MAILING LISTS MEET THE SEMANTIC WEB

Sergio Fernández and Diego Berrueta *Fundación CTIC, Parque Científico y Tecnológico, Cabueñes, Gijón, Spain. sergio.fernandez@fundacionctic.org, diego.berrueta@fundacionctic.org*

Jose E. Labra *Universidad de Oviedo, Computer Science Department, Campus de los Catalanes, Oviedo, Spain. labra@uniovi.es*

ABSTRACT

Mailing list archives (i.e., the compilation of the messages posted up-to-now) are often published on the web and indexed by conventional search engines. They store a vast knowledge capital. However, the ability to automatically recognize and process the information is mostly lost at publishing time. As a result, the current mailing list archives are difficult to query and have a limited use. This paper describes an usage of the Semantic Web technologies in order to avoid the information loss and to allow new applications to exploit the information in a more powerful way.

KEYWORDS

Mailing list, Semantic web, SIOC, RDF, ontology

1. INTRODUCTION

Electronic mail (e-mail) remains one of the most popular applications of the Internet. Besides direct messaging between individuals, mailing lists exist as private or public forums for information exchange in communities with shared interests. Mailing list archives are compilations of the previously posted messages that are often converted into static HTML pages for their publication on the web. They represent a noteworthy portion of the contents that are indexed by web search engines, and they capture an impressive body of knowledge that, however, is difficult to locate and browse.

The root of these problems can be traced back to the translation procedure that is run to transform the e-mail messages into static HTML pages. This task is fulfilled by scripts that create an static HTML page for each message in the archive. In addition, some indexes (by date, by author, by thread) are generated and usually splitted by date ranges to avoid excessive growth.

On the one hand, this fixed structure reduces the flexibility when users browse the mailing list archives using their web browsers. On the other hand, some of the meta-data that were associated to each e-mail message are lost when the message is rendered as HTML for presentational purposes.

We propose to use an ontology and RDF (Resource Description Framework (Klyne 2004)) to publish the mailing list archives into the (Semantic) web, while retaining the meta-data that were present in the messages. Additionally, by doing so, the information could be merged and linked to other vocabularies, such as FOAF.

The rest of the paper is organized as follows: Section 2 introduces the SIOC ontology and our extensions to it, and then some software applications are described in Section 3. We close the paper with the conclusions and a discussion on future plans in Section 4.

2. SIOC

An ontology to capture the meta-data of a discussion forum, such as a mailing list, was clearly recognized as the first milestone to fulfill the purpose of the project. Fortunately, DERI Galway has developed SIOC (Semantically-Interlinked Online Communities, http://sioc-project.org/), an ontology that provides a vocabulary to interconnect different discussion methods such as blogs, web-based forums and mailing lists (Breslin 2005, Breslin 2006). Indeed, SIOC has a wider scope than just mailing lists, and groups all kinds of online discussion primitives in a generic sioc:Forum concept. Each forum represents an online community of people that share a common interest. The goal of SIOC is to interconnect these online communities. Other relevant concepts of the ontology are sioc:User and sioc:Post, which model respectively the members of the communities and the content they produce.

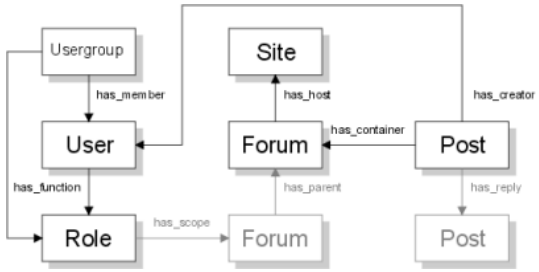

Figure 1. SIOC ontology terms

The SIOC ontology was designed to express the information contained both explicitly and implicitly in Internet discussion methods. Several software applications, usually deployed as plug-ins, are already available to export SIOC data from some popular blogging platforms and content management systems. The effort, however, is focused on web-based communities (weblogs, webforums), while little has been done so far to extend the coverage to legacy non-web communities, such as mailing lists and Usenet groups.

SIOC is specified in OWL, and their instances can be expressed in RDF. Therefore, they can be easily linked to other ontologies. The obvious choice here is FOAF (Brickley and Miller, 2005), which provides powerful means to describe the personal data of the members of a community.

2.1 Extending SIOC Ontology

SIOC is an almost perfect match for our purpose. Each mailing list becomes an instance of sioc:Forum, messages sent to the list become instances of sioc:Post (as well as their replies), and the people subscribed to the list are sioc:Users. The Dublin Core (Dublin Core Metadata Element Set, Version 1.1, 2006) vocabulary is used to capture meta-data such as the message date or title.

```
<rdf:RDF
  xmlns:dcterms='http://purl.org/dc/terms/'
  xmlns:sioc='http://rdfs.org/sioc/ns#'
  xmlns:rdf='http://www.w3.org/1999/02/22-rdf-syntax-ns#'
  xmlns:swaml='http://swaml.berlios.de/ns/0.2#'
  xmlns:dc='http://purl.org/dc/elements/1.1/'
  xml:base='http://swaml.berlios.de/demo/'>
  <sioc:Post rdf:about="2006-Oct/post-50.rdf">
    <dc:title>SIOC properties cardinality</dc:title>
    <sioc:has_creator rdf:resource="subscribers.rdf#s4"/>
    <dcterms:created>Thu, 12 Oct 2006 23:59:26 +0200</dcterms:created>
    <sioc:content><!-- ommitted --></sioc:content>
    <sioc:has_reply rdf:resource="2006-Oct/post-51.rdf"/>
    <swaml:previousByDate rdf:resource="2006-Oct/post-49.rdf"/>
    <swaml:nextByDate rdf:resource="2006-Oct/post-51.rdf"/>
  </sioc:Post>
</rdf:RDF>
```

Figure 2. SIOC Post example in RDF/XML

However, additional object properties were required in order to retain the sequence of messages published in a mailing list. Thus, we extended the SIOC ontology with two properties defined in a separate namespace: swaml:previousByDate and swaml:nextByDate. Both properties are defined with sioc:Post as their domain and range. An RDF representation of a sample message is shown in Figure 2.

3. SOFTWARE TOOLS

The ontology itself provides no service to end users. Software tools are required, and we built two of them as part of this project[1]:

- SWAML is a non-interactive, command-line application whose main purpose is to translate mailboxes into sioc:Forum instances in RDF.
- Buxon is a graphical browser for sioc:Forum instances.

1 Our applications are available at http://swaml.berlios.de/

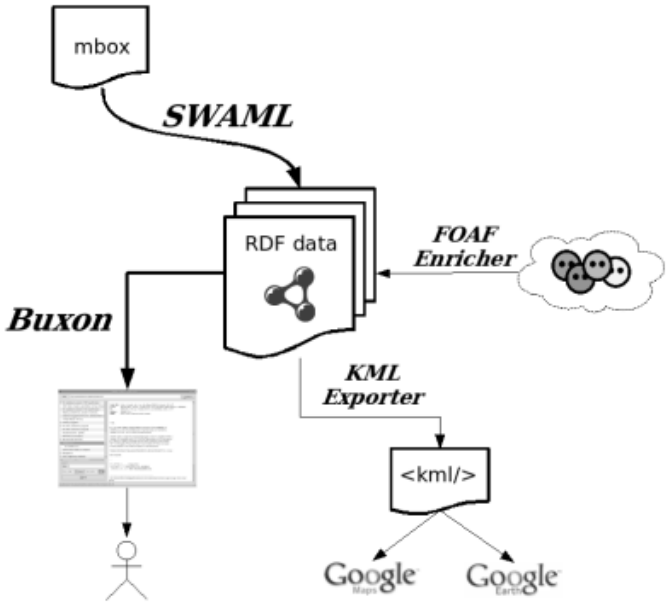

Figure 3. Buxon is an end-user application that consumes sioc:Forum instances, which in turn can be generated from mailboxes using SWAML.

Each tool has a precisely defined role, fulfilling the need to generate RDF data and to consume data, respectively, as depicted in Figure 3. The following paragraphs provide further detail on SWAML and Buxon.

3.1 SWAML

SWAML covers the data-generation phase, and it is intended to be used by mailing list administrators, who usually have access to the archives in raw format. The most popular format for mailing list archives is the "mailbox" (or "mbox"), as defined in RFC 4155 (Hall 2005). SWAML is essentially a mailbox parser implemented in Python. Its output is a number of SIOC instances (Forum, Posts and Users) in a set of RDF files. SWAML is a highly configurable, non-interactive application designed to be invoked by the system task scheduler.

Parsing the mailbox and rebuilding the discussion threads may be sometimes tricky. Although each mail message has a supposedly unique identifier in its header (Message-ID, defined by RFC 2822 (Resnick 2001)), in practice its uniqueness cannot be taken for granted. Actually, we have found some messages with repeated identifiers in some mailing lists, probably due to non-RFC compliant mail transport agents. Therefore, SWAML assumes that any reference to a message (such as those created by the In-Reply-To header) is in fact a reference to the most recent message with that ID in the mailbox (obviously, only previous messages are considered). Using this rule of thumb, SWAML builds an in-memory tree representation of the conversation threads, so sioc:Posts can be properly linked.

Actually, SWAML goes further than just a format-translation tool. A dedicated subroutine that runs as part of the batch execution, but may be also separately invoked on any sioc:Forum, tries to find a FOAF description for each sioc:User. To the best of our knowledge, there is not any web service to fetch FOAF descriptions from a given e-mail address, so we mocked it. Some of the authors of this paper are also currently working on a functional implementation of such a service as part of a different project.

The last step of the SWAML processing chain generates a KML (Ricket 2006) file that contains the geographical coordinates of the mailing list subscribers. The information is fetched from their FOAF descriptions, therefore it is only available for those subscribers whose FOAF description contains their coordinates using the basic geo vocabulary by Dan Brickley (Brickley 2006). Figure 4 depicts a graphical representation of the KML file for a sample mailing list.

Figure 4. Plotting the geographical coordinatesof the members of a mailing list using Google Maps.

3.2 Buxon

Buxon is a multi-platform desktop application written in PyGTK. It allows end users to browse the archives of mailing lists as if they were using their desktop mail application. Buxon takes the URI of a sioc:Forum instance (for example, a mailing list exported by SWAML, although any sioc:Forum instance is valid) and fetches the data, retrieving additional files if necessary. Then, it rebuilds the conversation structure and displays the familiar message thread list (see Figure 5).

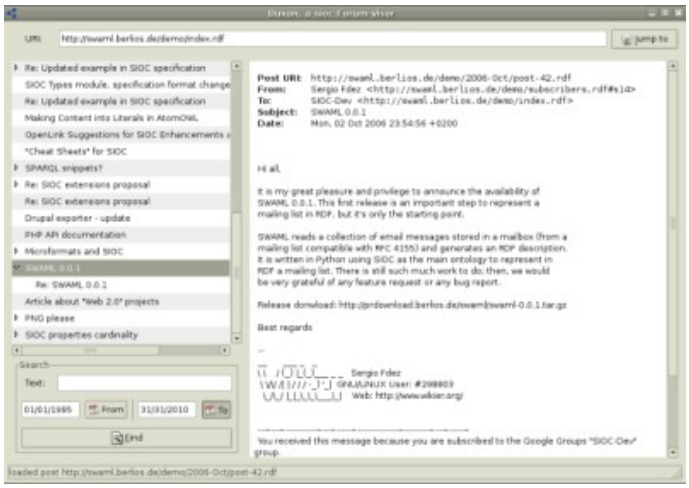

Figure 5. Buxon browsing SIOC-Dev mailing list.

Buxon also gives users the ability to query the messages, searching for terms or filtering the messages in a date range. All these queries are internally translated to SPARQL (Clark 2006) to be executed over the RDF graph, see Figure 6. Newer versions of Buxon can, at user's request, send the sioc:Forum URI to PingTheSemanticWeb.com, a social web service that tracks semantic web documents. That way, Buxon contributes to establish an infrastructure that lets people easily create, find and publish RDF documents.

```
PREFIX sioc: <http://rdfs.org/sioc/ns#>
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
SELECT ?title
FROM <http://swaml.berlios.de/demo/index.rdf>
WHERE
{
  ?x rdf:type sioc:Forum .
  ?x sioc:container_of ?message .
  ?message sioc:has_creator ?creator .
  ?creator sioc:name "Diego Berrueta" .
  ?message dc:title ?title
}
```

Figure 6. SPARQL query to extract all the posts sent by a given person to any sioc:Forum instance.

4. CONCLUSIONS AND FUTURE WORK

There is a lot of ongoing effort to translate data already reachable on the web into formats which are Semantic Web-friendly. Most of that work focuses on relational databases, micro-formats and web services. However, at the time of this writing and to the best of our knowledge, e-mail was almost excluded from the Semantic Web. This project, in combination

with the generic SIOC framework, fills this gap, conveniently providing an ontology and a parser to publish machine-readable versions of the archives of the countless mailing lists that exist on the Internet.

The SWAML project fulfills a much-needed requirement for the Semantic Web: to be able to refer to semantic versions of e-mail messages and their properties using resource URIs. By re-using the SIOC vocabulary for describing online discussions, SWAML allows any semantic web document (in particular, SIOC documents) to refer to e-mail messages from other discussions taking place on forums, blogs, etc., so that distributed conversations can occur across these discussion media. Also, by providing e-mail messages in SIOC format, SWAML is providing a rich source of data, namely mailing lists, for use in SIOC applications.

Some benefits arouse from the availability of these data. In the first place, data can be fetched by user applications to provide handy browsing through the archives of the mailing lists, providing features that exceed what is now offered by static HTML versions of the archives on the web.

Secondly, the crawlers of the web search engines can use the enhanced expressivity of the RDF data to refine search results. For instance, it becomes possible to filter out repeated messages, advance in the fight against spam, or introduce additional filter criteria in the search forms.

Another consequence of no lesser importance is that each e-mail message is assigned a URI that can be resolved to a machine-readable description of the message. This actually makes possible to link a message like any other web resource, and therefore enriches the expressivity of the web.

We are exploring some directions for future work. Some of them are:

- Integration of the SWAML process with popular HTML-based mailing list archivers, such as Hypermail or Pipermail, would be a giant push to speed up the adoption of SWAML. It is well known that one of the most awkward problems of any new technology is to gain a critical mass of users. The semantic web is not an exception. A good recipe to tackle this problem is to integrate the new technology into old tools, making a smooth transition without requiring any extra effort from users. Merging the SWAML process into the batch flow of tools such as Hypermail would allow to generate both HTML and RDF versions of the archives. Those could reside side-by-side on the web server, even sharing the same URI by means of content-negotiation (Miles 2006).
- Actually, integration could be pushed further away through RDFa (Birbeck 2006), embedding the RDF content into the XHTML documents.
- So far, no semantic annotation relative to the meaning of the messages is considered. Obviously, such information can not be automatically derived from a RFC 4155-compliant mailbox. However, it is conceivable that it can be added by other means, such as social tagging using folksonomies, or parsing the RDFa that may exist in the e-mail messages that are sent in XHTML format. The inherent community-based nature of mailing lists can be exploited to build recommendation systems (Celma 2006).
- The meta-data extracted from a mailing list archive can grow quite huge. Even if the body of the messages is omitted, the RDF/XML meta-data of a mailing list containing 1,000 messages may have a size of 4 MBytes, with a linear growth. It is not uncommon for a busy mailing list to generate such volume of messages monthly. Hence, it becomes imperative to provide a mechanism to fragmentate the dataset. The SWAML process splits each message in a separate RDF document, but this arbitrary decision clearly does not fit every application. A much better solution would be to create an easy-to-deploy SPARQL endpoint (Clark 2006), effectively translating the decision on how to partition the data to

the final application (Pan 2006).

- It is not always possible to obtain a mailbox file for a mailing list. For these cases, an alternative is envisaged: a high-capacity mail account can be subscribed to the mailing list with the unique purpose of collecting and storing the messages. A simple extension to SWAML that makes it possible to read the contents of a GMail account has been developed.

ACKNOWLEDGEMENTS

The authors would like to express their gratitude to Dr. John Breslin and Uldis Bojars from DERI Galway, whose support and contributions have been of great help to this project. Also to Ignacio Barrientos by his contribution packaging the project for Debian GNU/Linux.

REFERENCES

Birbeck, M. et al, 2006. RDFa Syntax, a collection of attributes for layering RDF on XML languages, Technical report, W3C.

Breslin, J. et al, 2006. SIOC: an approach to connect web-based communities. *International Journal of Web Based Communities*, Vol. 2, No. 2, pp 133-142.

Breslin, J. et al, 2005. Towards Semantically-Interlinked Online Communities. *Proceedings of the 2nd European Semantic Web Conference, ESWC 2005*, Heraklion, Crete, Greece.

Brickley, D., 2006. Basic geo (WGS84 lat/long) vocabulary, Technical report, W3C Informal Note.

Brickley, D. & Miller, L., 2005. FOAF Vocabulary Specification, Technical report.

Celma, O., 2006. Foafing the music: Bridging the semantic gap in music recommendation. *Proceedings of the 5th International Semantic Web Conference*, Athens, USA.

Clark, K. G., 2006. SPARQL protocol for RDF, Technical report, W3C Candidate Recommendation.

Dublin Core Metadata Element Set, Version 1.1, 2006. Technical report.

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

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

Miles, A. et al, 2006. Best practice recipes for publishing RDF vocabularies, Technical report, W3C Working Draft.

Pan, Z. et al 2006. An investigation into the feasibility of the semantic web, Technical Report LU-CSE-06-025, Dept. of Computer Science and Engineering, Lehigh University.

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

Ricket, D, 2006. Google Maps and Google Earth integration using KML, in American Geophysical Union 2006 Fall Meeting.