Ease of interaction plus ease of integration: Combining Web2.0 and the Semantic Web in a reviewing site

Tom Heath*, Enrico Motta

Knowledge Media Institute and Centre for Research in Computing, The Open University, Milton Keynes MK7 6AA, UK

Abstract

Web2.0 has enabled contributions to the Web on an unprecedented scale, through simple interfaces that provide engaging interactions. This wealth of data has spawned countless mashups that integrate heterogeneous information, but using techniques that will not scale beyond a handful of sources. In contrast, the Semantic Web provides the key to large-scale data integration, yet still lacks approachable interfaces allowing contributions from non-specialists. In this paper we present Revyu, a reviewing and rating site in the Web2.0 mould that is built on Semantic Web infrastructure and both publishes and consumes linked RDF data. This combination of approaches affords ease of interaction for regular users and ease of integration with external data sources.

Keywords: User interaction; Data integration; Reviews; Semantic Web; Web2.0

1. Introduction

Web2.0 and the Semantic Web have previously been viewed as mutually exclusive, competing paths to the Web of the future, each advocated by a distinct community [1]. In this paper we demonstrate that the two approaches are in fact complementary, and that each faces challenges the other can solve: Web2.0 data is not generally available in forms that facilitate its easy interlinking and reuse, whilst the Semantic Web has yet to embrace the ease of participation that has enabled Web2.0 to reach such wide audiences.

Drawing on examples from Revyu [16] we argue that core Semantic Web technologies provide a basis for integrating Web2.0 data on a large scale, and demonstrate how the interaction paradigms of Web2.0 can allow non-specialist users to create semantic annotations. Revyu is a public Web site launched in November 2006 at http://revyu.com, where people create reviews and ratings of anything they choose. The site follows the Web2.0 model of user contribution through form-based interactions. Use of Semantic Web technologies throughout the site (a way that is hidden from users) allows concurrent publication of reviews in HTML and RDF, and the interlinking and integration of Revyu content with data from other sources.

1.1. Web2.0 and the Semantic Web: definitions

The following definitions of Web2.0 and the Semantic Web will be used in this paper. Web2.0 is 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. High-profile services to which the label has been applied include Wikipedia, Facebook, Flickr, and Google Maps. The Semantic Web vision is one of data published on the Web in machine-readable formats, given formal semantics through the use of shared ontologies, and interlinked on a massive scale [23]. These three ingredients enable large-scale data integration, ultimately for the benefits of users [18].

2. Limitations of Web2.0 and the Semantic Web

2.1. Web2.0 data is hard to reuse and interlink

Web2.0 has produced many ‘killer-apps’: Wikipedia, Flickr, and Facebook have elicited vast numbers of wiki entries, tagged
photos, and links joining people in social networks. These forms of contribution are often referred to as ‘user-generated content’. However, at present much of this content is confined to ‘silos’ or ‘walled data gardens’, or published in formats that hinder its reuse. This prevents easy integration of the content with data from other sources, leading to the un-Web-like situation where a friend on Facebook is a stranger on MySpace, as the social network defined in one service cannot be used to populate the other. Overcoming this requires data to be published in formats that are easily processed by third parties, that are more expressive than simple syndication formats such as RSS, and that afford interlinking with other data on the Web.

APIs such as those offered by Amazon and Flickr go some way to addressing this issue, however barriers to the reuse of this data still exist. No common query language is implemented across Web2.0 APIs. Application developers must generally parse 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, and integrate data from, each API.

Publishing data using the Resource Description Framework (RDF) [26] conveys a number of benefits relative to ‘vanilla’ XML: it lowers the barriers to data reuse by third parties, makes data accessible via a standard query language (SPARQL [22]), eases the integration of data from different sources, and allows machine-readable links to be created between data sources. Integrating XML data from different sources into one document requires that all data conform to the same schema. In practice, much XML data from Web2.0 APIs is integrated at the level of programme code (at great cost in terms of development effort) and only republished in HTML, thereby hindering its reuse. RDF does not suffer these limitations; data can be arbitrarily combined into one document without this document needing to validate against a specific schema. Statements can be made anywhere on the Web about a particular resource; different statements may reference the same URI, or use different URIs but state that both identify the same resource.

Whilst an XML Schema may define a <uri></uri> element to be populated with the URI of some item, the semantics of this relationship are not explicit. Consequently, and in contrast to RDF, machines cannot infer links between data based on such elements. This situation is analogous to enclosing a URL in <span></span> tags within an HTML document (without using anchor tags <a href=””></a>) and expecting applications to interpret this string as a link.

2.2. The Semantic Web is (largely) closed to user contributions

Initiatives such as DBpedia [2] and the broader Linking Open Data project [6] are bootstrapping the Semantic Web, primarily by transforming to RDF and interlinking large, existing data sets. These initiatives are of great value in providing a base level of linked RDF data on the Web. However, few mechanisms currently exist that allow non-specialist users to contribute to the Semantic Web. This is in stark contrast to both the conventional Web and Web2.0. Early growth of the Web is widely attributed to individuals creating personal sites by copying and pasting HTML code. Whilst this approach may not be appropriate to a Semantic Web (novice users may not understand the semantics of statements contained in copied code), Web2.0 applications have demonstrated that regular users can 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.

In an evaluation of Semantic Web applications deployed to members of the Semantic Web community [14] it was found that the usability of applications hindered their uptake, even by those knowledgeable in the field. Consequently the usage of these tools to create semantic annotations was relatively low. In the light of these findings, tools that make semantic annotation accessible to non-specialist and specialist alike are required if the Semantic Web is to see the degree of user engagement enjoyed by previous generations of the Web.

3. Revyu combines Web2.0 interaction with Semantic Web data integration

3.1. System overview

Revyu is a Web site where users can create reviews and ratings of anything they wish. To the non-specialist Revyu appears like any regular Web site: little indication is given that the site is based on Semantic Web infrastructure (Fig. 1). Users can search or browse the site to read existing reviews, descriptions of things reviewed on the site, and profiles of reviewers. At the time of writing (June 2006), 100 reviewers have created a total of 381 reviews. The site currently receives between 2500 and 3000 unique page requests per day on average (a figure that is growing steadily), the majority of which originate from search engine queries. Such a high ratio of site usage to contribution suggests that whilst they are valued by many, a relatively low proportion of people are motivated to write reviews. This assertion has not been tested empirically; however it may indicate that user generation of content is not as prevalent as is widely assumed.

If a user wishes to contribute a review to Revyu they can do so by registering with the site and completing a simple web form. The form asks users to provide a name for the thing they wish to review, the text of their review, a numerical rating (on a scale of 1–5), some keyword tags describing the thing being reviewed, and one or more links to related Web resources. Users are presented with tag suggestions based on syntactically similar tags that have already been used in the system, helping to ensure consistency in tag usage. No further information is requested about the item being reviewed. Instead Revyu attempts to harvest relevant information from external Semantic Web data sources (a process described in more detail later in this paper).

Upon submission, the review and all related information (such as tags and Web links) is stored as RDF triples in the underlying triplstore. This is based on a de-normalised MySQL database and accessed using RAP, a PHP library for working
with RDF data [21]. Submitted data is immediately accessible on the site in both HTML and RDF/XML, as well as via a SPARQL ‘endpoint’¹ that allows remote querying of Revyu data. The SPARQL query language for the Semantic Web [22] enables standardised access to distributed data sources. 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. This contrasts with Web2.0 APIs where each requires custom code to handle query results.

Revyu uses the Review RDF vocabulary [3] to describe reviews, the FOAF ontology [7] to describe reviewers, and the Tag ontology [20] to describe bundles of tags associated with things reviewed on the site. Adopting these popular ontologies makes 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, that would have brought few benefits.

Revyu also exposes reviews using the hReview microformat[8] embedded in XHTML pages. This makes Revyu content accessible to applications that currently support microformats but not RDF. Whilst popular among sections of the Web2.0 community, microformats do not provide the same data integration and linking capabilities of RDF.

Users viewing the Revyu site with a conventional Web browser will never be exposed to the underlying RDF data unless they explicitly request it, either by clicking a link in HTML pages on the site or by sending appropriate Accept: headers in their HTTP request.

The primary significance of Revyu lies in its combination of an approachable interface with the creation of Semantic Web data, whilst also demonstrating (in a live system) current best practices in serving and consuming linked RDF. These themes will be discussed in the following sections.

3.2. Web2.0–style interaction

Web2.0 applications and services have enabled non-specialist users to contribute to the Web on a scale that, whilst in line with the original vision of a read-write Web, was previously unimaginable. This has been achieved by providing simple, well-structured Web forms through which users can, for example, tag photos or bookmarks, edit wiki entries, or write blog posts, using just their Web browser.

By adhering to this well-established interaction pattern, Revyu allows users to create content that is immediately usable on the Semantic Web, without requiring any knowledge of the underlying technologies or principles. In our view, such specific, focused applications that guide user input are the most promising means to elicit semantic annotations from regular Web users.

3.3. Tagging not classification

Revyu does not require users to classify reviewed items according to an existing taxonomy. Instead they can tag an item being reviewed with one or more descriptive keywords. This has a number of advantages: it lowers the barrier to contribution of reviews, as users do not need to locate the appropriate category in an existing, fixed classification; it avoids the need for one super-taxonomy of items that might be reviewed; it creates greater flexibility in what can be reviewed, as the user is not limited to reviewing certain classes of items; lastly it allows for sharing of knowledge that might be not be easily categorised but can be described with a few keywords.

Data about tags associated with reviewed items (e.g. when they were added, and by whom) is described using the Tag

¹ http://revyu.com/sparql.
ontologies, tag-based search, and other initiatives such as the Tag Commons [11].

3.4. From tags to semantics

Keyword tagging reduces the burden on the user by removing the need to classify items. Instead this burden is transferred to Revyu if we wish to provide additional functionality based on the type of reviewed item. Keyword tags alone are not a reliable basis for inferring type information: for example, the tag ‘film’ may refer to a movie film or a brand of photographic film. This ambiguity means we cannot assume that all items tagged ‘film’ are movies. Therefore, by default Revyu makes no assumptions about the type of reviewed items based on how they have been tagged, and adds no rdf:type statements other than owl:Thing to the triplestore. Instead we use a number of mechanisms to derive more detailed type information from a combination of tags and external data sources. At present heuristics exist for identifying books and films reviewed in Revyu (these are described below), with plans to add similar functionality for music albums, and amenities such as pubs, restaurants and hotels.

3.4.1. Identifying films on Revyu

The majority of contemporary films have homepages, which are generally provided by the film studio but carry little if any machine-readable data about the picture. However, coverage of films is very high in Wikipedia, which provides an external source against which we can verify Revyu data by querying the DBpedia SPARQL endpoint. We use the following heuristic to identify films: for each reviewed item tagged ‘film’ or ‘movie’, look for items in DBpedia with the same name and of type ‘Film’. For any items for which this heuristic returns a match, an rdf:type statement is added to the Revyu triplestore asserting that this item is a film. This type information is used to retrieve additional information about the reviewed item for display on the site, as described below.

3.4.2. Identifying books on Revyu

Whilst Wikipedia (and thus DBpedia) has extensive coverage of films, the coverage of books is less comprehensive; therefore we use a different heuristic to identify books reviewed on Revyu. When reviewing books, reviewers often place links to the book’s page on Amazon in the ‘Other Links’ field. These links are parsed and analysed to extract ISBN numbers. If a valid ISBN is identified then an rdf:type statement is added to the Revyu triplestore asserting that this item is a book. Again, we use this type information to retrieve additional information about the item, as discussed below. We may in future extend this heuristic to look up all items tagged ‘book’ against an external data source, however at present the current approach produces acceptable results.

3.4.3. Identifying related tags

It is not uncommon for tags to consistently co-occur with certain other tags. We use an algorithm to identify related tags (above a certain threshold of co-occurrence, to avoid identifying spurious connections), log them to the triplestore, and then republish them using the skos:related property of the SKOS vocabulary [19], asserting that these two concepts are related. This makes these conceptual relationships accessible to other applications wishing to find information about connections between tags. For example, our algorithm finds that ‘pub’ is related to ‘beer’ and ‘food’. Finding co-occurrence relationships between tags is certainly not unique to Revyu; what makes our work more novel is the republishing of these relationships to the Web in RDF. At present we do not attempt to link tags to other concepts in e.g. WordNet, as the results are too unreliable, especially when dealing with homonyms. However, as the techniques described in [24] mature we will apply these in order to better integrate Revyu tags with the Semantic Web.

3.5. Exploiting external data sources

Revyu uses a number of external data sources to supplement the basic reviews, tags, and links provided by reviewers. By exploiting these sources we provide rich information on the site but place minimal burden on the user to supply information.

3.5.1. Supplementing book and film reviews with external data

Having determined the rdf:type of reviewed books and films using the heuristics described above, we retrieve additional data about the item from external sources and use this to supplement reviews with information about the item. For items identified as books we automatically dereference their RDF Book Mashup URIs [5] to retrieve author information and the URL of a cover image (provided by the Amazon Web Services API) as an RDF graph. These additional pieces of information are then displayed alongside reviews of the item, as shown in Fig. 2.

Where items have been identified as films we automatically retrieve information such as the director and the URL of the film poster by querying the DBpedia SPARQL endpoint, and display this alongside reviews of the film. This mashup of review and film information is illustrated in Fig. 3.

We note that this form of human-oriented mashup provides no immediate user benefit over conventional Web2.0 mashups. However, this approach does bring two significant benefits, for the developer, and for the Semantic Web at large. Firstly the development effort expended in creating mashups is substantially reduced, as a common toolset (e.g. the SPARQL client of the RAP library) can be used to query both data sources. Secondly, links can be set between the Revyu URI for an item, and its identifiers at other services such as the RDF Book Mashup and DBpedia. By making and exposing these links in RDF (as described below) we help to populate the Semantic Web with links between data sets, creating a Web of Linked Data [4].
3.5.2. Supplementing reviewer information with FOAF data

When reviewers register with the site, they are only asked to supply minimal information: an email address, screenname, and password. Where a reviewer maintains their own RDF (i.e. FOAF) description in another location they may also supply its URI. In this case Revyu retrieves and processes this file to obtain additional information the reviewer chooses to share about themselves, such as photographs, homepage links, interests, and locations. This information is used to enhance the reviewer’s profile page (as illustrated in Fig. 4), thereby using the data integration capabilities of a Semantic Web to provide the kind of rich user profiles often associated with Web2.0 applications, without the information needing to be duplicated in Revyu. In addition, where a user knows another reviewer they can choose to add this person to their network by simply clicking a link. This creates additional foaf:knows statements in the Revyu triplestore which are then republished in the reviewer’s RDF description, and can be combined with other FOAF data from the Web to provide an integrated definition of the user’s social network.

3.6. Enabling and creating linked data

To enable linking between Revyu data and external sources, all entities on Revyu (things, reviews, reviewers, and tags) are given URIs. Adhering to the principles of Linked Data [4] these URIs can all be dereferenced, responding with HTTP 303 redirects according to the W3C TAGs finding on the httpRange-14 issue [27]. Where possible links are made between Revyu URIs and those minted by third parties. For example, where a reviewed film or book is found to exist in DBpedia or the RDF Book Mashup, owl:sameAs statements are added to the Revyu triplestore to record that both URIs identify the same item. Likewise, owl:sameAs statements are made between a reviewer’s Revyu URI and the URI they use in their FOAF description. These statements are then republished in the reviewer’s RDF description on Revyu. As more Semantic Web data is published according to Linked Data principles further linking opportunities will be created. This will in turn provide opportunities for increasingly compelling user applications.
3.7. Applications of Revyu data

Data from Revyu has many existing and potential applications. Providing multiple routes for accessing Revyu data (Javascript, RSS, RDF, and SPARQL) allows site users to easily syndicate reviews from the site for reuse in their own applications. At present uses of the data do not differ greatly in functional terms from Web2.0 syndication approaches using RSS. However, as increasing amounts of linked RDF data become available on the Web, Revyu will play a key role in an ecosystem of reusable review data which may be used to enhance existing sites with review-based functionality.

If other sites that support reviews and ratings could be persuaded to publish their data as linked RDF, a Web-wide aggregator of review data would become a possibility. The effort required to create such a system by scraping conventional Web sites or by integrating data from Web2.0 APIs is prohibitive on a very large scale. Semantic Web technologies provide the means to aggregate and integrate data in this way.

We are currently implementing a system that uses Revyu data to support personalised information seeking within one’s social network [15]. Not only do we aggregate reviews from networks of known individuals (using the information integration capabilities of RDF), we then rank the potential trustworthiness of individuals as information sources for a particular query. Ranking is based on a rich trust model of information- and recommendation-seeking in social networks [17].

The model, which resulted from a number of empirical analyses, defines five dimensions used by people to determine the trustworthiness of recommendation sources. These are experience, expertise, impartiality, affinity, and track record. Specifically our system exploits automatically generated trust metrics describing an individual’s experience of and expertise regarding a particular topic, and his or her affinity to others. How these metrics are applied in turn depends on the relative criticality and subjectivity of the task in question. RDF provides a common model with which we can aggregate data from Revyu and other sources as a basis for calculating these trust metrics. Once computed, these trust metrics can also be published on the Web in RDF, for consumption by other applications.

4. Related work

The idea of using RDF to publish reviews is not new. Revyu goes beyond the work of Guha [12] by implementing an open rating system that supports the reviewing and rating of anything, not just web content. The trust metrics used in our ongoing work (as described above) are more fine-grained than simply trust/distrust, and are computed automatically without relying on manual ratings of others in the network.

The FilmTrust system [10] solicited film reviews from users and republished them in RDF. However, the system is constrained to film reviews only, reviewed films are not annotated in any way beyond the rating/review, the accumulated reviews can not be queried programmatically, and the system does not integrate data from or link widely to other sources. This highlights the shortage of systems that take a comprehensive approach to the reviewing process. For example, Epinions3 provides a large number of reviews, but supports a limited notion of reputation, trust, and social networking, operates on a closed world of products and people, and does not integrate with external data. The same criticism applies to TrustedPlaces.4

The socially oriented music site Last.fm5 recommends music based on the taste overlaps of its users, mined from listening data. This approach arguably creates a more sensitive measure of trust between users than those based on manual trust/distrust ratings; however these affinities are music-specific, so cannot

---

3 http://www.epinions.com/
4 http://trustedplaces.com/
5 http://www.last.fm/
be guaranteed to scale to other domains. Once again the source
data for recommendations and trust calculations is limited to
a closed world. Floating the Music [9], another music recom-
mender system, does integrate data from a number of different
sources, such as user FOAF files, and profiles on Web2.0 music
sites. However, it does not provide obvious means to create addi-
tional annotations, nor does it link data from different sources
and publish aggregated data to the Semantic Web.

Whilst generic semantic annotation mechanisms such as
Semantic Mediawiki [25] have generated considerable inter-
est and gained some noteworthy uptake in sites such as
DiscourseDB, they may not be sufficiently usable or sufficiently
compelling to elicit semantic annotations from non-specialists.
Conversely, applications exist that, like Revyu, allow users to
create domain-specific annotations. PhotoStuff [13] is a desk-
top application enabling semantic annotation of photographs;
however, we argue that its implementation as a desktop applica-
tion limits its ease of interaction for users, compared to Webtop
applications such as Flickr.

5. Conclusions

This paper has sought to highlight distinct challenges facing
the Web2.0 and Semantic Web communities, and illustrate with
examples from Revyu how these may be resolved. In conclusion
we make the following recommendations to each community.
Firstly, that the Web2.0 community: gives serious considera-
tion to publishing data in forms that allow greater reuse and interlink-
ing, such as RDF; investigates the use of SPARQL, rather than
custom APIs, for remote data access; and mints dereferenceable
URIs that adhere to the httpRange-14 finding [27].

Secondly, we argue that the Semantic Web community must
give urgent attention to creating interfaces that allow 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 data from across the Web.

Humans have thousands of years of experience creating and
using textual documents, and decades of experience with
hypertext systems. Unlike the conventional Web of interlinked
documents, the Semantic Web is a Web of interlinked data. 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 deter-
mined. Mashups have set the standard for such interfaces and interactions. The next generation must demonstrate the unique benefits of a Web of data.

If other sites join Revyu in publishing reviews in RDF, and
reference the same URIs, large-scale aggregation of reviews
from many sources, that would be highly complex using Web2.0
approaches, becomes trivial using Semantic Web technologies.

The potential then exists to create RDF-based mashups that are
infinite in nature, integrating data from arbitrary sources as
required. To the best of our knowledge, no one has yet
implemented an open system that combines reviews, social net-
works, and recommendations with a task-sensitive, empirically
grounded, multi-dimensional trust model. Revyu represents the
most significant progress in this direction.

Acknowledgements

This research was partially supported by the Advanced
Knowledge Technologies (AKT) and OpenKnowledge (OK)
projects. AKT is an Interdisciplinary Research Collaboration
(IRC) sponsored by the UK Engineering and Physical Sciences
Research Council under grant number GR/N15764/01. OK is
sponsored by the European Commission as part of the Informa-
Society Technologies (IST) programme under grant number
IST-2001-34038.

References

[1] A. Ankolekar, D. Vrandecic, M. Krötzsch, D. Thanh Tran. The two cultures:
mashing up Web 2.0 and the Semantic Web, in: Proceedings of the 16th
Extracting Semantics from wiki content, in: Proceedings of the 4th Europe-
(accessed 1st June 2007).
LinkedData.html (accessed 1st June 2007).
to a web of data, in: Proceedings of the 3rd Workshop on Scripting for the
web, in: Proceedings of the Demonstrations Track, 4th European Semantic
xmils.com/foaf/0.1/ (accessed 1st June 2007).
0.3 draft specification. http://microformats.org/wiki/hreview (accessed 1st
June 2007).
[9] O. Celma, M. Ramirez, P. Herrera, Getting music recommendations and
filtering newsfeeds from FOAF descriptions, in: Proceedings of the 1st
Workshop on Scripting for the Semantic Web, 2nd European Semantic
[10] J. Golbeck, J. Hendler, FilmTrust: movie recommendations using trust in
web-based social networks, in: Proceedings of the IEEE Consumer Com-
Hendler, PhotoStuff—an image annotation tool for the semantic web, in:
Proceedings of the Poster Track, 4th International Semantic Web Confer-
ence (ISWC2005), 2005.
to successfully deploying semantic web technologies, in: Proceedings of the
3rd International Semantic Web User Interaction Workshop (SWUI2006),
[15] T. Heath, E. Motta, Personalizing relevance on the semantic web through
trusted recommendations from a social network, in: Proceedings of the
Workshop on Semantic Web Personalization, 3rd European Semantic


