Towards semantically-interlinked online communities

Towards Semantically-Interlinked Online Communities

Uldis Bojars, John G. Breslin, Andreas Harth, and Stefan Decker

Online community sites have replaced the traditional means of keeping a community informed via libraries and publishing. At present, online communities are islands that are not interlinked. Ontologies and Semantic Web technologies offer an upgrade path to providing more complex services. We present the SIOC (Semantically-Interlinked Online Communities) ontology which combines terms from vocabularies that already exist with new terms needed to describe the relationships between concepts in the realm of online community sites.

Keywords: Knowledge Management, RDF, Online communities, Ontologies, Semantic Web, Weblogs.

1 Introduction

At the moment, most online communities are islands that are not linked. Sites are hosted on stand-alone systems that cannot be interconnected due to application and interface differences. Parallel discussions on interrelated topics may exist on a number of sites, but their users are unaware of that. There is a huge amount of related information that could be harnessed across online communities, from similar member profile details to common-topic discussion fora.

The goal of SIOC (Semantically-Interlinked Online Communities, <http://rdfs.org/sioc>) is to interconnect these online communities. Community sites can include many discussion primitives, such as bulletin boards, weblogs and mailing lists, which we have grouped under the concept of forum.

SIOC will facilitate the location of related and relevant information; by searching on one forum, the ontology and interface will allow users to find information on fora from other sites that use a SIOC-based system architecture. Other uses include cross-site querying, topic-related searches, and the importing of SIOC data into other systems. Therefore, SIOC tries to overcome the serious limitations of current sites in making information accessible to their users in an efficient manner [6].

In a typical usage scenario, a user is searching for information on, for example, installing broadband on a Linux-based PC in their house in Galway. There is a post A discussing local ISPs (Internet Service Provider) on site 1, a bulletin board dedicated to Galway, that references (on the HTML, HyperText Markup Language, level) both a Usenet post B comparing broadband modems and a mailing list post C detailing how to install broadband on Linux. Previously the user would have had to traverse three sites to find the relevant information. However, by making use of the SIOC ontology and remote RDF (Resource Description Framework) querying, a search for broadband on the Galway bulletin board will also yield the relevant text from the interlinked Usenet and mailing list posts B and C.

There are some challenges for SIOC. The grand challenge is adoption by community sites, i.e. how can the users be enticed to make use of the SIOC ontology. By using concepts that can be easily understood by site administrators, and by providing properties that are automatically created by an end-user, the SIOC ontology can be adopted in a useful way. A second challenge is how best to use SIOC with existing ontologies. This can be partially solved by mappings and interfaces to commonly-used ontologies. Another challenge is how SIOC will scale. We will keep the scaling challenge in mind when creating a future architecture for an interconnected system of community sites.

The main contributions of this paper are the development of the SIOC ontology and mappings to other RDF vocabularies, and prototypes to produce SIOC metadata from community sites.

The remainder of this paper is organised as follows. In Section 2, we describe the SIOC ontology and mappings to other existing vocabularies. In Section 3, we discuss the exchange of SIOC instances. Section 4 describes some usages of the created instances, and related work is discussed in Section 5. Section 6 concludes the paper.

Uldis Bojars is currently studying for his PhD at DERI (Digital Enterprise Research Institute), National University of Ireland, Galway (NUI Galway). His research interests include semantic matching of skills, social networks and online community discussions. <Uldis.Bojars@deri.org>

John G. Breslin received his PhD at the National University of Ireland, Galway. He is a Postdoctoral Researcher at DERI (Digital Enterprise Research Institute), NUI Galway, Ireland. His research interests include social networks and online communities. <John.Breslin@deri.org>

Andreas Harth is currently studying for his PhD at DERI (Digital Enterprise Research Institute), NUI Galway, Ireland. His research interest is data interoperation on the Web. <Andreas.Harth@deri.org>

Stefan Decker received his PhD at the University of Karlsruhe, Germany. He is an Executive Director and Adjunct Lecturer at DERI (Digital Enterprise Research Institute), NUI Galway, Ireland. His research interests include the Semantic Web and P2P technologies. <Stefan.Decker@deri.org>

2 SIOC Ontology

In this section we present the SIOC ontology. The ontology consists of two major parts: classes and properties that describe the information in online community sites, and mappings that relate SIOC to existing vocabularies.

We have identified the main concepts in online communities. The ontology is available online at <http://rdfs.org/sioc/ns#> and its structure is shown in **Figure 1**.

2.1 Main Classes

We list the major classes that are used in the SIOC ontology, and describe their usage in more detail.

Site is the location of an online community or set of communities, with users in groups creating posts on a set of fora. While an individual forum or group of fora are usually hosted on a centralised site, in the future the concept of a 'site' may be extended (for example, a topic thread could be formed by posts in a distributed forum on a peer-to-peer environment).

Forum is a discussion area on which posts are made. A forum can be linked to the site that hosts it. Fora will usually discuss a certain topic or set of related topics. The hierarchy of fora can be defined in terms of parents and children, allowing the creation of structures conforming to topic categories. Examples of fora include mailing lists, online bulletin boards, Usenet newsgroups and weblogs.

Post is an article or message posted by a user to a forum. A series of posts may be threaded and are connected by reply relationships. Posts have content and may also have attached files. Posts may have one or many topics.

User is an online account of a member of an online community. They are connected to posts that they create or edit, to fora that they are subscribed to or moderate, to sites that they administer, and to other users that they know. Users can be grouped for purposes of allowing access to certain fora or enhanced community site features.

2.2 Important Properties

In the next paragraphs, we describe some important properties of SIOC concepts.

topic A topic definition applies to most of the concepts defined above, and topic metadata can be a useful way to match users and posts to each other. Users or group of users can define topics of interest when their profiles are created or modified. As regards posts, while it may be more difficult to require a user to assign a topic to a post at creation time, it is more likely that a forum will have an associated topic or set of topics that can be propagated to the posts it contains. Topics can also be assigned to posts via predefined category hierarchies and free-text keywords (using 'folksonomy' tagging). In order to enable the location of related information across sites, the SKOS framework [7] can be used to define the concepts represented by the topics or tags, and to link topics between community sites.

has_creator The has_creator property links a post to the user profile of its author. Thus, we can follow the link

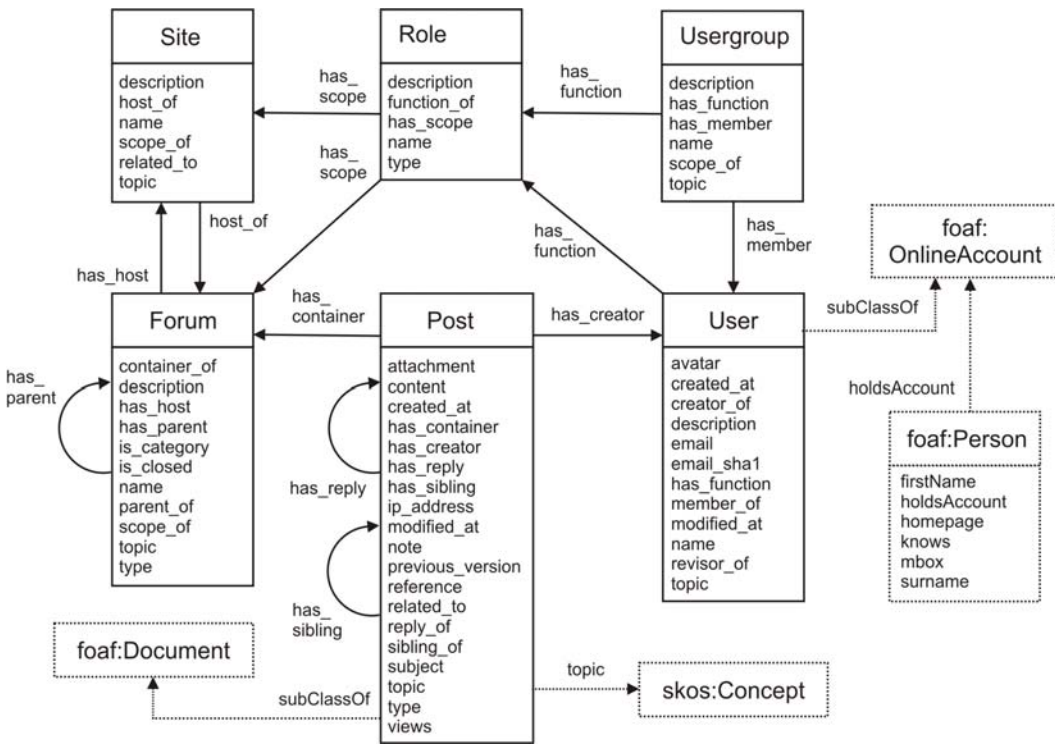

Figure 1: Overview of Classes and Properties Used in SIOC.

SIOC	FOAF	RSS 1.0	Atom
Site	-	-	-
Forum	-	channel	Feed
Post	Document	item	Entry
User	Online Account	-	-

Table 1: Selected SIOC Mappings.

from the post to the creator and locate the other posts by the same person. The community can be seen as a network of posts with users linked to each post, and there is also a network of other posts created by a given user stemming from there.

2.3 Mappings

One of the main functions of SIOC is to provide a means for exchanging community instance data. Since there are already a considerable number of classes and properties defined in RDF on the Web, we provide mappings in RDFs and OWL (Web Ontology Language), <http://rdfs.org/sioc/mappings>, to allow the import and export of SIOC instance data in different vocabularies, such as (Friend Of A Friend, <http://xmlns.com/foaf/0.1/>) and RSS 1.0 / Atom, <http://purl.org/rss/1.0/>. Therefore, we can leverage the instance data that is already available.

In **Table 1** we show how classes in FOAF, RSS 1.0 and Atom correspond to SIOC classes. Mappings of properties are described in a similar manner

Since mappings in SIOC are not only restricted to ontologies, we need to provide a means to extract information from simple data structures. For example, we can map from XML (eXtensible Markup Language) documents such as RSS 0.9x and 2.0 into the SIOC ontology using XSL (eXtensible Stylesheet Language) stylesheets. In this method, titles, descriptions and hyperlinks are extracted from XML documents, somewhat similar to how GRDDL (Gleaning Resource Descriptions from Dialects of Languages, <http://www.w3.org/2004/01/rdxh/spec>) is used to extract information from XHTML (eXtensible HyperText Markup Language) documents.

3 Exchanging Instances

The core use of SIOC will be in the exchange of instance data between sites. In the following, we elaborate on how the exchange, both importing and exporting data, can be carried out. We show how wrappers can help to achieve export functionality, either based on exporting documents containing the information or by rewriting queries. Another solution for incorporating the "document-based" wrapping to mirror exported RDF documents in an RDF store and thus allow for performing queries. We present a third solution, possibly for newly-developed applications, which uses a native RDF repository to store and retrieve statements, making import and export straightforward.

3.1 Wrappers to Existing Tools

Wrappers will allow us to export instances of community site concepts such as fora or posts in RDF. They can also allow us to import SIOC instances to other non-SIOC systems. Systems for which wrappers could be developed can be divided into two categories - legacy systems and web-based systems.

Legacy Systems A large number of systems preceding the current Web are still deployed and widely used on the Internet. Email is used for exchanging messages and files in an asynchronous way, Usenet is still used to exchange messages, and IRC (Internet Relay Chat, <http://www.ietf.org/rfc/rfc1459.txt>) is used for synchronous communication. Therefore, to really capture a large amount of data currently exchanged in online communities on the Internet, these legacy systems and protocols need to be considered for SIOC.

In contrast to web-based systems, where we just need to translate the data, we need to employ protocol wrappers for legacy protocols to HTTP (HyperText Transmissión Protocol). For example, for email we need to translate the data representation format from RFC822, <http://www.ietf.org/rfc/rfc822.txt>, to SIOC, and provide a wrapper to the access protocol for email stores (usually POP3 - *Post Office Protocol version 3* <http://www.ietf.org/rfc/rfc1939.txt>, or IMAP4 - *Internet Message Access Protocol*, <http://www.ietf.org/rfc/rfc1730.txt>).

The email export wrapper accepts a conjunctive query over HTTP GET and returns the results in SIOC. In a next step, the query is parsed and translated into IMAP4 to send to the original data source. The original data source then returns the results in RFC822 format, which is then translated back into RDF and returned to the original caller via HTTP. We have implemented the wrapper and the mapping using Java.

Web-Based Systems Providing mappings from web-based systems is easier than mapping from legacy systems since protocol translation is not needed here.

We will discuss two kinds of community sites here -

bulletin boards and weblogs. All these systems are based on content management systems. Therefore exporting and importing information from and to such systems can be accomplished by adding wrapper interfaces to these systems.

Some export functionality is already available for bulletin boards and CMSs (e.g. FOAF from vBulletin, <http://www.vbulletin.org/forum/>). Most of these systems use open source architecture and a wrapper for them will build on existing libraries such as Magpie RSS, etc. We have provided a module for Drupal CMS, <http://drupal.org/>, that exports SIOC information about Drupal 'nodes'.

Weblogs usually are small scale systems consisting of one or more contributors and a community of readers. Most weblog engines already have RSS export functionality and there are experimental implementations to export metadata, such as the Wordpress FOAF plugin, <http://www.wasab.dk/morten/blog/archives/2004/07/05/wordpress-plugin-foaf-output>. Since majority of these engines are open source software, it is straightforward to modify existing export functions to generate SIOC metadata.

The main challenge for using SIOC with web-based systems are not in the technical implementation of SIOC wrappers, but rather in the wide adoption of the SIOC ontology to gain incentives for people to provide data and tools for SIOC.

By making SIOC data available through exports, we are encouraging the adoption of SIOC concepts. To this end, we have created a SIOC metadata export facility for the WordPress weblog engine, <http://rdfs.org/sioc/wordpress/>. This makes use of existing WordPress PHP (Hypertext Preprocessor) functions to access the information about posts, users and fora (weblog channels) from the underlying relational database. SIOC metadata in RDF is generated for each concept instance. The export process is illustrated by example in Figure 2. Other export facilities are being written for phpBB, vBulletin, and b2evolution, <http://b2evolution.net/>.

3.2 Mirror Data in RDF Store

Most of the web-based wrappers just provide simple document-based export facilities. Since our goal is to make

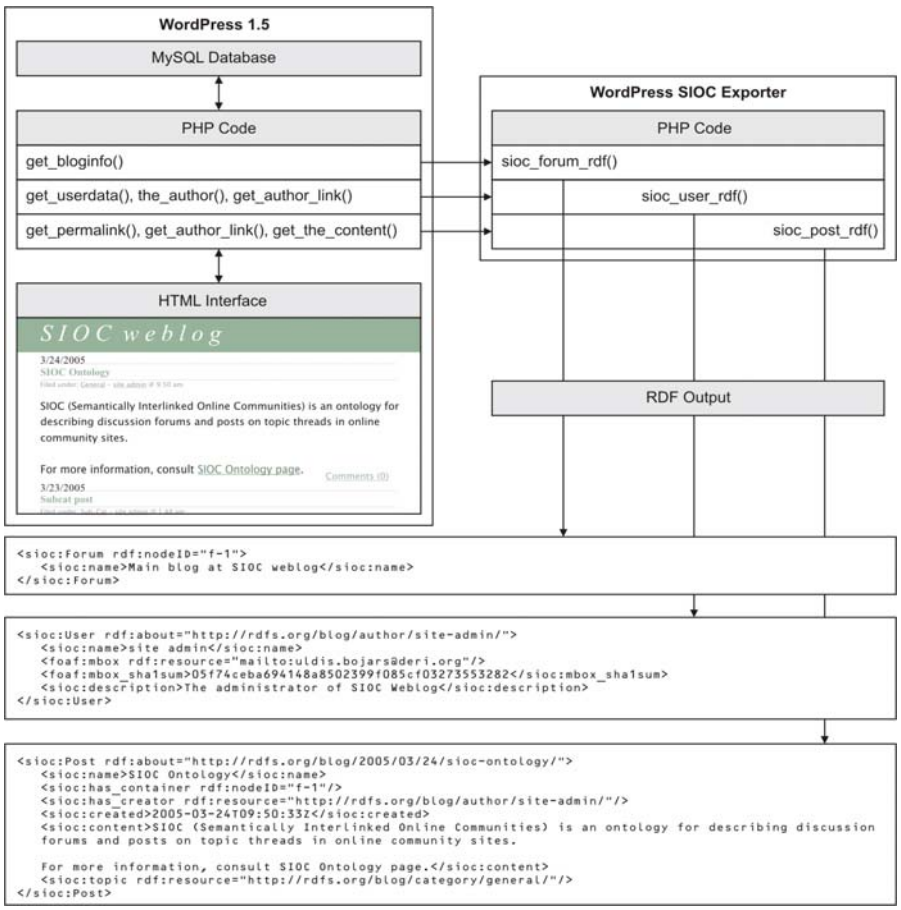

Figure 2: SIOC Metadata Export from WordPress.

SIOC data available for query and to entice people to use SIOC now, we need a method to allow querying of the information that sites publish in flat files.

A solution to provide query facilities for sites that have only simple data export facilities is to replicate the information in a data store that can process queries. Queries are then answered from the replica. The replica is updated either by an RDF crawler that traverses rdfs:seeAlso links, or by the original site that pushes updates and changes automatically into the mirror store.

Replicating the contents of the entire site from the relational database to an RDF store may work initially and create an easy upgrade path. However, in the longer term, storing and integrating data in a native RDF repository is the desirable solution.

3.3 Native RDF Store

The previous two subsections discussed tasks that concerned querying existing sites and their content. We will now describe how newly architectured sites can make use of a native RDF repository to store their data.

Exporting data is quite simple because RDF does not restrict you in the way data can be expressed. On the flip side, the flexibility of RDF creates a problem when importing data into systems with a fixed schema. Issues arise here, for example, when an application is importing data using a given schema, and certain mandatory data is missing.

Since community sites provide access to complex structures of information with different types, it is natural to store that information in RDF directly. Repositories such as Jena2 [10], Sesame [3], Redland [1], or YARS [5] can be used to store and retrieve the data. With an RDF store as the data repository, importing and exporting information is straightforward, and also data integration tasks can be facilitated. An API (Application Program Interface) similar to the RDF NetAPI [9] can be used as well. The route we chose for SIOC is to use HTTP methods such as PUT and DELETE for adding and removing data.

4 Using SIOC Data

Given the ontology, the mappings, and the wrappers, we are now able to pose queries and add data to individual SIOC sites.

4.1 Browsing

Once we have made the data available using a common query infrastructure, we can use various user interfaces to navigate SIOC data. The simplest solution is to use a mapping from SIOC to a data format where client programs already exist. For example, SIOC data can be mapped to email and then read in any email program. Also, a mapping from SIOC to RSS allows us to navigate a subset of SIOC information inside a regular RSS news reader. Since SIOC has a richer data model than RSS, some information will be lost during the conversion.

Another approach is to use existing RDF browsers such as BrownSauce, <http://brownsauce.sourceforge.net/>, or Node browser, <http://rdfs.org/search/>, to view arbitrary RDF data. Leveraging the full potential of SIOC requires the provision of custom programs and user interfaces specially tailored towards SIOC (e.g. for cross-site browsing).

4.2 Query

Representing data in SIOC enables users to pose structural queries against the collected data rather than just having keyword search. An implication of structural queries is that you get precise answers as a result, and not just pieces of documents that match the keyword.

One central problem in P2P networks is how to route queries [8]. We plan to exploit the link structure that connects fora or sites to route queries. The forum and site linkage inside SIOC makes it easier to do routing than in general-purpose peer-to-peer networks, since we have some (human-created) links that can be exploited. We expect a scale-free behaviour of these links once SIOC is widely used in practice.

By building the infrastructure for distributing queries into the different site management software or wrappers, we can perform queries without any central components. As a result, querying inside an intranet will be simple and already integrated into the tools used to manage the different community sites inside an organisation, such as mailing lists or fora.

4.3 Locating Related Information

Querying the community sites for information on demand is not the only model of end-user interaction. Another way to enhance the end-user experience is to prepare the data in advance, at creation time of a post.

Once a new post is created in a community site and the SIOC information is available, this site then queries the network of community sites to find related posts. A query is performed based on the post metadata, such as other posts by this person or other posts in the set of the post's topics. Results are then stored and can be reused to browse forum entries and navigate through the web of interlinked posts, independent of the underlying site structure that the fora and posts are hosted on.

The results of this information retrieval model are the enhanced functionality added to community sites, and better scalability since the information is prepared in advance.

5 Related Work

Harvest is an early system [2] that can be used to gather information from diverse repositories to build, search, and replicate indexes, and to cache objects as they are retrieved across the Internet. Harvest uses the Summary Object Interchange Format (SOIF) to exchange metadata about resources. In contrast, SIOC uses RDF as the exchange format and allows mappings between different vocabularies, which is not envisioned in SOIF.

Issue based information systems (IBIS) model [11] uses argumented discussions in the process of solving design issues and provides a detailed model for links between con-

versations. SIOC uses metadata and reply links to connect conversations on online community sites and can be extended to describe argumented discussions.

Various approaches for data integration on the Web, such as data representation languages, structural information retrieval, and query processing, are surveyed in [4]. However, advanced database techniques have failed so far to surface on the Web. SIOC is a first step in providing a common vocabulary for data representation across online communities.

RDF Site Summary (RSS 1.0) is widely used in weblog systems and news sites. RSS 1.0 defines a lightweight vocabulary for syndicating news items, but is used for all sorts of data exchange. Although RSS works well in practice, there are several issues: firstly, only the last "n" news items are typically exported in RSS. Secondly, most of the systems use non-RDF versions of RSS, which limit its use with other vocabularies.

6 Conclusion

We have presented the SIOC ontology and various mappings to and from other vocabularies that are already deployed on the Web. We have described how instance data in SIOC can be exchanged among online community sites. Our initial SIOC ontology can also be used to enable more complex use cases, for example cross-site structural queries, and integration based on the warehousing approach.

To tackle the challenge of adoption, we have provided an upgrade path that allows a gradual migration from existing systems to semantically-enabled sites. For combination with other ontologies, we have presented mappings to and from SIOC that allow the export and import of SIOC data using existing systems and tools. We have developed prototype SIOC exporters for a weblog engine and a content management system, with several more in development. In the future, we intend to exploit the characteristics of intra- and inter-site links to guide query routing in a P2P-like environment.

Acknowledgements

The authors would like to acknowledge the support of Science Foundation Ireland under Grant No. SFI/02/CE1/I131. This paper was originally presented at the 2nd European Semantic Web Conference (ESWC 2005).

References

[1] D. Beckett. The Design and Implementation of the Redland RDF Application Framework. Computer Networks, 39(5):577–588, 2002.

[2] C. M. Bowman, P. B. Danzig, D. R. Hardy, U. Manber, and M. F. Schwartz. The Harvest information discovery and access system. Computer Networks and ISDN Systems, 28(1–2):119–125, 1995.

[3] J. Broekstra, A. Kampman, and F. van Harmelen. Sesame: A Generic Architecture for Storing and Querying RDF and RDF Schema. In 1st International Semantic Web Conference, pages 54–68, 2002.

[4] D. Florescu, A. Y. Levy, and A. O. Mendelzon. Database Techniques for the World-Wide Web: A Survey. SIGMOD Record, 27(3):59–74, 1998.

[5] A. Harth, S. Decker. Optimized Index Structures for Querying RDF from the Web. 3rd Latin American Web Congress, Buenos Aires, Argentina, October 31 to November 2, 2005, pp. 71-80.

[6] R. Lara, S.-K. Han, H. Lausen, M. Stollberg, Y. Ding, and D. Fensel. An Evaluation of Semantic Web Portals. In IADIS Applied Computing International Conference 2004, Lisbon, Portugal, March 23-26, 2004.

[7] A. J. Miles, N. Rogers, and D. Beckett. SKOS Core RDF Vocabulary. 2004. <http://www.w3.org/2004/02/skos/core/>

[8] W. Nejdl, B. Wolf, C. Qu, S. Decker, M. Sintek, A. Naeve, M. Nilsson, M. Palmer, and T. Risch. EDUTELLA: a P2P networking infrastructure based on RDF. In WWW , pages 604–615, 2002.

[9] A. Seaborne. An RDF NetAPI. In 1st International Semantic Web Conference, pages 399–403, 2002.

[10] K. Wilkinson, C. Sayers, H. A. Kuno, and D. Reynolds. Efficient RDF Storage and Retrieval in Jena2. In Proceedings of SWDB'03, In 1st International Workshop on Semantic Web and Databases, Co-located with VLDB 2003, pages 131–150, 2003.

[11] H. Rittel, W. Kunz. Issues as elements of information systems. Working Paper 131, Berkeley Ca, University of California Center for Planning and Development Research, 1970.