Semantics-based Publication Management using RSS and FOAF

Peter Mika and **Michel Klein** and **Radu Serban**
Department of Computer Science, Vrije Universiteit Amsterdam
[pmika|mcaklein|serbanr]@cs.vu.nl

Abstract

Listing references to scientific publications on personal or group homepages is a common practice. Doing this in a consistent and structured manner either requires a lot of discipline or a centralized database. Scientific publication, however, is a distributed activity by nature. We present a completely distributed and RDF-based implementation for disseminating references to scientific publications. Our application only uses existing information sources and allows for different output formats, e.g. HTML, RSS and RDF.

1 Collecting and Publishing References

Information about scientific publications is often maintained by individual people. To present this distributed information in different selections and in different formats at different locations usually requires a lot of manual work. We demonstrate an application that performs this task using Semantic Web based techniques.

Our application collects several sources of information from several locations, in particular information about publications of authors from their homepage, information about group-membership from the department website and information about people by crawling FOAF-profiles. All sources are—if not yet in this format—translated to RDF and uploaded to an RDF store, in our case Sesame [Broekstra *et al.*, 2002].

In the repository we apply several unification and reasoning steps to link the different data sources, to derive additional facts and to remove redundant information. In addition, a separate web service can be used to query for publications based on specific criteria and to produce a variety of output formats, including BuRST (a compatible extension of RSS 1.0) and HTML.

Figure 1 presents a schematic representation of the approach, which is introduced in detail in the following section.

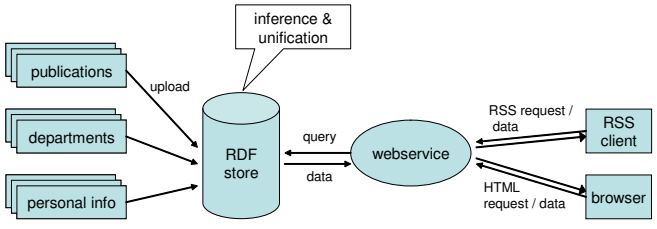

Figure 1: A schematic representation of the approach.

2 Sources of Information

For information about publications, we rely on the common BibTeX format. We ask authors to include a BibTeX file with their own publications on a publicly accessible part of their website. For many authors this does not require additional work, as they already maintain such a file themselves. A simple crawler collects all files from the `www.few.vu.nl` domain. The BibTeX files are translated to RDF using the BibTeX-2-RDF service,[1] which creates instance data for the "Semantic Web Research Community" (SWRC) ontology.[2]

Personal information is collected via the web as well, using the FOAF profiles [Brickley and Miller, 2005] that people linked from their homepage. The FOAF files contain RDF statements describing personal information such as the individual's homepage, workplace, image and relationships to other people.

To know which researchers are member of which department, we have implemented a web service that translates the content of the department mailing lists to a FOAF format with statements about group membership. We do not reveal the email addresses of people, but use a hash of the email address as identifier. By using the mailing lists as a source for the group membership information, we do not have to maintain this information ourselves, but rely on the existing infrastructure in the department (i.e. the computer system adminstration).

3 Aggregation

Mapping Schemas

Using distributed, web-based knowledge technologies, we have to deal with the arising semantic heterogeneity of our information sources. Heterogeneity effects both the schema and instance levels.

As the schemas used are stable, lightweight web ontologies, mappings on the class level cause little problem: such mappings are static and can be manually inserted into the knowledge base. An example of such a mapping is the subclass relationship between the `swrc:Person` and `foaf:Person` classes or the subproperty relationship between `swrc:name` and `foaf:name`.

Although we used existing RDF schemas for describing the instance data, a simple extension of the SWRC ontology was necessary to preserve the sequence of authors of publications. To this end we defined the `authorList` and `editorList` properties, which have `rdf:Seq` as range, comprising an ordered list of authors..

[1] See `http://www.cs.vu.nl/~mcaklein/bib2rdf/`.
[2] See `http://ontoware.org/projects/swrc/`.

Unifying Instances

Heterogeneity on the instance level arises from using different identifiers in the sources for denoting the same real world objects. This effects FOAF data (where typically each personal profile also contains partial descriptions of the friends), but also publication information, as the same author may be referenced in a number of BibTeX sources.

The solution is provided by instance reasoning (smushing) using ontological features. The FOAF ontology defines a number of inverse-functional properties of the Person class which can be used to determine whether two instances of Person are the same. (Functional properties, on the other hand, can be used to prove that two instances are not the same.) For example, if two Persons have the same value for the `mbox-sha1sum` (hash of the email address), we can conclude that both instances are the same. In this way, we can relate the statements from the FOAF files to the statements about the mailinglist-membership. Besides the inverse-functional properties, we also apply fuzzy string matching to compare person names, following a step of normalization (e.g. to be able to compare 'Harmelen, F.' and 'Frank van Harmelen'). Similarly, publications are matched based on an exact match of the date of the publication and a tight fuzzy match of the title. Matching publications based on author similarity is among the future work.

The matches that we find are recorded in the RDF store using the owl:sameAs property. Since Sesame doesn't natively support OWL semantics at the moment, we expanded the semantics of this single property using Sesame's custom rule language. These rules express the reflexive, symmetric and transitive nature of the property as well as the intended meaning, namely the equality of property values. The rules add several statements to give all the equivalent resources the same the set of properties. These rules are executed by the custom inferencer during uploads, which means that queries are fast to execute. (On the downside, the size of the repository greatly increases.)

4 Presentation

After the information has been merged, the triple store can be queried to produce publications lists according to a variety of criteria, including persons, groups and publication facets. An online form helps users to build such queries against the departmental publication repository. The queries are processed by another web-based component, the Publication webservice.

This tool takes the location of the repository, the query, the properties of the resulting RSS channel and optional style instructions as parameters. In a single step, it queries the repository and generates an RSS channel with the publications matching the query. This RSS channel follows the BuRST specification[3] for mixing in publication metadata into the RSS channel. The resulting channel appears as a RSS 1.0 channel for compatible tools while preserving RDF metadata.

The presentation service can also add XSL stylesheet information to the RSS feed, which allows to generate different HTML layouts (tables, short citation lists or longer descriptions with metadata). The HTML output can be viewed with any XSLT capable browser and it can be tailored even further by adding a custom CSS stylesheet.

[3] `http://www.cs.vu.nl/~pmika/research/burst/BuRST.html`

5 Use Cases

Our system for semantics-based bibliography management can be used by individuals and groups alike in a variety of modes. It can be used to provide a search interface to publication collections on personal homepages or departmental websites such as the homepages of the AI and BI groups of the VUA (information pull).

More interestingly, the use of RSS technology allows others to be notified of changes to these collections (information push) by subscribing to publication feeds. A number of generic tools are available for reading and aggregating RSS information, including browser extensions, online aggregators, news clients and desktop readers for a variety of platforms. While these software are not aware of the SWRC and FOAF schemas, they are still able to process BuRST feeds by ignoring the information they do not understand. (This behaviour is mandated by the RSS specification and is the basis of modularization in RSS.) Mozilla FireFox also natively supports RSS feeds as the basis for creating dynamic bookmark folders. These folders refresh their contents from an RSS feed whenever the user opens them.

The reliance on RDF and lightweight, widely used web ontologies also makes it possible to access personal profiles and publication information by generic RDF tools such as the Piggy Bank browser extension. Piggy Bank allows users to collect RDF statements linked to Web pages while browsing through the Web and to save them for later use. FOAF information can be processed by a growing number of tools, while the SWRC data can be easily converted back to BibTeX to complete the knowledge cycle.

6 Discussion

In summary, we presented a semantic-based system for publication management that builds on web technology, well-known ontologies and by reusing existing information requires no additional effort from the individual. In comparison to centralized approaches, our system leaves the control over publication management and presentation in the hands of the individual researcher, while still allowing for information push. On the other hand, our system is more lightweight than P2P networks that require users to install and run specific software on their computers. The Java object models for the FOAF, RSS and BuRST formats as well as the tools for crawling and smushing FOAF data have been made available as part of the open source Elmo API for Sesame. Elmo can be downloaded from `www.openrdf.org`. The interface to the tools themselves and some examples can be found at `http://prauw.cs.vu.nl:8080/burst/`.

References

[Brickley and Miller, 2005] Dan Brickley and Libby Miller. FOAF vocabulary specification. Namespace document, June 3, 2005.

[Broekstra *et al.*, 2002] Jeen Broekstra, Arjohn Kampman, and Frank van Harmelen. Sesame: An architecture for storing and querying RDF and RDF Schema. In Ian Horrocks and James A. Hendler, editors, *Proceedings of the First International Semantic Web Conference (ISWC 2002)*, volume 2342 of *Lecture Notes in Computer Science*, pages 54–68, Sardinia, Italy, June, 9–12, 2002. Springer-Verlag.