Publishing Mailing Lists on the Semantic Web

Diego Berrueta <diego@berrueta.net>

June 2005

Abstract

Mailing list archives (i.e., 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 hard to query and have a limited use. This paper describes the use of the Semantic Web technology in order to avoid the information loss and to allow new applications able to exploit this information in a more convenient way.

Keywords

Mailing list, Semantic Web, archives, RDF.

1 Introduction

Mailing lists are one of the best communication tools on the Internet. They cover any possible topic. Nowadays, mailing list archives (i.e., the previously posted messages) are often published on the web, making them available for indexation by the search bots. Users can read the archives using their web browser, and no previous subscription is required. We can also query the archives using general search engines like Google.

These archives make a huge knowledge base, in particular for technical areas. A common pattern usage is to type or paste an error message from a software application into the Google search box. In many cases, mailing lists posts are part of the search results, probably because somebody has already asked for help in a mailing list. If we are lucky, we can then browse to a reply post containing hints to solve the problem.

2 Problems of the Current Model

Unfortunately, browsing the list archives with a web browser is not as flexible as doing it with our favorite e-mail client. For instance, there are some simple tasks that, in general, cannot be performed with a web browser:

- To show the conversation thread as a tree.
- To print the whole thread.
- To filter the messages in a given date range.
- To hide the messages without replies.
- To show only the messages from a given author.
- To search a string just in one thread.
- To download a whole thread into a file for offline browsing.
- To reply to a message using an e-mail client (or a webmail), quoting parts of the original one.

It is not uncommon for mailing list archives to be mirrored at different sites. Unfortunately, search bots are rarely aware of mirrors, and it is very difficult to detect duplicated messages. As a consequence, sometimes the same message appears more than once in the results page of the search engine. This is annoying for users. Obviously, the ideal behavior would be not to show any duplicate result.

3 The Source of the Problems: Information Loss

The origin of these problems can be traced to the information loss that happens when e-mail messages are transformed to HTML for the web.

Some of the most popular mailing list managers (such as Mailman or Majordomo) store the messages into a file in Mailbox (mbox) format. Later, other independent applications read this file and create static HTML webpages, usually one page for each message. In addition, they can generate complex indexes by date, author or thread, and cross-references to the previous and next messages. But even it the best case, this information is of use only for humans, but not recognizable as such by machines. As a consequence, any possibility of further automatic processing is almost lost.

Some of the information pieces lost in this process are:

- The message subject.
- The name and address of the author.
- The date of the message.
- The reference to the mailing list in which the message was published.
- The reference to the previous message, if it exists.
- The references to the replies by other users.

4 Our Proposal

The technology of the Semantic Web, in particular, the RDF framework, is perfectly suited for preserving all the information described above. As all this information is already available in the original format (the Mailbox file), a minimal amount of additional processing is required (just some work for generating cross-references and indexes). The only requirement is that the transformation process must publish this information in RDF beside the HTML version. With little effort, the mailing lists can be aware of the Semantic Web.

5 Applications

The enrichment of the mailing list archives with machine-processable data opens the door to a number of new applications:

- Search engines may avoid showing duplicated results with the same mailing list message. In order to achieve this, their crawlers should recognize the mirrors of the web archives.
- New features, such as those pointed in Section 2, can be implemented in the web browsers. These new features, probably in the form of plug-ins or extensions to current web browsers, would enhance user's experience when reading the mailing list archives through the web.
- Specialized software agents could gather more information about mailing list subscribers, like a compilation of all the mailing lists in which a given user has posted messages. Linking the information with FOAF [1] would make possible to retrieve the photographs of all the posters in a given mailing lists (like it is currently done -by hand- in some blog aggregators[2]), or plot their position on a map[3].
- Easier localization of the user's interface, without being dependent on the language in which the HTML archives are published on.
- Improved accessibility for disabled people, by means of alternative presentations of the information (for instance, text-to-speech).

6 Related Work

There exist some similar approaches.

The DOAML vocabulary [4] is a RDF vocabulary designed to describe mailing lists. The project homepage includes some examples using the W3C mailing lists. Nevertheless, the expressiveness is not enough. For instance, references to past posts are just links to their HTML version.

[1] http://www.foaf-project.org/

[2] See GNOME Planet, http://planet.gnome.org/heads/

[3] Similar to the Debian Developer map at http://www.debian.org/devel/developers.loc

[4] http://www.doaml.net/

There are also RDF schemes to describe mail messages (without being specific for mailing lists), namely EMiR[5] and XMTP[6].

7 Conclusions

The mailing list archives can enter the Semantic Web just by using a suited application to retain as much information as possible using the proper language (RDF), as a complement to the HTML version. Deployment of such application would require little effort by mailing list administrators, so adoption could be fast[7]. It is not required any data enrichment by an expert, so huge volumes of information can be easily and quickly processed. This is an important factor because some mailing lists have a high activity and many years of existence.

Development of such application requires, first of all, the definition of an information scheme, probably a combination of the existing ones (see Section 6). Secondly, Mailbox files must be processed in order to extract the already-existent information. This second step can take as a starting point some of the current free software tools that generate HTML from a Mailbox.

[5] http://xmlns.filsa.org/emir/

[6] http://www.openhealth.org/xmtp/

[7] Actually, any subscriber (not just the administrator) would be able to use such application. The only requirement is to have access to all the prior posts, probably already stored in his own e-mail client.