may lead some to believe that the Semantic Web will need not much expressiveness beyond RDFS. However, more powerful ontology languages have proven relevant in many application areas, and are likely to become increasingly important on the Web.

Indeed, complex knowledge often cannot be encoded without further expressivity, as became evident in many practical uses of semantic technologies. Examples include applications in medicine (see, e.g., [53]), or natural sciences (see, e.g., the *Halo* project[32] [25]). But additional expressivity is desirable even in cases of simple semantic data. On the one hand, ways of describing facts declaratively are crucial for querying knowledge bases. On the other hand, ontological knowledge can be exploited for query simplification and for formalising constraints, similar to the use of schematic information in relational databases [14, 40].

Clearly, expressive ontologies bear many challenges, some of which are also listed below. The majority of those challenges is actively addressed in current Semantic Web research, but typical Web 2.0 applications may bear additional requirements for the use of expressive ontologies. For instance, the use of complex ontologies in semantic wikis [50] requires users to be able to understand and control automatic inferences, supported by adequate software interfaces.

4.2 Scalability and tractability

Scalability and performance is a huge issue for the Semantic Web, as will be apparent when moving beyond a few semantically annotated websites. The sheer amount of data on the Web is as challenging as the desire for higher expressiv-

[31] `http://microformats.org`

[32] `http://www.projecthalo.com/`

ity. On an engineering level, this problem is addressed by increasingly powerful implementations. Classical Web pages, even if dynamic, often rely on controlled data sources that allow them to make good use of caching. With data sources being interconnected and dispersed all over the Semantic Web, the assumption of controlled data sources breaks, and caching must be reinvestigated and reimplemented based on the novel interaction model that arises with dynamic data-driven websites.

On a more foundational level, the careful design of powerful yet computationally manageable ontology languages needs to be continued. Recently, this research has lead both to an extension of the expressiveness of OWL DL [21], and to the identification of a number of much simpler but still useful ontology languages, such as $\mathcal{EL}$++ [3] or RDFS++ [35].

Investigations of the semantics and complexity of query languages also are an important contribution (see, e.g., [2]). On the Semantic Web, queries may also refer to data from different data sources, possibly even physically distributed. Aggregation of data, and the federation or distribution of queries are possible ways of addressing these problems.

4.3 Usability

Ease of use of both the user interface and the developer interface is essential. Many current Semantic Web tools still require expertise in semantic technologies and Web standards in order to be used, which can easily repel Web developers. It is possible – and necessary – to hide the complexity of the underlying technologies from the users and developers just as today's users are often unaware of the intricacies of XHTML, HTTP, or different encoding systems. It is therefore necessary to incorporate semantics into applications in ways that allow intuitive usage, as promoted for instance by tools like Semantic MediaWiki [33].

Still Semantic Web applications tend to burden people cognitively with their own internal semantic models and ontologies. Instead, there needs to be more understanding of the *"human-semantics interaction"* aspects of how people approach semantically rich applications, and ways for easing people into working with the semantic models underlying their software and tools. Unfortunately, there is a real dearth of work in this area with a few notable exceptions [7].

One of the strengths of the Semantic Web is its easy extensibility. You need a vocabulary to describe your collection of salt and pepper shakers, but can't find one? Go ahead, create it yourself! But creating a good vocabulary or ontology is hard, and users may rely on existing ontologies rather than to create their own. If none of the given ontologies truly fits the user's needs, this may reduces the quality of semantic annotations. Therefore, simple creation and evaluation of ontologies for the Web will become a much more practically important issue than it is now. There are promising results in this field of research, e.g. *Diligent* as a method [49], and [9] as an AJAX-based application for the collaborative construction of ontologies.

4.4 Trust and control

A particular challenge for distributed information systems is to be able to trace the origin of data. On the consumer side, this allows people to trust into data by trusting into its source. Content creators, on the other hand, have many reasons for being interested in tracking the exchange data that they published. In general, however, there is are few ways of establishing the trustworthiness of a particular creator or consumer in an open Web environment. One effect of this is that HTML meta-tags, though potentially useful and long established, are basically ignored by most search engines who cannot ensure that the specified information is trustworthy. Yet, humans can well pick reliable data sources – RSS feeds are an example of such selective use of data.

Provenance can either be established by digital signatures, as are massively used for signing emails or securing HTTP, or through a chain of trustworthy content providers that can be selected by users. In both cases, it is easy to remove data from those trustworthy contexts: consumers can find trustworthy sources, but creators still have no means of tracking their content. The latter observation has important ramifications. Controlling semantic data becomes very difficult, and private, confidential, or proprietary information can hardly be restricted. This is a known problem on today's Web, but the fine granularity of semantic data also prevents *watermarking* and similar methods that are currently used. Of course, any such discussion also includes a wide range legal aspects that is not being addressed yet.

4.5 Mapping and integration

For the discussed scenario, we assumed that the data being shared used one agreed-upon ontology that is common to applications in the domain. Similar domain-specific ontologies have been created for various purposes, examples including FOAF and SIOC [11], and they might be very suitable for basic data exchange. But in a true Web setting heterogeneous or overlapping conceptualisations are bound to appear. Since the exchange process requires a shared common understanding of the involved data, differences in the ontologies need to be aligned and reconciled. This is addressed by substantial work in the field of *ontology alignment*, see, e.g., [23], [44], and [24]), but further steps are needed to produce reliable mapping systems. It is also important to explore how far one the automatic identification of mappings can actually reach, and how semi-automatic methods could make efficient use of human judgements.

While correct mappings of any origin address the problem of data exchange, data integration adds some additional requirements. For instance, extensive use of ontologies and semantic data often requires extensions and modifications. Even if it is known in principle how two sources of knowledge should be integrated, they might have extended the common assumptions in incompatible ways. In the worst case, this might lead to logical inconsistency of information that would need repair [34], in other cases, it might require at least proper versioning of independently updated semantic data [28]. Both problems are relevant for the creation of advanced semantic mashups.

Finally, the alignment of instance data is also important. Most instances are not part of widely adopted ontologies, but are abundant in applications. There is a need for automatic fusion of data, including object identification and resolution of conflicts among data entries. As opposed to schema mapping, the large amount of instance data involved in normal use cases renders manual approaches unmanageable. Currently, much effort is devoted to this topic and promising results have been achieved. Linking Open Data[33]

[33] `http://esw.w3.org/topic/SweoIG/TaskForces/`

for instance is a Semantic Web community project that aims to produce more and link existing semantic data sources by means of equivalence mining and the development of publishing tools and converters.

5. CONCLUSIONS

The ideas underlying the Semantic Web and the Web 2.0 are often presented as competing visions for the future of the Web. Both communities have their own assumptions, cultures, and focal points. 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.

In this paper, we argued that basic web application scenarios, such as semantic blogging, are worthwhile goals for further developing 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. This does not say that foundational topics are banned from our vision of tomorrow's Semantic Web – without expressive ontology languages and the associated technologies and methodologies, one would quickly arrive at a downgraded *semantic web* that adds little on top of RSS feeds. Arguably the latter is a useful first step, but it would fail to live up to the full potential of semantic technologies.

We also think that semantic technologies, in turn, bear a great potential of providing a robust and extensible basis for emerging Web 2.0 applications. Interchange, distribution, and creative reuse of data can be greatly facilitated by the infrastructures that the Semantic Web offers. Web 2.0 efforts should take the opportunity to embrace those freely available technologies. 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.

6. ACKNOWLEDGEMENTS

We want to thank everybody who has engaged in fruitful discussions over the ideas described in this position 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 reviews. Research reported in this paper was supported by the EU in the IST projects X-Media (IST-FP6-026978, `www.x-media-project.org`) and NeOn (IST-2006-027595, `www.neon-project.org`). The view presented in this position paper is the authors' and not of the projects as a whole.

7. REFERENCES

[1] A. Ankolekar and D. Vrandečić. Personalizing Web surfing with semantically enriched peronal profiles. In M. Bouzid and N. Henze, editors, *Proc. Semantic Web Personalization Workshop*, Budva, Montenegro, June 2006.

[2] M. Arenas, J. A. Perez, and C. Gutierrez. Semantics and complexity of sparql. In I. Cruz and S. Decker, editors, *Proc. 5th International Semantic Web Conference (ISWC06)*, pages 30–43, Athens, GA, USA, 2006.

`CommunityProjects/LinkingOpenData`

[3] F. Baader, S. Brandt, and C. Lutz. Pushing the EL envelope. In *Proc. 19th Int. Joint Conf. on Artificial Intelligence (IJCAI'05)*, Edinburgh, UK, 2005. Morgan-Kaufmann Publishers.

[4] F. Baader, D. Calvanese, D. McGuinness, D. Nardi, and P. Patel-Schneider, editors. *The Description Logic Handbook: Theory, Implementation and Applications*. Cambridge University Press, 2003.

[5] G. Beged-Dov, D. Brickley, R. Dornfest, I. Davis, L. Dodds, J. Eisenzopf, D. Galbraith, R. Guha, K. MacLeod, E. Miller, A. Swartz, and E. van der Vlist. RDF Site Summary 1.0, 9 December 2000. Available at `http://web.resource.org/rss/1.0/spec`.

[6] T. Berners-Lee, J. Hendler, and O. Lassila. The Semantic Web. *Scientific American*, 5, 2001.

[7] A. Bernstein, E. Kaufmann, A. Göhring, and C. Kiefer. Querying ontologies: A controlled english interface for end-users. In *Proc. 4th International Semantic Web Conference (ISWC05)*, pages 112–126, November 2005.

[8] C. Bizer and A. Seaborne. D2RQ – treating non-RDF databases as virtual RDF graphs. In *Proc. 3rd International Semantic Web Conference (ISWC04)*, 2004.

[9] S. Braun, A. Schmidt, and V. Zacharias. Ontology maturing with lightweight collaborative ontology editing tools. In *Proc. Workshop on Productive Knowledge Work: Management and Technological Challenges (ProKW)*, 2007.

[10] T. Bray, J. Paoli, and C. M. Sperberg-McQueen. Extensible markup language (XML) 1.0 (second edition). W3C Recommendation REC-xml-20001006, World Wide Web Consortium (W3C), Oct. 2000. Available at `http://www.w3.org/XML/`.

[11] J. G. Breslin, A. Harth, U. Bojārs, and S. Decker. Towards semantically-interlinked online communities. In *Proc. 2nd European Semantic Web Conference (ESWC05), Heraklion, Greece, Proceedings, LNCS 3532*, pages 500–514, 2005.

[12] D. Brickley and R. V. Guha. RDF Vocabulary Description Language 1.0: RDF Schema. W3C Recommendation, 10 February 2004. Available at `http://www.w3.org/TR/rdf-schema/`.

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

[14] A. Cali and M. Kifer. Containment of conjunctive object meta-queries. In *Proc. 32nd Int. Conf. on Very Large Data Bases (VLDB06)*, pages 942–952. VLDB Endowment, 2006.

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

[16] S. Cayzer. What next for semantic blogging? Technical Report HPL-2006-149, Hewlett-Packard Laboratories, Bristol, UK, Oct. 2006.

[17] Creative Commons. "Some Rights Reserved": Building a layer of reasonable copyright. `http://creativecommons.org`.

[18] F. Dawson and T. Howes. vcard mime directory profile. RFC 2426, Internet Engineering Task Force,

Sept. 1998.

[19] F. Dawson and D. Stenerson. Internet calendaring and scheduling core object specification (icalendar). RFC 2445, Internet Engineering Task Force, Nov. 1998.

[20] L. Ding, T. Finin, A. Joshi, R. Pan, R. S. Cost, Y. Peng, P. Reddivari, V. Doshi, and J. Sachs. Swoogle: A search and metadata engine for the Semantic Web. In *Proc. 13th ACM Conf. on Information and Knowledge Management*, pages 58–61, 2004.

[21] B. C. G. (ed.). OWL 1.1 web ontology language, November 2006. Available at `http://owl1_1.cs.manchester.ac.uk/`.

[22] T. C. (ed.). hReview 0.3, 22 February 2006. Available at `http://microformats.org/wiki/hreview`.

[23] M. Ehrig and S. Staab. QOM – Quick ontology mapping. In *Proc. 3rd International Semantic Web Conference (ISWC04)*, pages 683–697. Springer, 2004.

[24] M. Ehrig and Y. Sure. Ontology mapping – an integrated approach. In *Proc. 1st European Semantic Web Symposium*, volume 3053, pages 76–91. Springer, 2004.

[25] N. Friedland, P. Allen, G. Matthews, M. Witbrock, D. Baxter, J. Curtis, B. Shepard, P. Miraglia, J. Angele, S. Staab, E. Mönch, H. Oppermann, D. Wenke, B. Porter, K. Barker, J. Fan, S. Y. Chaw, P. Yeh, D. Tecuci, and P. Clark. Project Halo: Towards a digital Aristotle. *AI Magazine*, 2004.

[26] J. Golbeck and J. Hendler. FilmTrust: movie recommendations using trust in web-based social networks. In *Proc. IEEE Consumer Communications and Networking Conference*, 2006.

[27] T. Heath and E. Motta. Reviews and ratings on the semantic web. In *Poster Track, 5th International Semantic Web Conference (ISWC2006)*, Athens, Georgia, USA, 2006.

[28] J. Heflin and J. Z. Pan. A model theoretic semantics for ontology versioning. In *Third International Semantic Web Conference*, pages 62–76, Hiroshima, Japan, 2004. Springer.

[29] U. Hustadt, B. Motik, and U. Sattler. Reducing $\mathcal{SHIQ}^-$ description logic to disjunctive datalog programs. In *Proc. of KR2004*, pages 152–162. AAAI Press, 2004.

[30] A. Kalyanpur, B. Parsia, E. Sirin, and J. Hendler. Debugging unsatisfiable classes in OWL ontologies. *Journal of Web Semantics*, 3, 2005.

[31] D. R. Karger and D. Quan. What would it mean to blog on the semantic web? In S. A. McIlraith, D. Plexousakis, and F. van Harmelen, editors, *Proc. 3rd International Semantic Web Conference (ISWC04), Hiroshima, Japan*, pages 214–228. Springer, November 2004.

[32] H. Knublauch, R. W. Fergerson, N. F. Noy, and M. A. Musen. The Protégé OWL plugin: An open development environment for Semantic Web applications. In *Proc. 3rd International Semantic Web Conference (ISWC04*. Springer, 2004.

[33] M. Krötzsch, D. Vrandečić, and M. Völkel. Semantic mediawiki. In I. Cruz and S. Decker, editors, *Proc. 5th International Semantic Web Conference (ISWC06)*, pages 935–942, Athens, GA, USA, 2006.

[34] J. Lam, J. Z. Pan, D. Sleeman, and W. Vasconcelos. A fine-grained approach to resolving unsatisfiable ontologies. In *Proc. of the 2006 IEEE/WIC/ACM International Conference on Web Intelligence (WI-2006)*, 2006.

[35] O. Lassila. Identity crisis and serendipity, May 2006. Available at `http://www.lassila.org/publications/2006/IdentityCrisisAndSerendipity.pdf`.

[36] A. Maedche, B. Motik, and L. Stojanovic. Managing multiple and distributed ontologies in the Semantic Web. *VLDB Journal*, 12(4):286–302, 2003.

[37] F. Manola and E. Miller. Resource Description Framework (RDF) primer. W3C Recommendation, 10 February 2004. Available at `http://www.w3.org/TR/rdf-primer/`.

[38] C. Marlow, M. Naaman, d. boyd, and M. Davis. HT06, tagging paper, taxonomy, flickr, academic article, to read. In *Proc. 17th Conf. on Hypertext and Hypermedia (HYPERTEXT'06)*, pages 31–40, 2006.

[39] P. Mika. Ontologies are us: A unified model of social networks and semantics. In *Proc. 4th International Semantic Web Conferences (ISWC05)*, pages 522–536, 2005.

[40] B. Motik, I. Horrocks, and U. Sattler. Integrating description logics and relational databases, Dec 6, 2006. Technical Report, University of Manchester, UK.

[41] M. Nottingham and R. Sayre. The atom syndication format. RFC 4287, Internet Engineering Task Force, Dec. 2005.

[42] 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`.

[43] E. Oren, R. Delbru, and S. Decker. Extending faceted navigation for rdf data. In I. Cruz and S. Decker, editors, *Proc. 5th International Semantic Web Conference (ISWC06)*, pages 559–572, 2006.

[44] R. Pan, Z. Ding, Y. Yu, and Y. Peng. A Bayesian network approach to ontology mapping. In *Proc. 4th International Semantic Web Conference (ISWC05)*, 2005.

[45] A. Seaborne and E. Prud'hommeaux. SPARQL query language for RDF. Technical Report `http://www.w3.org/TR/2006/CR-rdf-sparql-query-20060406/`, W3C, April 2006.

[46] J. Seidenberg and A. Rector. Web ontology segmentation: Analysis, classification and use. In *Proc. 15th Int. Conf. on World Wide Web (WWW 2006), Edinburgh, Scotland, May 23–26*, 2006.

[47] M. K. Smith, C. Welty, and D. McGuinness. OWL Web Ontology Language Guide, 2004. W3C Recommendation 10 February 2004, available at `http://www.w3.org/TR/owl-guide/`.

[48] Y. Sure, S. Staab, and R. Studer. On-to-knowledge methodology. In S. Staab and R. S. (eds.), editors, *Handbook on Ontologies*, Series on Handbooks in Information Systems, chapter 6, pages 117–132. Springer, 2003.

[49] C. Tempich, H. S. Pinto, Y. Sure, and S. Staab. An argumentation ontology for distributed, loosely-controlled and evolving engineering processes

of ontologies (DILIGENT). In *Proc. 2nd European Semantic Web Conference (ESWC05), LNCS 3532*, pages 241–256. Springer, 2005.

[50] D. Vrandečić and M. Krötzsch. Reusing ontological background knowledge in semantic wikis. In *Proceedings of 1st Workshop "From Wiki to Semantics" (SemWiki'06)*, 2006.

[51] D. Vrandečić, H. S. Pinto, Y. Sure, and C. Tempich. The DILIGENT knowledge processes. *Journal of Knowledge Management*, 9(5):85–96, Oct 2005.

[52] M. Völkel, M. Krötzsch, D. Vrandečić, H. Haller, and R. Studer. Semantic Wikipedia. In *Proc. 15th Int. Conf. on World Wide Web (WWW 2006), Edinburgh, Scotland, May 23–26*, 2006. Available at `http://www.aifb.uni-karlsruhe.de/WBS/hha/papers/SemanticWikipedia.pdf`.

[53] K. Wolstencroft, P. Lord, L. Tabernero, A. Brass, and R. Stevens. Using ontology reasoning to classify protein phosphatases. *8th Annual Bio-Ontologies Meeting 2005*, 24, 2005.

[54] V. Zacharias and M. Sibler. Semantic announcement sharing. In *Proc. Fachgruppentreffen Wissensmanagement*, 2004.