# Bootstrapping the Semantic Web of Social Online Communities

Diego Berrueta Sergio Fernández Lian Shi

{diego.berrueta,sergio.fernandez,lian.shi}@fundacionctic.org
Fundación CTIC
Gijón, Asturias, Spain

## ABSTRACT

Mining and searching the social web is hardly possible without a noteworthy amount of data available in an interoperable format. This paper enumerates and compares several techniques which can be applied to obtain large quantities of RDF data describing social web sites. Advantages, drawbacks and potential issues of each of these methods are discussed. Practical experimentation permits to illustrate and to discuss the convenience of each approach.

## Categories and Subject Descriptors

M.7 [**Knowledge Management**]: Knowledge Retrieval; H.3.1 [**Information Storage And Retrieval**]: Content Analysis and Indexing

## Keywords

semantic mining, online community, mailing list, rdf, xsl, foaf, sioc, semantic web, social web

## 1. INTRODUCTION

Effective large-scale mining of the social web requires the availability of big amounts of well-defined data [16]. The semantic web provides a convenient platform to publish and consume this data. There are a couple of semantic web vocabularies which are particularly suited to represent the information of the social web using an interoperable formalism. However, currently only a small portion of the social web is represented in these vocabularies.

In this paper we survey and classify a number of methods that are targeted to create semantic web enabled representations of the information of the social web. We focus on online communities and discussion forums, although the methods described here are also valid for other social web sites. The combination of all of them may provide enough momentum to the semantic social web and it can help to reach the critical-mass that enables a virtuous cycle of applications and data.

The rest of the paper is organized as follows: in Section 2 the FOAF and SIOC vocabularies are introduced in the context of applying the semantic web to the description of the social web. Section 3 enumerates and classifies some methods to produce large quantities of RDF descriptions. Section 4 and Section 5 address two paradigmatic methods and their experimental application. Some issues which appear frequently are described in Section 6, and finally Section 7 discusses the future of the semantic social web and concludes the paper.

*WWW2008*, April 21–25, 2008, Beijing, China.
.

## 2. SEMANTIC WEB VOCABULARIES TO DESCRIBE THE SOCIAL WEB

The Semantic Web initiative uses RDF (Resource Description Framework [14]) as the (meta)data representation model. Ontologies are artifacts that define the meaning of the symbols which appear in the RDF assertions. Two ontologies are specially relevant to describe the social web: FOAF and SIOC.

FOAF [5] is one of the most used vocabularies in the documents that constitute the current Semantic Web [8, 11]. It provides the essential classes and properties necessary to describe individuals and their relationships. However, FOAF descriptions can be found only for a very small portion of the web users. Bootstrapping the FOAF-web by means of mining the document-web is the topic of [16]. Consequently, the focus of our work is not the description of people. Instead, this paper focuses on the description of the interactions between people in online discussion communities.

DERI NUI Galway leads the development of SIOC (Semantically-Interlinked Online Communities[1]), an ontology which defines a vocabulary to interconnect different discussion methods such as blogs, web-based forums and mailing lists [4]. SIOC is now an official W3C member submission [3]. SIOC provides the foundations to describe online discussion communities using RDF (users, forums and posts), as illustrated in Figure 1.

In the rest of the paper, we describe methods to create a large quantity of RDF instances of the SIOC ontology.

## 3. METHODS TO MASS-PRODUCE SIOC INSTANCES FROM EXISTING SOURCES

Since SIOC is a recent specification, its adoption is still low, and only a few sites export SIOC data. There exist a number of techniques that can be used to bootstrap a network of semantic descriptions from current social web sites. We classify them in two main categories:

- On the one hand, methods which require direct access to the underlying database behind the social web site are *intrusive techniques.*

[1] `http://sioc-project.org/`

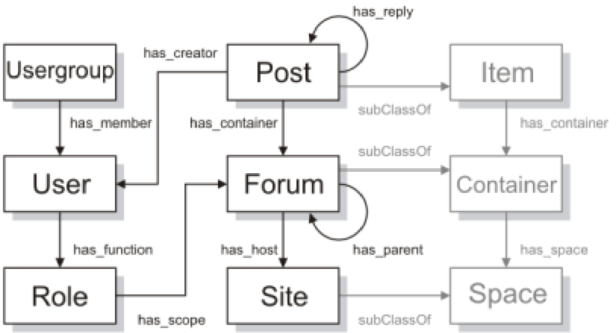

**Figure 1: SIOC main classes and properties, reproduced from the SIOC specification.**

- On the other hand, methods which do not require direct access to the database and can operate on resources already published on the web are *non-intrusive*.

We further describe each kind in the following paragraphs.

## 3.1 Intrusive techniques

It is safe to say that every social web site is built on top of a database that serves as its information model. The web application acts as the controller and publishes different views of the model in formats such as HTML and RSS. In terms of this pattern, publishing SIOC data is as simple as adding a new view. From a functional point of view, this is the most powerful scenario, because it allows a lossless publication due to the direct access to the back-end database.

The SIOC community has contributed a number of plugins for some popular web community-building applications, such as Drupal, WordPress and PhpBB[2]. Mailing lists are also covered by SWAML, which is described later in this paper.

There is, however, a major blocker for this approach. All these software components need a deployment in the server side (where the database is). This is a burden for system administrators, who are often unwilling to make a move that would make it more difficult to maintain, keep secure and upgrade their systems. This is particularly true when there is no obvious immediate benefit of exporting SIOC data (*chicken-and-egg* problem).

## 3.2 Unintrusive techniques

In absence of direct access to the database, unintrusive techniques exploit the information already available on the web:

- Cooked HTML views of the information, the same ones that are rendered by web browsers for human consumption. Even if this source is always available, its exploitation poses a number of issues described in Section 6.

- RSS/Atom feeds, which have become very popular in the recent years. They can be easily translated into SIOC instances using XSLT stylesheets (for XML-based feeds) or SPARQL queries[3] (for RSS 1.0, which is actually RDF). Unfortunately, these feeds often contain just partial descriptions.

- Public APIs. The Web 2.0 trend has pushed some social web sites to export (part of) their functionality through APIs in order to enable their consumption by third-party mash-ups and applications. Where available, these APIs offer an excellent opportunity to create RDF views of the data.

A shared aspect of these sources is their ubiquitous availability through web protocols and languages, such as HTTP and XML. Therefore, they can be consumed anywhere, and thus system administrators are freed of taking care of any additional deployment. In contrast, they cannot compete with the intrusive approaches in terms of information quality, as their access to the data is not primary.

# 4. CASE STUDIES

In this section we describe an example of each of the two approaches aforementioned.

## 4.1 From mboxes to RDF: SWAML

SWAML [10] is a Python script that reads mailing list archives in raw format, typically stored in a “mailbox” (or “mbox”), as defined in RFC 4155 [12]. It parses mailboxes and outputs RDF descriptions of the messages, mailing lists and users as instances of the SIOC ontology. Internally, it re-constructs the structure of the conversations in a tree structure, and it exploits this structure to produce links between the posts.

This script is highly configurable and non-interactive, and has been designed to be invoked by the system task scheduler. This low-coupling with the software that runs the mailing list eases its portability and deployment.

SWAML classifies as an intrusive technique because it requires access to the primary data source, even if in this case it is not a relational database but a text file. Anyway, it is worth mentioning that some servers publish these text files (mailboxes) through HTTP. Therefore, sometimes it is possible to retrieve the mailbox and build a perfect replica of the primary database in another box. In such cases, SWAML can be used without the participation of the system administration of the original web server.

## 4.2 HTML Scraping with XSLT

Web scraping is a well-known, non-intrusive and widely used technique (see [7] for a review of several HTML scraping applications), although it should be applied only when other approaches are not viable. Screen-scraping applications are difficult to maintain and often produce low-quality information.

A popular language to write HTML scrapers is XSLT. As a prerequisite, the mark-up must be converted to XHTML (an XML dialect) if it is not already in this format. Fortunately, open source utilities such as Tidy[4] do a decent job to clean and fix HTML files with a poor mark-up.

Scraping functions are often tied to a web crawler to follow the links between HTML pages.

As each web site uses a different, customized template to publish their cooked HTML files, it is difficult to develop a generic scraper, even for a single social web appli-

[2] A more complete and up-to-date list is available at `http://sioc-project.org/exporters`.

[3] `http://tinyurl.com/2u3e6k`

[4] `http://tidy.sourceforge.net/`

cation. Moreover, there are lots of different social and web community-building applications, and thus the portability of scrapers is very low.

The output of a web scraper implemented in XSLT is usually RDF/XML, but another interesting possibility has already been explored. In mle [13], the authors use XSLT to decorate the DOM tree of an XHTML page with RDFa attributes [2]. This creates an hybrid representation which is readable for humans *and* semantic web agents.

## 5. EXPERIMENTATION

Some experiments were run following the case studies described above. We chose the Free Software communities as the target of our experimentation because they are characterized for their openness and they offer a huge number of very popular online discussion forums. Among those, we picked two clusters: the GNOME project mailing lists and the Debian mailing lists. Although both of them contain the same kind of forum (mailing lists), they are tackled with different methods, as explained below. The result, anyway, is the same in both cases: a big dataset of RDF instances that can be uploaded to an RDF store such as Sesame [6]. In this way, they can be queried and mined. Moreover, it is also possible to execute rules or inference engines to obtain new knowledge.

### 5.1 GNOME mailing lists

GNOME is a graphical desktop environment available as free software. It has a vibrant community of users and developers who communicate over the Internet using IRC and mailing lists. The web site of the GNOME project publishes the complete archive of near 200 mailing lists during the ten year of activity of the project[5]. These archives are published not just as cooked HTML files for human consumption, but also as gzipped mailboxes split by month and mailing list.

A simple shell script was run to fetch, unpack and concatenate the mailboxes into a single file for each mailing list. These files were provided as input data to SWAML. The result was a dataset that contains more than 25 million RDF triples.

### 5.2 Debian mailing lists

Debian is a compilation of free software that constitutes a complete operating system [15]. It is the result of a collaborative effort by a thousand developers since 1993, and it has millions of users (as often happens with open source software, it is very difficult to estimate the actual number of users). Together, developers and users constitute a very active community, and they use the web and mailing lists as communication channels. Some of the 180 official mailing lists have almost 30,000 members and up to 7,000 messages/month[6].

The mailboxes of these mailing lists are not available on the web, but the complete archive of the messages is published as a set of HTML files generated by MHonArc. We crawled a subset of these files (11 mailing lists in the period 2005-2006) and collected almost 220,000 messages. The mark-up of each file was fixed and converted to XHTML Strict using Tidy. Finally, an XSLT stylesheet was applied to produce a RDF description for each message. In order to simplify the task of translating dates to a uniform ISO format, we extended the Xalan XSLT processor with custom functions implemented in Java.

The resulting dataset sums up more that 3 million RDF triples. The memory space required to store this dataset can be notably reduced if the body of the messages is dropped, i.e., if only the meta-data is kept. We envision that many mining applications will not need the body of the messages.

## 6. COMMON ISSUES

When put into practice, some shared problems and limitations of the approaches described above are revealed:

- *Same person, multiple identities.* A single individual can participate in several social web sites, often under a different virtual identity. Over the years, this individual can own a number of user accounts and e-mail addresses, which are modelled as different entities in SIOC. If each of these were taken as a different person, social web mining would lead to imprecise conclusions.

  FOAF separates the description of a person (`Person`) from the description of her user accounts (`OnlineAccount`). This makes it possible to establish one-to-many relationships between these entities.

  From the perspective of an automatic processing of the information, the challenge is to build these links. In an ideal scenario, the FOAF description of an individual would contain such links. In practice, these links must be inferred from coincident values of some properties such as the e-mail address or the URL of the personal home page. In the worst case, the only way to go is to perform heuristic matching using the person's name or nickname.

  A comprehensive knowledge base of FOAF descriptions can prove very useful in this task. However, it introduces another related issue: there may exist more than one instance of `foaf:Person` in the knowledge base to describe the same person. Consolidation of these instances is a similar problem to the one just described, and receives the name of "instance smushing"[7]. For our experiments, we crawled a dataset 4,000 FOAF descriptions from Advogato[8], a social network for free software developers.

- *Hashed e-mail addresses.* Both FOAF and SIOC rely on the sha1 algorithm [9] to represent e-mail addresses in an opaque way. The main purpose to do so is to block spammers, who otherwise would find it easy to collect e-mail addresses. It is assumed that hashed e-mail addresses retain their capability as unique identifiers of the resources. However, neither the FOAF nor the SIOC specification describe a normalization procedure to be applied to the e-mail address, besides adding the `mailto:` prefix. This is unfortunate, because it fails to prevent equivalent e-mail addresses from producing different hashed values. For instance, it is common to find spelling variants of the same e-mail address which only differ on the use of lower and

[5] `http://mail.gnome.org/archives/`

[6] Information collected from `http://lists.debian.org/stats/`, Feb 10, 2008.

[7] `http://esw.w3.org/topic/RdfSmushing`

[8] `http://www.advogato.org/`

upper-case letters. These variants produce irreconcilable values when the hash function is applied, thus making them unusable to spot equivalent instances of the same resource.

- *Flat threads.* Typically, web-based discussion forums and blogs have flat threads, i.e., all the replies are attached to the original post that started the thread. However, discussions hosted in those forums often violate this restrictive pattern, and some messages are in fact replies to some of the precedent ones. Users often quote the actual message they are replying to in order to clarify the flow. Unfortunately, it is difficult to automatically re-build the actual hierarchical structure of the conversation. Therefore, when converted to SIOC, some information about the sequence of the discussion is lost.

  The situation is completely different for mailing lists, because each new post contains a header (`In-Reply-To`) that points to the immediate parent in the thread hierarchy.

- *Repeated primary keys.* Every online discussion community assigns an identifier (primary key) to each message and user. These identifiers are locally unique, and can be used to coin a URI for each resource. Mailing lists also use identifiers (`Message-ID`, as defined in RFC 2822 [17]) for each e-mail message, although in this case, such identifiers are supposed to be globally-unique. However, non RFC-compliant or improperly configured mail transport agents can potentially produce repeated identifiers for e-mail messages. Our experimentation has revealed that it is possible to find clashes among the messages of a mailing list. This fact leads to two consequences. Firstly, `Message-ID`s cannot be used to coin URIs. Secondly, links between messages, such as those created by the `In-Reply-To` header, are not fully reliable.

- *Pagination.* Long discussions and indexes are often paginated into several inter-linked HTML files. Although web crawlers can retrieve all parts, this fragmentation poses a challenge for scraping the information. It is often necessary to re-join the different pages in order to produce a consistent RDF representation of the information.

## 7. CONCLUSIONS

Machine-readable descriptions of online communities enable them to be mined in a more efficient way. So far, the availability of such descriptions has been low. The semantic web provides the best framework to publish and consume formalized descriptions of the artifacts that are part of the social web. We contribute to the social semantic web by reviewing and evaluating some approaches to produce RDF instances, and by providing a large amount of instances.

Each scenario dictates different requisites, and thus, a different technique. The intrusive ones are clearly preferred due to their closeness to the primary source (the database), but sometimes they may be unpractical because of deployment issues.

In the long term, we foresee that FOAF and SIOC-enabling plug-ins will become commodities in the software that supports the social web. The RSS is the most immediate precedent of a similar semantic web technology which has permeated into the mainstream web. These information-rich descriptions will be available for both machines and humans by means of HTTP content negotiation [1] or hybrid representations (RDFa).

## 8. REFERENCES

[1] D. Berrueta and J. Phipps. Best practice recipes for publishing RDF vocabularies. Working draft, W3C, 2008.

[2] M. Birbeck, S. Pemberton, and B. Adida. RDFa Syntax, a collection of attributes for layering RDF on XML languages. Technical report, W3C, 2006.

[3] U. Bojars and J. G. Breslin. SIOC core ontology specification. Member submission, W3C, 2007.

[4] J. Breslin, S. Decker, A. Harth, and U. Bojars. SIOC: an approach to connect web-based communities. In *International Journal of Web Based Communities*, 2006.

[5] D. Brickley and L. Miller. FOAF Vocabulary Specification. Technical report, 2005.

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

[7] P. Coetzee, T. Heath, and E. Motta. SparqPlug: Generating linked data from legacy HTML, SPARQL and the DOM. In *Proceedings of Linked Data on the Web*, 2008.

[8] L. Ding, L. Zhou, T. Finin, and A. Joshi. How the semantic web is being used: An analysis of foaf documents. In *Proceedings of the 38th International Conference on System Sciences*, 2005.

[9] D. Eastlake and P. Jones. RFC 3174: US Secure Hash Algorithm 1 (SHA1). Technical report, IETF, 2001.

[10] S. Fernández, D. Berrueta, and J. E. Labra. Mailing lists meet the semantic web. In *BIS 2007 Workshop on Social Aspects of the Web*, 2007.

[11] T. Finin, L. Ding, L. Zhou, and A. Joshi. Social networking on the semantic web. *The Learning Organization: An International Journal*, 12(5):418–435, May 2005.

[12] E. Hall. RFC 4155 - the application/mbox media type. Technical report, The Internet Society, 2005.

[13] M. Hausenblas and H. Rehatschek. mle: Enhancing the exploration of mailing list archives through making semantics explicit. In *Semantic Web Challenge 2007*, 2007.

[14] G. Klyne and J. J. Carroll. Resource Description Framework (RDF): Concepts and abstract syntax. Technical report, W3C Recommendation, 2004.

[15] M. Krafft. *The Debian System*. No Starch Press, 2005.

[16] P. Mika. Bootstrapping the FOAF-web: An experiment in social network mining. In *1st Workshop on Friend of a Friend, Social Networking and the Semantic Web*, 2004.

[17] P. Resnick. RFC 2822 - internet message format. Technical report, The Internet Society, 2001.