Available online at www.sciencedirect.com
ScienceDirect
Web Semantics: Science, Services and Agents on the World Wide Web 6 (2008) 21–28
JOURNAL OF Web Semantics
www.elsevier.com/locate/websem

Using the Semantic Web for linking and reusing data across Web 2.0 communities

U. Bojārs *, J.G. Breslin, A. Finn, S. Decker

Digital Enterprise Research Institute, National University of Ireland, Galway, Ireland

Received 20 June 2007; received in revised form 17 September 2007; accepted 6 November 2007
Available online 4 December 2007

Abstract

Large volumes of content (bookmarks, reviews, videos, etc.) are currently being created on the "Social Web", i.e. on Web 2.0 community sites, and this content is being annotated and commented upon. The ability to view an individual's entire contribution to the Social Web would be an interesting and valuable service, particularly important as social networks are often being formed through created content and things that people have in common ("object-centred sociality"). SIOC is a Semantic Web research project that aims to describe online communities on the Social Web. This paper describes how SIOC and the Semantic Web can enable linking and reuse scenarios of data from Web 2.0 community sites, and introduces a SIOC Types module to further specify the type of content items and act as a "glue" between user posts and the content items created and annotated by users.

Keywords: RDF; Semantic Web; SIOC; Social software; Web 2.0

1. Introduction

The Web is increasingly becoming a social place: there has been a shift from just *existing* on the Web to *participating* on the Web. Community applications such as collaborative wikis, blogging, photo and bookmark sharing, and online social networks have become very popular recently, both in personal/social and professional/organisational domains [1]. Most of these collaborative applications provide common features such as content creation and sharing (images, user profiles, bookmarks, articles, etc.), provisions for discussions related to the content (comments, talk pages) and user-to-user connections (circle of friends, private messaging, etc.) and networks of users are also forming through content items of common interest (in what has been termed "object-centred sociality" [2]).

Moreover, applications are going beyond just data to provide categorising and interlinking for better search and retrieval. As examples of this, there has been huge growth in taxonomy and folksonomy usage [3] on sites like the Wikipedia, del.icio.us, CiteULike and Flickr and within some application areas interconnections between people as well as content have been formed through social networks, trackbacks, blogrolls and interwiki links. However, these applications are hitting boundaries in terms of information integration. For example, many people have multiple user accounts through which they will create new or replicated content across sites, and there is little in terms of connections between these user accounts and the associated content.

Why one would choose the Semantic Web for enhancing their Web 2.0 experience? The Semantic Web offers a generic infrastructure for interchange, integration and creative reuse of structured data, which can help to cross some of the boundaries that Web 2.0 is facing. Current Web 2.0 sites offer poor query possibilities apart from searching by keywords or tags. Microformats allow embedding of structured information into web pages but lack a generic data representation (other than embedding in HTML) and are limited in representing connections between different types of objects. Adding semantics to Web 2.0 sites aims to tackle some of these issues by creating a web of linked, "mashable" data: facilitating better (i.e. more precise) querying when compared with keyword matching, providing more reuse possibilities and creating richer links between content items. Existing efforts to represent structured data on Web 2.0, on the other hand,

* Corresponding author. Tel.: +353 91 495079; fax: +353 91 495541.
E-mail addresses: uldis.bojars@deri.org (U. Bojārs), john.breslin@deri.org (J.G. Breslin), aidan.finn@deri.org (A. Finn), stefan.decker@deri.org (S. Decker).

1570-8268/$ – see front matter © 2007 Elsevier B.V. All rights reserved.
doi:10.1016/j.websem.2007.11.010

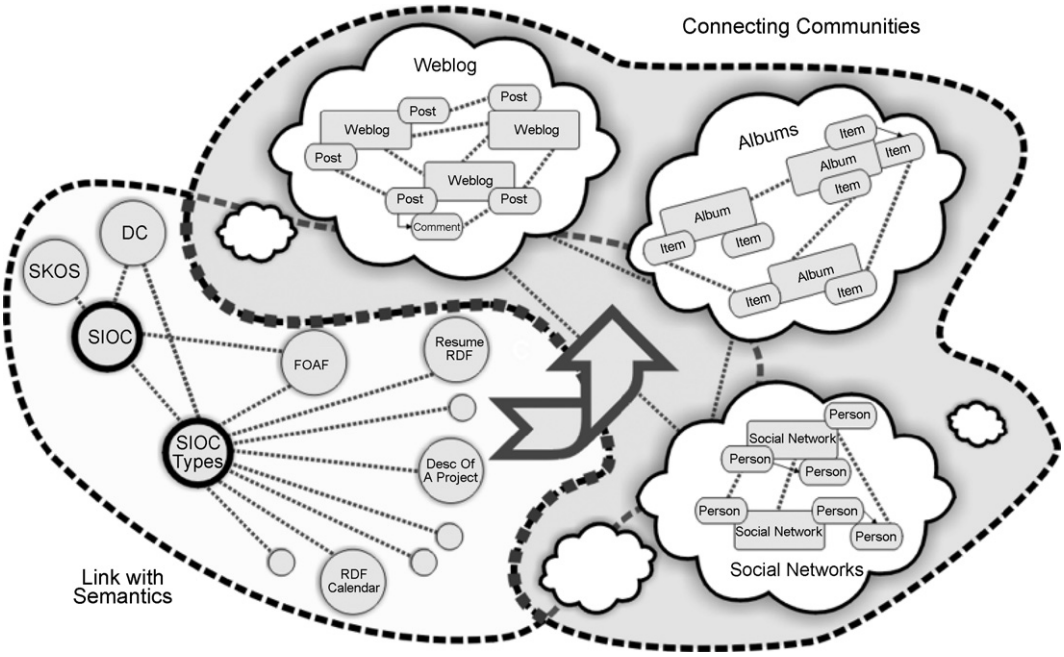

Fig. 1. Connecting communities through linked semantic data.

offer a large amount of data that we can use. By exploiting each other's achievements the Semantic Web and Web 2.0 together have a better opportunity to realise the full potential of the web [4].

In this paper, we will describe how a combination of the SIOC (Semantically Interlinked Online Communities) ontology [5] and other projects that aim to add semantic information to the current Web can be used to bring various social applications together and take them beyond some of their current limitations towards the vision of a "Social Semantic Web" (see Fig. 1). Through the use of Semantic Web data, searchable and interpretable content is retrieved from existing Web 2.0 collaborative infrastructure and intelligent use of this content can then be made. The SIOC Types module[1] introduced in this paper extends core SIOC classes with additional types needed for describing different Web 2.0 objects and aims to facilitate locating appropriate RDF vocabularies and classes suitable to describe these objects.

We will begin with some background summaries of related projects, describe the visions of this work, and detail our implementations to date, and will finish with conclusions and future work.

2. Background

The motivation for this work is to combine SIOC with other initiatives to expand the potential of the Social Web. We will now describe some of these initiatives and detail how we can augment them to create a linked Web of Data that is often locked within various social spaces.

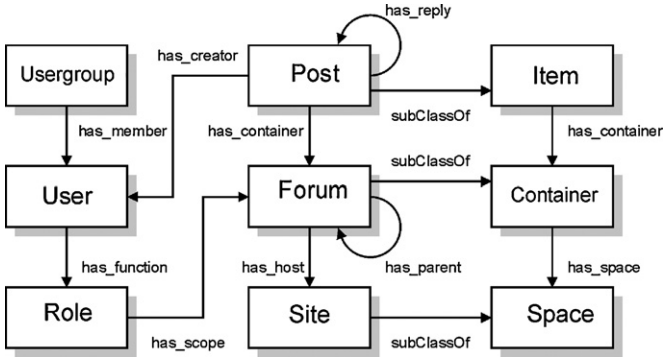

Fig. 2. Main classes and properties in SIOC.

2.1. SIOC core ontology

The SIOC core ontology[2] defines the main concepts and properties required to describe information from online communities on the Semantic Web. Through this ontology and the initial set of applications that make use of its terms, SIOC aims to meet the needs of communities and users on the evolving Web, as community-centric content sites become more prevalent and finding relevant items from these communities is now more important than ever.

The main terms in the SIOC core ontology are shown in Fig. 2. In brief, Users create content Items (e.g., Posts) that reside in Containers (e.g., Posts in Forums) on data Spaces (e.g. Sites). Initially, the SIOC Project was created to describe the realm of Web-based discussions, occurring on message boards, blogs and web archives of mailing lists. However, it soon became obvious that SIOC can span various applications for online communities,

[1] http://rdfs.org/sioc/types

[2] http://www.w3.org/Submission/2007/02/

and can be tailored towards very specific domains. The SIOC Types module was created to extend it, and various subtypes of core classes were created to describe the various types of content items that people are creating, annotating and talking about on Web 2.0 platforms.

One of the problems with combining social media data is in knowing what accounts the user holds on different social media sites so that one can access information about the content created by the user on each of these sites. A combination of the FOAF (Friend of a Friend) vocabulary and SIOC can be used to describe content created by a person across several different sites by including a list of her social media site accounts in personal FOAF profiles and using SIOC to express user-created content on these sites.

Existing SIOC exporter tools can be used to export RDF information about the content and structure of Web 2.0 platforms (blogs, wikis, forums, etc.) and are available for several common content creation platforms.[3] An important property of these SIOC exporters is that information from every page of a site is represented in RDF making all the main information contained within a site is available in a machine readable form and ready for reuse.

2.2. Microformats

Microformats[4] allow specific pieces of structured information to be embedded within HTML markup that makes up web page. This information can then be reused by various applications. Microformats have been successful in bringing semantic metadata to the current Web through a vibrant developer community. Through this community, several microformats have been created and are currently in use. The hCard microformat enables to describe information about a person such as name and contact details; the hReview microformat describes information about reviews and the hAtom microformat allows to describe information about content items available for syndication, such as blog posts and comments.

There are some limitations with microformats, especially in representing relationships between individual fragments of data, which limits the ability to properly describe the linked, Web nature of data (e.g., hAtom is sometimes used to represent blog comments, but does not have a property to indicate what blog post the comment is in a reply to). Parsing of microformats can also be a difficult task where a significant number of exceptions and special cases have to be taken into account. References to objects (such as people, content items, etc.) can often be ambiguous.

A generic approach for storing the information contained within microformats is needed if we are to store and query information about all different kinds of Web 2.0 objects in a uniform way. One option would be to store microformats in their native HTML format, but these would be difficult to process and query. Alternatively, domain specific data stores and applications could be used for each particular kind of microformat object, but they may lack flexibility and limit the ability to query over links between different object types. The third option is to use RDF, which has advantages over the two as it is more generic and allows to store and process information about all types of resource and relations between them.

2.3. APIs

Social media sites such as Flickr, Twitter or Facebook have started to open APIs that can be used by other applications to interact in new ways with the site and its data. Such APIs often provide richer data models than is possible via metadata embedded into web pages and can be a good basis for building data exporters for the Semantic Web.

However, traditional APIs have a number of shortcomings [6]. Some of these limitations include: (1) they do not work with clients that have not been designed with the specific API in mind; (2) their content cannot be accessed by search engines and other generic web agents and (3) each mashup only allows access to data from a limited number of sources chosen by the developer. In contrast, information on the Semantic Web can be used by generic clients, including RDF browsers, RDF search engines, and web query agents. Therefore, applications that can lift the data from such APIs to the Semantic Web may become useful.

2.4. Structured and semantic blogging

There have been some approaches for adding more information to blog posts, so that this information can be reused in other applications. The structured blogging effort[5] has created tools to provide microformat data from blogging platforms such as WordPress and Moveable Type. In structured blogging, structured data about people, reviews, events and other objects are becoming a part of blog posts. Sometimes a person will need for more structure in their posts (e.g. when doing a review) and may best be served by filling in an appropriate form during the post creation process. An advantage of microformats and structured blogging is that they can serve as an introduction to semantics for non-technical users: users simply choose their post type and some semantic content is generated in the background. A little bit of structure added by the user allows us to generate a lot more semantics.

The semantic blogging [7] aims to describe semantic information about individual content items within blog posts (internal semantic) using RDF. It is similar to structured blogging, but is more flexible as a result of using RDF as a data model. Some semantic blogging applications, allow a user to "drag and drop" items from the desktop and automatically generate their descriptions in RDF. This allows to describe precise semantics of data items in blog posts, but needs to reach a larger user base before it becomes a considerable source of data.

[3] http://rdfs.org/sioc/applications/
[4] http://microformats.org/
[5] http://structuredblogging.org/

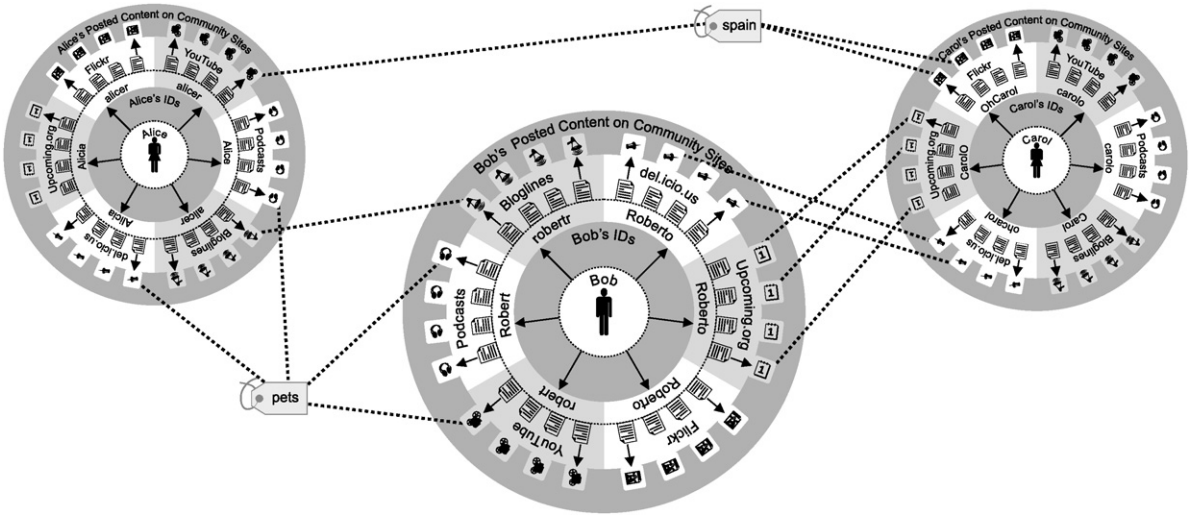

Fig. 3. Creating social networks via object-centred sociality.

3. Vision

Our work combines the benefits of different approaches described in Section 2 – we propose to use existing information on Web 2.0 and convert it to RDF which can be used as a flexible model for describing and integrating data. SIOC vocabulary is useful to describe user-created content and acts as a core to which additional structured information (e.g. data about items described in the blog post) can be added to. SIOC Types module facilitates locating appropriate RDF vocabularies and classes suitable to describe these items. Using these tools, we can describe what a post is about (sioc:about), what type of post we have created (e.g. an idea, a review, etc.) and any attachments (sioc:attachment) or other parts (dcterms:hasPart) that may be contained in it.

3.1. Consolidating user-created content

As mentioned in the introduction, an interconnection of Web 2.0 content by using Semantic Web technologies such as SIOC can lead to many interesting possibilities on the individual and community level. We will now describe one of them.

Imagine an example where a user (Bob) has created content on Flickr, YouTube, etc., through his various user identities on those sites. We could also say that each Web 2.0 content item is a user-contributed post, with some attached or embedded content. This can be modeled as a content "circle" where a person (described using FOAF) is at the inner layer of a content "circle", the next layer is formed of its user accounts (sioc:User) and the outer layer is the content – text, files, associated metadata – created by them on community sites (described using SIOC, its Types module and other relevant vocabularies).

Object-centred sociality [2] illustrates one usage of the SIOC Types module in relation to Web 2.0 sites. This idea is conceptually illustrated in Fig. 3 where the model of content "circles" is extended by showing a person being linked across communities to other people (via their user profiles) connected together by the content they create together, co-annotate or for which they use similar annotations.

For example, Bob and Carol are connected via bookmarked URLs that they both have annotated and also through events that they are both attending, and Alice and Bob are using similar tags and are subscribed to the same blogs. SIOC and FOAF can be used together to describe the objects in this social network of users. The SIOC Types module, described in Section 5, acts as a glue by pointing to external vocabularies to use for each particular type of content. All this information, integrated together, allows us to build a picture of all the objects that a user has interacted with, discussed and commented upon across different social network sites, from which the links between the users themselves emerge.

4. Describing data from Web 2.0 sites with RDF

The Resource Description Framework (RDF) allows semantic information to be expressed as a graph consisting of resources (objects) and properties used to describe objects' attributes and relationships between them. RDF is a universal model that all information, including real-world objects and their representation on Web 2.0 sites, can be expressed in. It is designed to facilitate integration of data from different sources, expressed in a number of vocabularies, and allows expressing links between various objects, e.g., a person and a software project she created.

The ability to create links between objects and to point to additional machine-readable information about these objects can often be very useful. For example, if a person is working on a software project and this project is described in Wikipedia, we can create a link to DBPedia – a machine-readable representation of Wikipedia data [8] – with some RDF data about this project.

We will now illustrate how information about a typical online community site content – blog post with comments – can be described in RDF, and how meaningful queries can be asked over this linked data set. The snippet below presents a blog post and its comment described in RDF (using Turtle notation):

```
_:post1 a sioc_t:BlogPost ;
    ... content; other properties ...
    sioc:has_reply _:comment1 ;
    sioc:has_creator _:user1 .
_:comment1 a sioc_t:Comment ;
    ... content; other properties ...
    foaf:maker _:person2 .
_:person2 a foaf:Person ;
    foaf:name "Aidan Finn" ;
    foaf:homepage <http://www.aidanf.net/> .
```

This is a basic example which describes the source information as a set of linked objects (a post, a comment and a person) and their properties. This information can be enhanced with additional linked data, located anywhere on the Web, e.g., we could add a rdfs:seeAlso property to _:person2 pointing to a location of this person's FOAF RDF profile, containing information about other social media site accounts this person has (e.g., Flickr), people he knows and topics he is interested in. The RDF data model allows to seamlessly join data coming from all these distributed locations on the Web.

As soon as all the information is described in RDF, we can ask queries over this heterogeneous data set. For data described using microformats, you would usually extract a particular microformat (e.g., hCard) and store it in a domain-specific application (e.g. an address book). In such a scenario, a user would be limited to using only queries on the properties of address book entries, but will not be able to tap into a much richer information contained in the relations between different types of objects. By converting Web 2.0 data into RDF (e.g. by using GRDDL[6]) we can use make a richer use of this information. Here is an example query for retrieving information about all persons who have replied to posts created by a particular user, represented in a human readable pseudocode:

```
return a distinct set of values of
  names and (optionally) homepage URLs
    of all persons
      who have commented on posts
        created by _:user1
```

This query returns information about a set of people whom a user is connected to via comments to his posts. If the user shares the same identification (e.g., a URI) across a number of community sites then relevant information from all of these sites will be returned. This and other queries over RDF data are used by the Social SIOC Explorer [9] to extract social relations and context from online community sites.

While ability to query linked data already gives us powerful tools to explore the Social Web, new information can also be created from the existing information, e.g., by using rules to create new, derived data which can also be published back to the web. Details about using rules on RDF data is outside the scope of this paper, and therefore we will just provide a simple example, relevant to our work. Object-centred sociality, introduced earlier, considers user-created objects as an indication of relations between people. If the information on who created different kinds of Web 2.0 objects is available, then we can define simple rules that will add new information about relations between people, for example:

```
(a has_bookmarked url_1) and
(b has_bookmarked url_1)
    => (a is_related_to b)
```

The SPARQL RDF query language [10] can be used both for querying RDF (using SELECT statements) and to express simple rules like the one in the example above (using CONSTRUCT statements) with additional RDF rule languages available for more complex use cases.

5. Implementation of SIOC Types module

SIOC follows a modular design where additional ontology modules can be created for specializing and further extending classes and properties contained within the SIOC Core ontology. Currently there are two modules defined: (1) SIOC Services module and (2) SIOC Types module. The Services module allows one to indicate web services that are associated with (located on) a sioc:Site or a part of it, and is not directly relevant to this paper. In this section we will concentrate on the SIOC Types module and describe it in more detail.

The SIOC Types module extends the core ontology by introducing subtypes of SIOC classes such as sioc:Container, sioc:Item, sioc:Forum and sioc:Post. This module has two roles:

(1) to define subtypes of SIOC objects needed for more precise representation of various elements of online community sites (e.g., sioc_t:Comment is a subclass of sioc:Post and sioc_t:MessageBoard is a subclass of sioc:Forum);
(2) to introduce new subtypes for describing different types of Web 2.0 objects in SIOC and pointing to existing ontologies suitable for describing details of these objects (e.g., a sioc_t:ReviewArea may contain sioc_t:Review(s) which can be described in detail using the Review Vocabulary[7]).

This second role of the SIOC Types module aims to bring together tools – different RDF vocabularies – to describe Web 2.0 objects and sites in RDF. While the vocabularies may exist for some of these types, due to the distributed nature of the Semantic Web it is not a simple task to find these vocabularies. Sometimes a single vocabulary will not cover all the information needed, and a number of vocabularies may need to be combined or new terms and/or vocabularies created. The SIOC Types module does not aim to replace these vocabularies but rather adds value by acting as a "one-stop shop"—a single location that can point to other suitable vocabularies for many common content types.

Fig. 4 lists main sub-types of sioc:Container and sioc:Forum we identified as necessary to represent collections of popular types of Web 2.0 objects. These subtypes are listed on the left side of the figure while the right side of the figure lists relevant

[6] http://www.w3.org/TR/grddl/

[7] http://dannyayers.com/xmlns/rev/

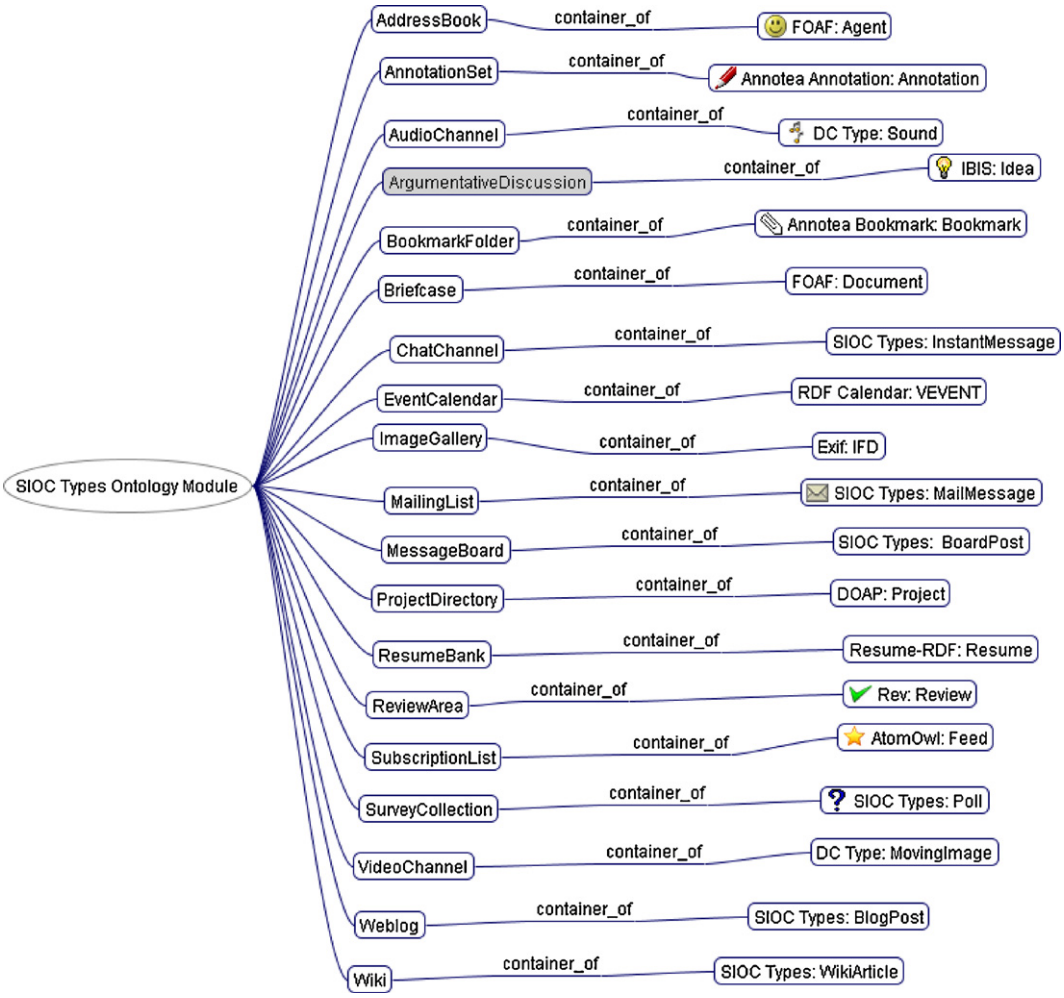

Fig. 4. Container classes in SIOC Types and related content items which they may contain.

ontologies that can be used to represent these objects. Containers and objects contained within them are linked together using sioc:has_container and sioc:container_of properties.

Currently, the initial version of the SIOC Types module uses an rdfs:seeAlso property to point SIOC Types objects to related vocabularies and classes to use for describing individual items contained within them. We chose this property as a weak link pointing to related objects and are exploring more formal ways how to link vocabularies together. One option is to subclass individual item classes in SIOC Types module from the relevant classes identified in an external ontology. The downside of this option is that one particular class has to be chosen contradicting the distributed nature of the Semantic Web where there can a number of suitable ontologies which users may wish to choose from.

After social media site data are described in RDF using SIOC, its types module and other relevant vocabularies, the advantages of producing RDF data can be reaped through semantically enabled applications for browsing, reusing and sharing. For example, WordPress SIOC Importer can import any sioc:Post item into a WordPress blog entry, and generic RDF browser applications such as Disco and Tabulator may be used for exploring the Web of Data.

6. Example—representing reviews

Structured blogging allows the creation of a group of pre-defined content types and assists a user in entering and publishing structured information about this content. The hReview microformat schema defines several fields that can be used to describe a review: summary, item type, item info, reviewer, dtreviewed, rating, description, tags, permalink and license. The fields that are defined for hReview are fixed in advance and are limited to describing data defined in the hReview schema or by one of the other microformats.

The Review vocabulary is designed to represent a review in RDF. The vocabulary properties are createdOn, hasReview, maxRating, minRating, rating, reviewer and text. The number of fields defined for the Review vocabulary is less than for the hReview microformat, but RDF makes it possible to combine a number of ontologies in a well defined way thus allowing to express all the information in hReview and some additional data. External object descriptions such as DBPedia or FOAF profiles can be linked to. E.g., a rev:Review may link to an external FOAF file on the author's website that describes the author and his or her online accounts.

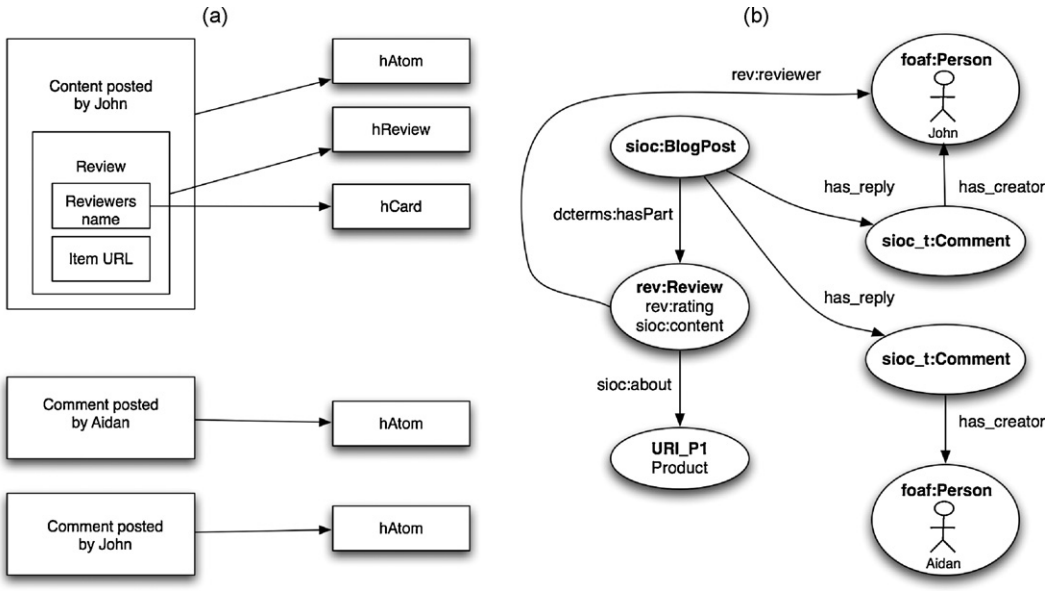

Fig. 5. Describing a review (a) in hReview and (b) as linked RDF data.

Table 1
Mapping between hReview and RDF vocabularies

hReview field	RDF field(s)
Summary	dc:title
Item type	Classes linked from SIOC Types
Item info	sioc:about
Reviewer	foaf:maker, foaf:Person, rev:reviewer
dtreviewed	dcterms:created
Rating	rev:rating
Description	sioc:content, rev:text
Tags	sioc:topic
Permalink	sioc:link, URL
Licence	cc:license

An example of a structured review post represented using hReview is shown in Fig. 5a while Fig. 5b shows the same review data represented using linked RDF data. Notice that the RDF data shown here resembles the basic example from Section 4 (a post and a comment linked to it) with a rev:Review object added.

Table 1 lists how the terms in the hReview microformat map to a combination of Review, SIOC, FOAF and other vocabularies used as needed. This shows how using RDF gives us greater flexibility as we can combine different vocabularies to describe items of interest. SIOC Types allows us to specify detailed information about the types of objects and to connect different types of RDF classes (describing real-world and Web 2.0 objects) together by linking to the relevant vocabularies to use to describe them. The roles of all the different RDF vocabularies and their namespace abbreviations used in the mapping above are as follows:

- FOAF (*foaf*) is used to describe information about a person who created the review;
- SIOC (*sioc*) is used to describe information about the blog post that a review is contained in, some of the information contained within a review and other kinds of online community site information (e.g., comments to a review post);
- SIOC Types module (*sioc_t*) is used to define different types of items that a review is about and also to specify that a container is a sioc_types:ReviewArea;
- Review RDF (*rev*) is a domain specific vocabulary used to describe the main properties of a review, and CC is a domain-specific RDF vocabulary used to describe Creative Commons licenses;
- Dublin Core (*dc* and *dcterms*) is used to describe general properties such as review title and creation date and to connect a sioc:BlogPost with a review that is a part of a post.

Information from microformats can be reused using utilities such as the Firefox Operator plugin. Operator currently only detects a single fragment of information it can use from a review—the hCard used to describe the reviewer. Even if extended to extract more, it will still be working with individual pieces (fragments) of the information available, e.g., losing an important link between a comment and a post that a comment responds to. A review expressed in RDF, on the other hand, is designed with relations between objects in mind and can be easily extended with more properties and pointers to where additional structured data are available (e.g. if URL2 is some software the we can link to a detailed DOAP description for this product).

7. Conclusions and future work

Many social media sites already express structured information through open APIs and microformats, but both have their limitations. In this paper, we have demonstrated how information about Web 2.0 can be combined with the Semantic Web technologies in a mutually beneficial way: with open APIs and microformats acting as a way to get structured information from the social media sites, and Semantic Web technologies offering a generic infrastructure for interchange, integration and creative reuse of structured data.

In this paper, we described the use of SIOC, FOAF and other vocabularies for describing social media site information as linked RDF data. We introduced the SIOC Types ontology module which can act as a glue that brings together various vocabularies needed for describing information about different types of objects in RDF. Main challenges to be addressed in future work are in exploring better techniques for describing combinations of RDF vocabularies to use to describe Web 2.0 objects and in defining concrete mappings from these objects into RDF.

We hope that this work will help in bridging the efforts of Semantic Web and Web 2.0 communities and help us all to achieve more that can be done by each of these efforts individually.

Acknowledgements

This work was supported by Science Foundation Ireland under Grant No. SFI/02/CE1/I131. We gratefully acknowledge Conor Hayes for his valuable feedback and all members of the SIOC developer community for their contribution in adding semantics to online community sites.

References

[1] J. Kolbitsch, H. Maurer, The transformation of the web: how emerging communities shape the information we consume, J. Universal Comput. Sci. 12 (2) (2006).

[2] Y. Engeström, Collaborative Intentionality Capital: Object-Oriented Interagency in Multiogranizational Fields, University of California, San Diego, 2004.

[3] A. Hotho, R. Jaschke, C. Schmitz, G. Stumme, Information retrieval in folksonomies: search and ranking, in: Proceedings of the 3rd European Semantic Web Conference, Budva, Montenegro, 2006.

[4] A. Ankolekar, M. Krötzsch, T. Tran, D. Vrandecic, The two cultures: mashing up Web 2.0 and the Semantic Web, in: Proceedings of the 16th International Conference on World Wide Web, Banff, Alberta, Canada, May 2007, pp. 825–834.

[5] J.G. Breslin, A. Harth, U. Bojārs, S. Decker, Towards semantically interlinked online communities, in: The 2nd European Semantic Web Conference Proceedings (ESWC '05), Heraklion, Greece, May 2005.

[6] C. Bizer, R. Cyganiak, T. Gauß, The RDF Book Mashup: from Web APIs to a web of data, in: The 3rd Workshop on Scripting for the Semantic Web (SFSW 2007), Innsbruck, Austria, 2007.

[7] K. Möller, U. Bojārs, J.G. Breslin, Using semantics to enhance the blogging experience, in: Proceedings of the 3rd European Semantic Web Conference (ESWC '06), LNCS, vol. 4011, Budva, Montenegro, June 2006, pp. 679–696.

[8] S. Auer, C. Bizer, G. Kobilarov, J. Lehmann, R. Cyganiak, Z. Ives, DBpedia: a nucleus for a web of open data, in: The 6th International Semantic Web Conference (ISWC 2007), Busan, Korea, November 2007.

[9] U. Bojārs, B. Heitmann, E. Oren, A prototype to explore content and context on social community sites, in: SABRE Conference on Social Semantic Web (CSSW 2007), Leipzig, Germany, September 26–28, 2007.

[10] E. Prud'hommeaux, A. Seaborne, SPARQL Query Language for RDF, W3C Candidate Recommendation, 14 June 2007, http://www.w3.org/TR/rdf-sparql-query/.