Weaving SIOC into the Web of Linked Data

Uldis Bojārs
Digital Enterprise Research Institute
National University of Ireland, Galway
uldis.bojars@deri.org

Alexandre Passant
Electricité de France, R&D
Clamart, France
& LaLIC, Université Paris-Sorbonne
Paris, France
alexandre.passant@edf.fr

Richard Cyganiak
Digital Enterprise Research Institute
National University of Ireland, Galway
richard@cyganiak.de

John Breslin
Digital Enterprise Research Institute
National University of Ireland, Galway
john.breslin@deri.org

ABSTRACT

Social media sites can act as a rich source of large amounts of data by letting anyone easily create content on the Web. The SIOC ontology and tools developed for it allow us to express this information as interlinked RDF data. This paper describes approaches for making this Social Web information an integral part of the Web of Linked Data.

1. INTRODUCTION

By envisioning a Web of Data, in addition to the Web of Documents, the Semantic Web provides a way to represent real-world data, as well as virtual objects in a uniform manner. The Social Web can also benefit from a common way to link and represent content created through online communities and social media websites.

The Web currently uses untyped HTML hyperlinks to connect web pages and mainly considers documents without any semantics. RDF allows us to describe information about resources, including real-world objects, and the different types of relations between them. The Linked Data initiative takes this a step further and defines best practices for publishing linked data on the web [2]: use URIs as names for things; use HTTP URIs so that people can look up those names; when someone looks up a URI, provide useful information; include links to other URIs so that people can discover more related things.

SIOC (Semantically Interlinked Online Communities) [7] is a step towards providing linked data from social media sites (forums, blogs, wikis, etc.) and provides a common vocabulary to describe meta-data of such sites in RDF. The project itself is built in a collaborative way, since it involves many partners and welcomes any suggestions on its SIOC-DEV mailing list workgroup[1], and its work has been recently published as a W3C member submission[2].

This paper describes how Social Web data, described in SIOC, can be further weaved together with other kinds of linked data. We describe the structure of SIOC data and give a description of different approaches to interlink SIOC data with other sources of linked data. Finally, we conclude the paper with some interesting uses of such data.

2. SIOC DATA

The SIOC implementations list[3] contains around 35 applications for creating and using SIOC data. By installing relevant SIOC export plugins, online community sites can generate linked data and start forming a critical mass of RDF data about user-created content in the same way as LiveJournal did for FOAF data. Other tools allow users to browse SIOC data or to translate existing data, such as mailing list archives, to SIOC. Moreover, SIOC is also used for enterprise data integration[9], and some popular Web 2.0 sites such as Seesmic[4] start using it to model their data.

To demonstrate the linked nature of SIOC data created by these sites, let us look at data generated by a SIOC export plugin for a blog site (e.g. the WordPress SIOC plugin[5]). It creates a set of RDF documents describing the blog itself and every post, comment and user on this blog. Note that data instances exported from a larger site (e.g., a bulletin board) are similar, but larger in number and may contain multiple `sioc:Forum` objects whereas a blog has just one.

An important property of SIOC data is that all the RDF documents generated by an exporter are interlinked using `rdfs:seeAlso` links (Fig. 1):

[1] `http://groups.google.com/group/sioc-dev`
[2] `http://www.w3.org/Submission/2007/02/`
[3] `http://rdfs.org/sioc/applications/`
[4] `http://seesmic.com`
[5] `http://sioc-project.org/wordpress`

LDOW2008, April 22, 2008, Beijing, China.

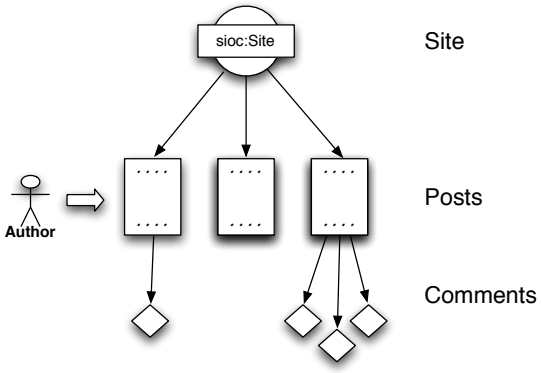

Figure 1: Structure of linked SIOC data

- a blog profile links to all posts on the blog;
- posts link to replies / comments;
- posts and comments link to the profile of a user who created them (if a user is registered on a site).

This makes SIOC-enabled social media sites a rich source of interlinked data and enables an RDF crawler to start at the main profile of a site and retrieve all SIOC RDF pages that this site provides from a single entry point, i.e. the RDF file describing the `sioc:Site`. Almost all SIOC-enabled sites also have RDF autodiscovery links in their web pages, making it easier to find RDF data for people and applications when opening a page. Another interesting use of this autodiscovery feature is implemented in the Semantic Radar Firefox extension[6]. The extension automatically sends a notification to the Ping The Semantic Web service[7] whenever the user accesses an autodiscovery-enabled web page. This allows people to participate in the distributed discovery of Semantic Web content [6].

While these features makes a SIOC-enabled site a good source of linked data, at this point it can still be viewed as a "walled garden" and not really as an integral part of a larger Web of Linked Data. On its own, SIOC data from a single exporter may have a limited connectivity to the "outside world". Links to other websites are mainly achieved at the moment thanks to the extraction of HTML hyperlinks from post content. These hyperlinks are republished in the post's SIOC profile using the `sioc:links_to` property.

3. LINKING DATA WITH SIOC

There are two paths to making online community data, described in SIOC, a more integrated part of the Web of Data: linking into "SIOC space", and referencing linked data from inside online community sites. We will now describe both of these paths, starting with linking to SIOC data.

3.1 Linking to SIOC Data

The question of receiving links from the "outside world", strictly speaking, is not something SIOC exporters can directly influence as this depends on other sources of data

[6] `http://sioc-project.org/firefox`
[7] `http://pingthesemanticweb.com`

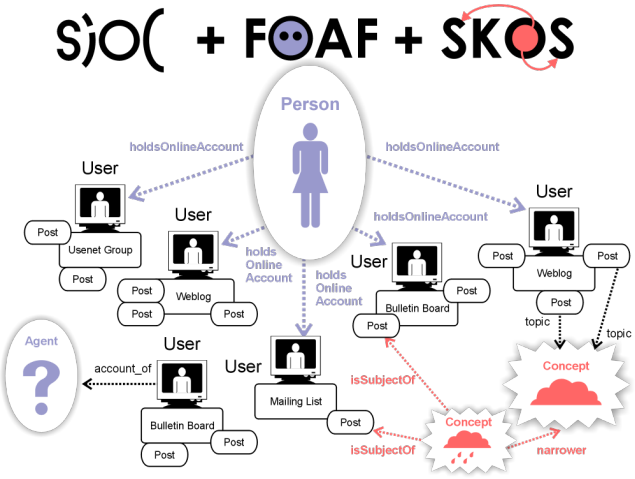

Figure 2: Interlinking SIOC, FOAF and SKOS

linking to SIOC Sites(s). Yet, there are some things that can be done to facilitate linking to SIOC:

- owners of FOAF profiles can link to social media sites and their user accounts on these sites;
- SIOC exporters can be optimized to make SIOC data easier to discover;
- Semantic Web indexing and lookup services can find and provide access to SIOC data.

A simple and effective way to link to SIOC profiles is to point to them from personal FOAF profiles. For example, people often already link to their blogs from FOAF using the `foaf:weblog` property. In this case they just need to add an `rdfs:seeAlso` link pointing to the weblog's SIOC profile, which also mentions the weblog's homepage URI and thus allows a client to connect the information in the two documents.

People can also use the `foaf:holdsOnlineAccount` property to point to user accounts (`sioc:User`) that they have registered on online community sites (Fig. 2). As a part of SIOC data, these user accounts are normally linked to content items created by the user.

If we can make it easier for applications and users browsing the Web to find SIOC data, they are more likely to use this data. RDF autodiscovery links play an important role in helping one to find Semantic Web data, and almost all SIOC exporters use them, making it a reliable way to detect the presence of SIOC data. As a result, all web pages on a SIOC-enabled site contain RDF autodiscovery links to relevant RDF data (e.g., a blog post will point to data about this particular post).

Another way of improving discoverability is content negotiation. The idea is that client requests to URIs of resources such as forum posts are answered with either an HTML or an RDF representation, depending on the client's preferences. This lets existing URIs of the forum's web pages,

which are already used as targets of HTML links, double as access points to the SIOC data when used by RDF-aware clients. We have experimented with adding content negotiation to some SIOC exporters, but this has proven to be challenging. First, implementing full content negotiation as described in the HTTP specification is fairly complex. Second, it is hard to verify that an implementation works correctly with the large number of different versions of Web and data browsers in use on the Web, some of which express questionable preferences. Third, the plugin systems of some content management systems do not provide the necessary hooks for implementing content negotiation. Therefore, we recommend that linked data browsers such as Tabulator [3] also make use of RDF autodiscovery information as an alternative to content negotiation.

SIOC exporters can also use RDFa[8] to embed SIOC data directly in HTML documents, e.g. blog posts. A first approach to embed RDFa in Drupal has been explored[9], and future work may go in this direction, since it provides another way to discover RDF triples related to a document.

Finally, some site administrators have decorated their pages with small SIOC icons that link directly to the related SIOC profile. These icons can be considered redundant as we already have RDF autodiscovery links, or can even be considered harmful as users who click the icon in non-RDF-aware Web browsers will be subjected to raw RDF/XML code. On the other hand, such icons are a useful marketing tool that increases awareness of SIOC in general.

In general, it can be said that we desire a user experience where URIs of Web resources such as blog posts can simply be used in HTML hyperlinks *and* RDF statements to refer to the resource. Mechanisms such as RDF autodiscovery, content negotiation or RDFa parsing should be transparent to the user and entirely reliable.

An important aspect of linked data is being able to find what data is linking into a resource. Online community sites sometimes notify each other of links they make by using pingbacks or trackbacks, but other publishers usually do not send such notifications. This is where Semantic Web indexing services such as Sindice [12] play an important role. Such a service can find the incoming RDF statements that reference a resource. These "backlinks" can be displayed to users, either on SIOC-enabled sites or directly in the user interface of linked data browsers, thus helping users to navigate these links in both directions.

3.2 Linking From SIOC Data to the Outside

Some resources that users may want to link to include:

- Persons - authors of posts and comments => link to their FOAF URI and FOAF profile;
- Topics - categories and tags => link to other linked data about these topics;
- Other data - information associated with / embedded within content.

FOAF is one of the most successful Semantic Web vocabularies and is often used to describe personal information and social network relations. One of the first use cases for linking to other RDF data is to link posts or comments to FOAF profiles of their creators. Currently `sioc:Post(s)` are linked to a creator's user profile on a community site, but export tools can be easily extended to point to a provided FOAF profile and to the URI of a creator of this content.

In order to point to FOAF profiles, the application has to know what URI to point to. Users of a community site (e.g., post authors) can supply this information when registering on the site, but comment authors may not be motivated enough to provide the necessary FOAF information and an automated process for finding this data is preferable. This can be achieved automatically using OpenID[10]: A user authenticates using OpenID and a blog engine checks an OpenID URI for links to person's FOAF profile. If such a link is found, statements pointing to the person's FOAF profile (e.g., `owl:sameAs` and `rdfs:seeAlso`) can be added to SIOC metadata describing the author of the comment. Some work regarding this point are currently done within the SparqlPress[11] project, which aims to provide Semantic Web functionalities, in both exporting, crawling and linking RDF data.

Categories and tags which describe the topic of the content are also good candidates for pointing to additional information. SIOC exporters use a `sioc:topic` property to point to a topic URI. Topic category hierarchies can be described in SKOS and topics can be linked to more information about them such as RDF data from DBpedia [1]. Tag vocabularies such as the Tag Ontology, MOAT (Meaning Of A Tag)[12], or SCOT [8] can be used to associate more information with tags. However, SIOC exporters need to know what URIs to point to for a given tag or category.

One option is to add this knowledge to the site in advance. Adding annotations every time a new topic is created may work for categories that do not change very often, but is not feasible for tags which are more dynamic. Another option is to rely on the MOAT framework, which allows us to assign meaning to tags in a collaborative way using existing URIs from datasets such as DBpedia or GeoNames [11]. For example, some users in a community may have indicated that the `sparql` tag means `http://dbpedia.org/resource/SPARQL`. When a user tags a post with the `sparql` tag, the DBpedia URI will be offered as a possible meaning of the tag, and can be selected with a single click for exporting as a `sioc:topic` link. This provides an efficient way to interlink community site posts and DBpedia. Thanks to semantic relationships within DBpedia itself, useful information about how the topic of different posts relate is only a link away (Fig. 3).

[8] `http://www.w3.org/TR/xhtml-rdfa-primer/`
[9] `http://groups.google.com/group/sioc-dev/browse_thread/thread/b1585e9ef3a17665`
[10] `http://apassant.net/blog/2007/09/23/retrieving-foaf-profile-from-openid/`
[11] `http://wiki.foaf-project.org/SparqlPress`
[12] `http://moat-project.org/`

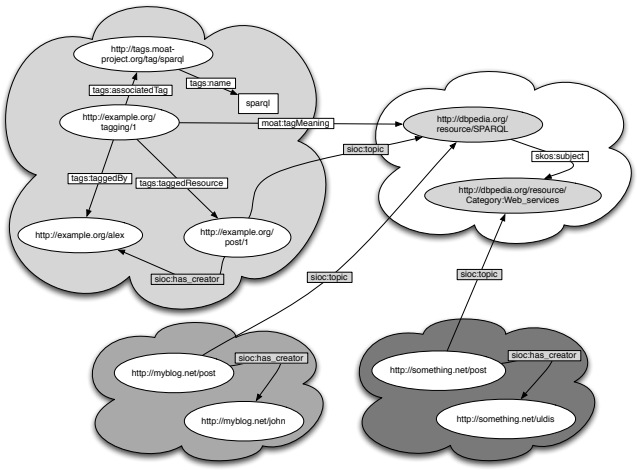

Figure 3: Interlinking blog posts thanks to SIOC, MOAT and DBpedia.

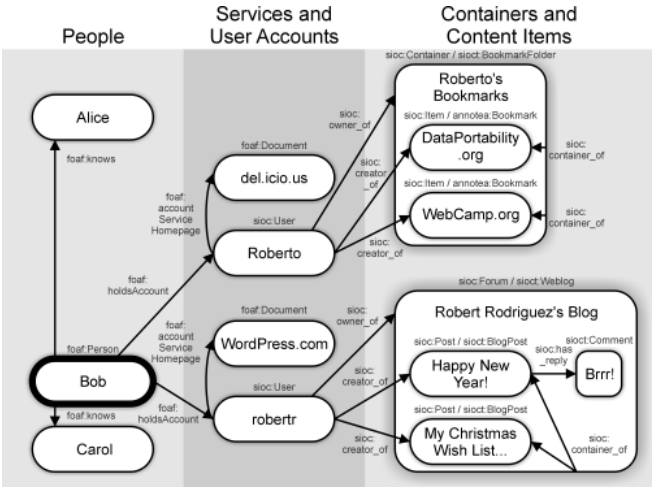

Figure 4: FOAF, SIOC and Data Portability.

There are other resources that authors may want to describe or link to. Some SIOC exporters have built-in functionality for defining links to additional RDF data. The Drupal and WordPress exporters pass post content through a filter to find hyperlinks with MIME type `application/rdf+xml`. Appropriate `rdfs:seeAlso` statements are added to the generated SIOC RDF data for this post in addition to the usual `sioc:links_to` property. This simple mechanism allows us to add links to RDF data to existing blog posts with very little effort. As a possible extension of this mechanism, some information about the linked resources may be copied into the post's SIOC profile, e.g. `rdfs:label` and `rdf:type` information which assists users of data browsers in navigation.

Sometimes, having a link to external RDF data is not enough. Within a post, a user may want to express some additional machine-readable information about the objects discussed in the post. As example use case an author describes a software project, points to DOAP (Description Of A Project) data describing this project, and includes a review of this project. This information can be added to SIOC data, but we need a way to add additional information to content items and to be able to retrieve it later.

Such embedded annotations are not widespread yet and current SIOC tools do not implement this functionality, but they can be extended to support such a use case. Filters for extracting metadata from post content can be executed one after another and RDF data can be extracted and added to the generated SIOC data. Data can be embedded in content in a number of different ways – RDFa, GRDDL[13], microformats – as long as appropriate content extraction modules are available.

4. USING SOCIAL MEDIA DATA

The Social Web is not limited to forums or blogs. There are different kinds of social media and Web 2.0 sites, such as Flickr, Twitter and Facebook, which offer interesting content that can be described in RDF using FOAF, SIOC and related vocabularies [4]. Developers have already created exporters for such tools and some of them use SIOC to describe user account and content, as in exporters for Twitter[14] and Flickr[10].

Information from users' contribution to these sites can be interesting at different levels. There are social networks, which are formed by users on different sites. These networks can be described in FOAF. Web users may also have a FOAF profile which points to their different accounts on different sites. Users create content items (`sioc:Item`) such as videos, bookmarks, etc. which can be organised in containers (Fig. 4). When expressed in RDF, this information forms an interlinked web of rich social data, ready for reuse. Moreover, the use of the SIOC types module[15] allow people to decribe exactly the type of their content (eg: `sioct:BlogPost`, `sioct:VideoChannel`).

Two interesting and emerging applications of such information are object-centred sociality and social media portability. Object-centred sociality looks at objects such as content items which people create and co-annotate as a medium through which people are connected together. The use of SIOC data for object-centred sociality is explored in [5]. Data or social media portability[16] is an initiative aimed at providing open standards for discovery, import, export and synchronisation of user profiles, relationships, content and media. SIOC and FOAF, combined with domain specific ontologies, allows to describe most of such information and can form a solution for social media portability in an open and machine-redeable way (Fig. 4).

5. CONCLUSION

SIOC data created by online community sites are highly interlinked and ready for weaving into a larger Web of Data. In this paper we described some approaches for facilitating the linking to SIOC data and for using SIOC to link back to other RDF data.

[13] `http://www.w3.org/TR/grddl-primer/`

[14] `http://sioc-project.org/node/262`

[15] `http://rdfs.org/sioc/types`

[16] `http://www.dataportability.org/`

While some may consider social media content just an "endpoint" in a journey for exploration of linked data, interesting possiblities arise when these content items are both linked to and contain links to other RDF data. In this case social media content and associated SIOC data act as a linking point by connecting together different parts of the linked data universe.

The authors are looking forward to feedback and suggestions from other implementers of Linked Data on the Web for enabling interoperability and reuse of the SIOC data and tools described here.

6. REFERENCES

[1] S. Auer, C. Bizer, G. Kobilarov, J. Lehmann, R. Cyganiak, and Z. G. Ives. DBpedia: A Nucleus for a Web of Open Data. In *ISWC/ASWC*, volume 4825 of *Lecture Notes in Computer Science*, pages 722–735. Springer, 2007.

[2] T. Berners-Lee. Design Issues–Linked Data. Published online, May 2007. `http://www.w3.org/DesignIssues/LinkedData.html`.

[3] T. Berners-Lee, Y. Chen, L. Chilton, D. Connolly, R. Dhanaraj, J. Hollenbach, A. Lerer, and D. Sheets. Tabulator: Exploring and Analyzing linked data on the Semantic Web. In *Proceedings of the The 3rd International Semantic Web User Interaction Workshop (SWUI06)*, Nov 2006.

[4] U. Bojārs, J. Breslin, A. Finn, and S. Decker. Using the Semantic Web for Linking and Reusing Data Across Web 2.0 Communities. *The Journal of Web Semantics, Special Issue on the Semantic Web and Web 2.0 (Forthcoming)*, 2008.

[5] U. Bojars, B. Heitmann, and E. Oren. A Prototype to Explore Content and Context on Social Community Sites. *SABRE Conference on Social Semantic Web (CSSW 2007)*, 2007.

[6] U. Bojārs, A. Passant, F. Giasson, and J. G. Breslin. An Architecture to Discover and Query Decentralized RDF Data. *Proceedings of the 3rd Workshop on Scripting for the Semantic Web (SFSW 2007), Innsbruck, Austria*, Jun 2007.

[7] J. G. Breslin, A. Harth, U. Bojars, and S. Decker. Towards Semantically-Interlinked Online Communities. In *The 2nd European Semantic Web Conference (ESWC '05), Heraklion, Greece, Proceedings*, May 2005.

[8] H. L. Kim, J. G. Breslin, S. K. Yang, and H. G. Kim. Social Semantic Cloud of Tag: Semantic Model for Social Tagging. *Proceedings of the 2nd KES International Symposium on Agent and Multi-Agent Systems: Technologies and Applications (Forthcoming), Incheon, Korea*, 2008.

[9] A. Passant. A Collaborative Semantic Space for Enterprise. *Proceedings of the Knowledge Web PhD Symposium 2007 (KWEPSY 2007), Innsbruck, Austria*, Jun 2007.

[10] A. Passant. :me owl:sameAs flickr:33669349@N00. *Proceedings of the WWW 2008 Workshop Linked Data on the Web (LDOW2008), Beijing, China*, Apr 2008. Demo presentation.

[11] A. Passant and P. Laublet. Meaning Of A Tag: A collaborative approach to bridge the gap between tagging and Linked Data. *Proceedings of the WWW 2008 Workshop Linked Data on the Web (LDOW2008), Beijing, China*, Apr 2008.

[12] G. Tummarello, R. Delbru, and E. Oren. Sindice.com: Weaving the Open Linked Data. *Proceedings of the International Semantic Web Conference (ISWC 2007)*, 2007.