How to Combine the Best of Web2.0 and a Semantic Web: Examples from Revyu.com

Tom Heath and **Enrico Motta**
Knowledge Media Institute, The Open University, UK
t.heath@open.ac.uk, e.motta@open.ac.uk

It is not uncommon to view Web2.0 and the Semantic Web as mutually exclusive, competing paths to the Web of the future, each advocated by a distinct community. We argue that the two approaches are in fact complementary, and that both face challenges the other can solve, such as how to integrate Web2.0 data on a Web scale, and how to enable users to create semantically rich annotations. Here we will use examples from Revyu.com to demonstrate how features of Web2.0 and the Semantic Web can be combined in a single service that overcomes these challenges. Revyu[1] is a Web site where people can review and rate anything they choose. The site is built on Semantic Web technologies, but also uses common features of Web2.0, such as keyword tagging.

Web2.0 provides an umbrella label for myriad applications that elicit and reuse user-generated content, support social and collaborative interaction on the Web, and provide engaging user interactions based on AJAX. The Semantic Web vision is one of data published in machine-readable formats, given formal semantics through the use of shared ontologies, and interlinked on a Web scale. By making Web data more open to processing by machines, the Semantic Web fundamentally aims to bring tangible benefits to users. Despite both approaches' revolutionary potential, barriers exist to this being fully realised.

Web2.0 applications have elicited vast amounts of user-generated content, such as wiki entries, tagged photos, and links joining people in social networks. However, at present, most such applications represent walled data gardens from which information cannot be easily set free and combined with other sources. This can lead to the un-Web-like situation where my friend in Orkut is a stranger on MySpace. Overcoming this requires services to publish data in formats easily processed by third parties.

The Semantic Web offers a platform on which to do just this. Publishing data in RDF lowers the barriers to its reuse by others. Two applications may choose to describe their data using the same schema, or they may not. This is immaterial, provided mappings can be defined between the two data models. However, using elements from existing ontologies does significantly streamline the data integration process. Key to the power of RDF is the ability within one document to mix statements that use elements from any number of ontologies, without the document itself needing to validate against a fixed schema.

For example, in addition to XHTML, Revyu exposes reviews in RDF using elements from the Review vocabulary[2] and the FOAF ontology[3] (for describing people). Adopting these popular ontologies make Revyu data instantly interoperable with that from other sources. Creating a Revyu-specific ontology that was then mapped to others would have been an equally valid, albeit more complex process. Acknowledging that any semantics are better than none, Revyu also exposes reviews using the hReview microformat[4] embedded in XHTML pages.

Services that offer data APIs, such as Amazon or Flickr, go a long way to dismantling the walls of the data garden. However, barriers to the reuse of this data still exist. Data structures may be truly novel, requiring new mappings between every additional data source that is integrated. Even more critically, where vanilla XML is used, integrating data from different sources may involve rewriting existing schemas if new information is to be incorporated and republished in the same document. RDF does not suffer the same limitation, allowing statements to be arbitrarily combined in one document, made in different locations but referencing the same identifier, or made in different locations using different identifiers but stating that both identifiers refer to the same thing.

This last scenario is nicely illustrated by Revyu's RDF descriptions of reviewers. New URIs are *minted* for each reviewer, who may already have several others. Where a reviewer maintains their own RDF description in another location, this is processed to retrieve their additional URIs. Statements are then added to their Revyu RDF, stating that all URIs identify the same person. Human users also benefit, as additional information about the reviewer (e.g. photos or homepage links) can be retrieved from external RDF files and used to enhance their profile page (as illustrated below), without it being duplicated in Revyu.

Web Services that publish vanilla XML present application developers with the additional challenge of parsing XML trees to retrieve the desired data. Whilst most programming languages make this task trivial, data processing remains tied to the underlying syntactic rather than semantic structure of the data. Creating Web2.0 mashups consequently requires the writing of custom handlers to interact with each API. No common language is available for querying and integrating such data sources. This issue must be overcome if mashups are to be created on a truly flexible, Web scale.

The SPARQL query language for the Semantic Web enables standardised access to distributed data sources.

[1] http://revyu.com
[2] http://purl.org/stuff/rev#
[3] http://xmlns.com/foaf/0.1/
[4] http://microformats.org/wiki/hreview

SQL-like queries can be executed as HTTP GET requests against remote "endpoints", returning data that can be processed using standard code, irrespective of the endpoints underlying implementation. Developers must simply know the structure of the RDF graph behind the endpoint in order to write the appropriate query. Revyu exposes review, people, and tagging data via its SPARQL endpoint. At present, additional information about books reviewed on Revyu is retrieved via the Amazon Web Services API and displayed to users of the site, as shown in the figure below. However, creation of the RDF Book Mashup[5] means Amazon-specific code will shortly be replaced by a single SPARQL query.

Underpinning such data integration and Web-scale mashup exercises is the need for widespread minting of URIs identifying non-Web resources, to enable linking of data from disparate locations. As for reviewers, URIs are assigned to every item reviewed on Revyu, the review itself, and all tags used. Consequently, many things in the offline world now have URIs against which further RDF statements can be made, on Revyu or elsewhere. Things such as restaurants or pubs are unlikely to mint their own URIs in the near future. Consequently Revyu provides valuable infrastructure on which to build next generation mashups. Revyu URIs can all be dereferenced, responding with HTTP303 redirects, according to the W3C TAG's finding on the httpRange-14 issue[6].

Having discussed how Semantic Web approaches may address challenges facing Web2.0, we will now examine the inverse relationship. We argue that the Semantic Web vision faces challenges of equal significance if it is to reach widespread adoption. In our view these are focused on two issues: availability of data and interaction design.

Initiatives such as dbpedia[7] are bootstrapping the Semantic Web by RDF-ising existing data sets. However, in stark contrast to the conventional Web and Web2.0, few mechanisms currently exist allowing non-specialist users to contribute to the Semantic Web. Early growth of the Web is widely attributed to individuals creating personal sites by copying/pasting HTML code or using visual editors. Such approaches may not be appropriate to a Semantic Web. However, there are many Web2.0 applications enabling regular users to contribute content without specialist skills. With few exceptions, similar tools enabling grassroots publishing on the Semantic Web are not currently available. Revyu is one exception.

By adhering to the well established interaction pattern of completing forms in a Web browser, Revyu allows users to create content that is immediately usable on the Semantic Web. This occurs without any user knowledge of RDF, ontologies, or even the principles of the Semantic Web. In our view, specific, focused applications that guide user input through the use of forms represent the most promising way to elicit semantic annotations from regular Web users.

In addition, Revyu lifts the burden of classifying reviewed items according to complex, existing, or relatively fixed taxonomies, through the use of keyword tagging. This creates greater flexibility in what can be reviewed, and we argue, lowers the barrier to contribution of reviews. A less desirable consequence of this feature is that machine-readable statements regarding the nature of reviewed items cannot be made with any confidence. Further investigation is required as to how this may be achieved.

Assuming that sufficient data can be produced, the Semantic Web faces a further challenge: creating interfaces allowing non-specialist users to exploit it. Humans have thousands of years of experience creating and using textual documents, and decades of experience with hypertext systems. The Semantic Web is not constrained to the notion of a document in the same way as the conventional Web of HTML pages. The question remains of how we design compelling, coherent, and usable interactions based on data from multiple sources, in such a way that its source, trustworthiness, and value can be determined. Map-based Web2.0 mashups provide some clues in this direction, by presenting alternative visualisations for complex data sets. How such approaches may scale on a Web of infinitely interlinked data is not clear at present. This remains an open and pressing research question for the Web community.

Conclusions: This paper has sought to highlight distinct challenges facing the Web2.0 and Semantic Web communities, identify potential solutions each community may offer the other, and illustrate with examples from Revyu how these may be realised. In conclusion we make the following recommendations to each community.

Firstly, that the Web2.0 community: gives serious consideration to publishing data in forms that are more easily reusable, such as RDF; investigates the use of SPARQL for remote data access rather than custom APIs; and mints URIs for offline items that are distinct from the URIs of documents describing them.

Secondly, we argue that the Semantic Web community must give urgent attention to creating interfaces allowing regular Web users to contribute to the Semantic Web. This should not take the form of more usable editors for ontologies or RDF instance data (whilst these would undoubtedly be useful), but seek to exploit familiar interaction patterns. Revyu's form-based approach is no doubt just one of many options. In tandem, significant effort must be given to developing compelling interfaces able to display structured, linked data from across the Web. Mashups have set the standard for such interfaces and interactions. The next generation must demonstrate the unique benefits of a Web of interlinked data.

Lastly, whilst acknowledging that we have made distinctions of this nature here, we suggest that viewing Web2.0 and Semantic Web approaches as mutually exclusive is detrimental to all those involved, and to the development of the Web as a whole.

[5] http://sites.wiwiss.fu-berlin.de/suhl/bizer/bookmashup/
[6] http://www.w3.org/2001/tag/issues.html#httpRange-14
[7] http://dbpedia.org/