

A Publication Aggregation System Using Semantic Blogging

Aman Shakya¹, Hideaki Takeda², Ikki Ohmukai², Vilas Wuwongse¹

¹ Asian Institute of Technology, Klong Luang, Pathumthani
12120, Thailand

{aman.shakya@ait.ac.th, vw@cs.ait.ac.th}

² National Institute of Informatics, 2-1-2 Hitotsubashi, Chiyoda-ku,
Tokyo, 101-8430, Japan
{takeda, i2k}@nii.ac.jp

Abstract. Researchers need to share information about their publications. They also desire to share opinions and comments about each other's publications. The paper describes a system which demonstrates how semantic blogging can be used for the purpose. The SWRC ontology has been incorporated in the blogging system for entering the metadata of publications. The semantic blogging system provides means to annotate publications. The system is a decentralized publication aggregation system. It utilizes the RSS technology to aggregate posts. The publication metadata is embedded in the RSS to produce BuRST feed. The RSS aggregator has been extended to handle the BuRST feeds. The system uses the FOAF links of the authors and friends to explore the social network of the research community for aggregation of relevant information.

1 Introduction

In research communities there is always need of posting and sharing information about publications. The bibliographic information about publications should follow some standard metadata schema for being machine processable and interoperable. SWRC (Semantic Web for Research Communities) [1] is an ontology for modeling entities of research communities such as persons, organizations, publications (bibliographic metadata) and their relationships. Besides posting structured data about publications, there is often need of posting comments or annotations about the publications. Blogging can be a mechanism for such purposes. There is need of sharing information in the community in a decentralized way. It is not feasible for researchers worldwide to be bound to a single centralized information system. Publications from various relevant sources need to be aggregated. The researcher should be able to search required publication information from the postings of the community. The paper describes a publication aggregation system for research communities using semantic blogging.

2 Semantic Blogging

Blog, a short form for weblog, is a publicly accessible web-based publication of periodic articles, usually in reverse chronological order that often serves as a personal journal. Blogging makes publishing information on the web very easy, unlike web sites which involve much more elaborate plan and technical effort. It can be a powerful tool for establishing and maintaining an online community [2,3]. Blogs are user-oriented providing personal spaces for users on the web. It motivates higher rate of timely information publication. However, filtering, organizing and navigating through the blogosphere are a challenge in traditional blogging [2].

Semantic blogging is a technology that builds upon blogging and adds semantic structure to the blog items. The blog items are enriched with metadata. Semantic blogging uses desirable features of both blogging and the Semantic Web to deal with the challenges of traditional blogging. The semantic web is well suited for incrementally publishing structured and semantically rich information. On the other hand, the easy publishing nature of blogging can boost the semantic web by publishing enough data and resources. Semantic blogging can extend blogging from simple diary browsing to informal knowledge management [2,3].

2.1 Semantic Blogging for Research Communities

Traditionally blog entries are just unstructured passages of text. However, for publishing information such as about research publications, there is need of some structure. Semantic blogging provides this. Research publications can be posted described by standard metadata.

Often blog entries are written about some publication or resource on the web. Semantic blogging can provide a way to write blog entries as annotations or comments to other blog entries or publications.

Blogs provide RSS feeds which are understood by RSS readers. The metadata in semantic blogs can be embedded in the RSS feeds. If a standard vocabulary is used for the metadata, RSS readers can be extended to process the metadata as well. Ontologies such as SWRC [1] exist for the research community. Semantic blogging with RSS aggregation forms a powerful web-based decentralized system. Semantic blogging can offer an easy, lightweight and flexible mechanism for information publishing, sharing and aggregation for research communities.

Further blogs provide effective mechanisms to foster communities. Blogrolls list the friends and associates of the blog-owner and connects his/her blog to theirs. Moreover, hyperlinks, annotations, comments and trackbacks also serve to tightly link elements in the community.

3 Relevant Work

A semantics based publication management system using RSS and FOAF has been described in [4]. The application collects information in BibTex format from several locations and information about people by crawling FOAF files. All the information is stored in RDF format in a central store. The SWRC ontology has been used. The shortcoming of the system is that it is centralized. We have to access the central unified store or the web service provided. Besides, the authors still have to write the BibTex files which is harvested by the system.

Bibster [5] is a peer-to-peer application for sharing bibliographic metadata among researchers. Bibster also uses SWRC as the application ontology and uses the ACM topic hierarchy as the domain ontology for classification of metadata entries. Bibster is decentralized and doesn't require the authors to produce BibTex files. However, the system is limited to a peer-to-peer network. Blogs make a wide community on the web instead of being restricted to a peer network. Like peer-to-peer systems blogging can also be decentralized. Moreover, blogging supports commenting.

The Semantic Blogging Demonstrator¹ developed by HP labs as a part of the Semantic Web Advanced Development-Europe (SWAD-E) project also uses the bibliographic domain [2,3]. The blog entries and bibliographic items have been modeled as different things. The fact that the blog entry annotates the bibliographic item has been modeled by a 'contains' property. Blog entries *contain* bibliographic items [3]. The demonstrator offers semantic navigation, view and query capabilities. The system also provides a metadata import facility - SemBlogIT! It provides a way to attach items to blog entries and a category chooser functionality also. The implementation of the RDF store is a single file containing metadata for all blog items. The system is centralized and doesn't offer cross-blog data aggregation.

The semantic blogging system described in [6] is based on egocentric methods. The closeness between the contents is determined by the distance between authors on the human network. The project extends the concept of RSS to describe metadata like inter-site relation. The metadata has been called "RDF Content Summary (RCS)". RCS uses additional modules for ordinary hyperlinks and trackbacks. The personal publishing suite [7] consists of two software - "RNA" and "glucose". RNA is a web application which subscribes RSS files and builds a site tree based on RDF. Glucose is a standalone RSS aggregator.

BuRST. BuRST (Bibliography Management using RSS Technology) [8] is a lightweight specification for publishing bibliographic information using RSS 1.0 and bibliography-related metadata standards. The specification defines guidelines about how to use existing vocabularies instead of defining additional vocabulary. An RSS module for bibliographic items has been defined using the SWRC and FOAF ontologies. The SWRC ontology should be used to provide metadata about the publication. It is recommended to use the FOAF vocabulary to provide detailed descriptions of persons related to publications.

¹ <http://www.semanticblogging.org/semblog/blog/default/>

4 Requirements and Design

The requirements of the system are based on the needs of research community as pointed out in the introduction. These can be summarized as below.

- Users should be able to post blog entries and metadata about publications easily from anywhere on the web.
- It should be possible to post blog entries as comments or annotations to other publications or entries.
- The system should be fully decentralized as ordinary blogging systems.
- The system should be able to aggregate relevant information from multiple blogs in the community

To meet the above requirements we have the following design policies.

Integration of RDF metadata store with blogs. The existing blogging infrastructure needs to be supplemented by an RDF metadata store following a standard metadata schema.

Use of trackbacks and annotations. Trackback facility offered by blogs can be utilized for commenting on posts. Standard blogging tools would be able to trace the trackback to determine the commenting blog entry. Trackbacks provide backward link to the commented post. Additionally, an annotation mechanism could be developed to comment in the forward direction.

Extended RSS. RSS mechanism is well established for the decentralized operation of blogs. RSS can be extended to incorporate the publication related metadata. Maintaining compatibility with RSS would allow the system to reap the benefits of RSS and have interoperability with blogging systems.

Powerful aggregation. Aggregation plays an important role for a decentralized system like this. Existing aggregation tools can be adapted for the purpose. Aggregation for the research community can be realized using FOAF links as FOAF is a popular mechanism for forming communities. Further searching the aggregate can be done by parsing the RDF metadata.

5 System Architecture

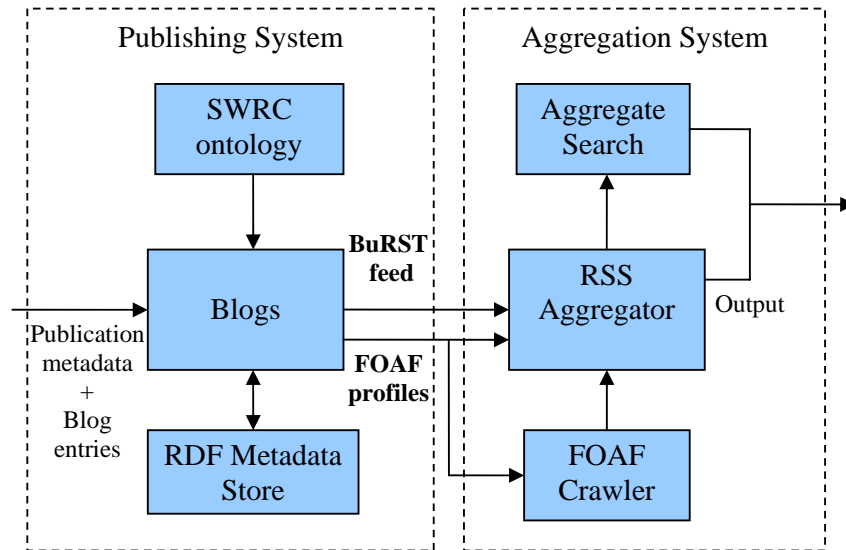


Fig. 1. System architecture of Semantic Blogging Publication Aggregation System

Fig. 1 shows the architecture of the developed system. The architecture consists of two sub-systems – Publication system and Aggregation system.

The *publication system* facilitates the user to publish blog entries and metadata about publications. The blogs are enriched semantically by using the SWRC ontology for metadata about publications. The metadata is stored in an RDF metadata store connected to the semantic blog. BuRST feeds publish the blog contents in interoperable and machine processable form. The publication system also helps to publish FOAF profiles of the blog-owners.

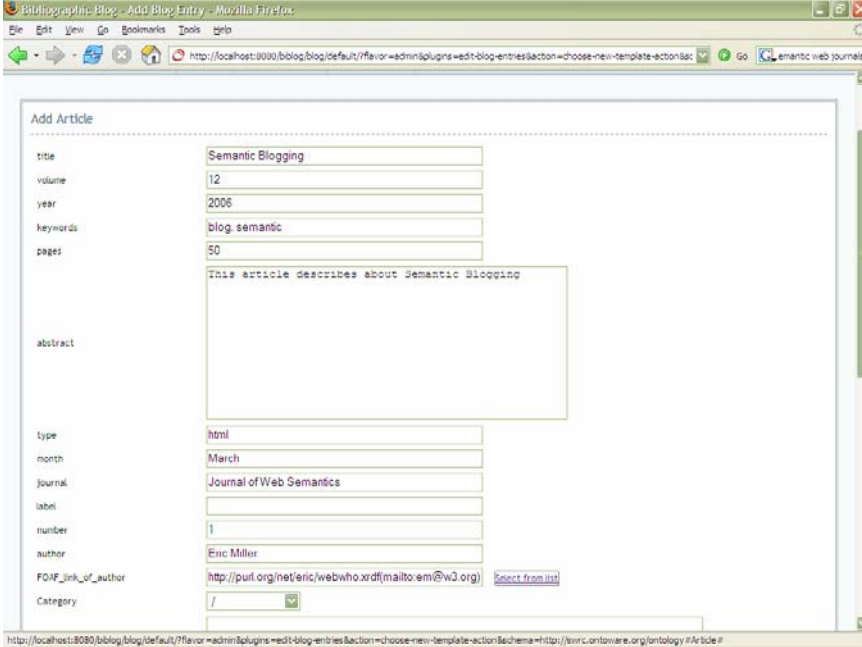
The decentralized *aggregation system* uses RSS technology to aggregate publication data from multiple blogs. FOAF profiles and crawled FOAF links are used to determine appropriate feeds to be aggregated. The aggregated output is shown on the blog interface itself. The aggregate search allows the user to search for publications and posts by various criteria.

5.1 Publication Metadata and Blog Entry

The publication Metadata follows the SWRC ontology. The semantic blog developed provides the entry forms for the metadata of different SWRC publications. Fig. 2 shows an example of the metadata entry interface for SWRC article publication type. The metadata schema has been taken from SWRC expressed in OWL. Additionally,

few elements have been added to the schema file to control the GUI of the forms. For instance, input control for the abstract should be a text area of enough size. The system provides a lightweight metadata interface to the users, unlike sophisticated interfaces of ontology editors.

Blogging has been made convenient by employing javascript bookmarklets. The bookmarklet has to be saved in the bookmarks list of the browser. If the user wants to blog a web page or a blog entry, he simply clicks the bookmarklet link from the browser. The bookmarklet captures the title, URL, and trackback ping URL of the current blog entry which is being annotated. The form for the new blog entry is then automatically populated with these items as shown in fig. 3. The trackback ping URL is discovered if the blog embeds the information about the trackback ping URL in an RDF snippet as specified in the trackback technical specification by Six Apart [9]. Though some blogging systems provide similar mechanism of easy data import, they don't extract the trackback ping URL from the page.



The screenshot shows a web browser window titled "Bibliographia: Blog - Add Blog Entry - Mozilla Firefox". The address bar shows the URL: "http://localhost:8080/biblog/blog/default/?flavor=admin&plugins=edit-blog-entries&action=choose-new-template-action&schema=http://swrc.ontoware.org/ontology#Article#". The main content area displays a form titled "Add Article" with the following fields and values:

Field	Value
title	Semantic Blogging
volume	12
year	2006
keywords	blog semantic
pages	50
abstract	This article describes about Semantic Blogging
type	html
month	March
journal	Journal of Web Semantics
label	
number	1
author	Eric Miller
FOAF_link_of_author	http://purl.org/net/eric/webwho.xrd/mailto:em@w3.org Select from list
Category	/

Fig. 2. Metadata entry interface for SWRC article

Blog entries as annotations. The bookmarklet also captures the URL of the current page so that the user can annotate the URL. The blog entry would be about the URL, possibly a publication. A trackback ping is also sent to the annotated blog entry if desired. Traditionally, blogs support trackbacks only and they just act as one way links. If user A writes a blog entry about a post by B, B receives a trackback from A. B and the readers of B's blog know that A has written something related. However, readers of A's blog do not know that it is related to B's blog. An *annotates* element has been introduced in the RDF metadata to model this action of annotation. Fig. 3 shows the normal blog entry interface where the *annotates* field can be seen. The trackback URL is also seen populated in the figure. The annotation is manifested as a link in the blog page as shown in fig. 4. Clicking on the link navigates the user to the annotated blog entry or resource. Fig. 4 shows both a normal blog entry and a publication.

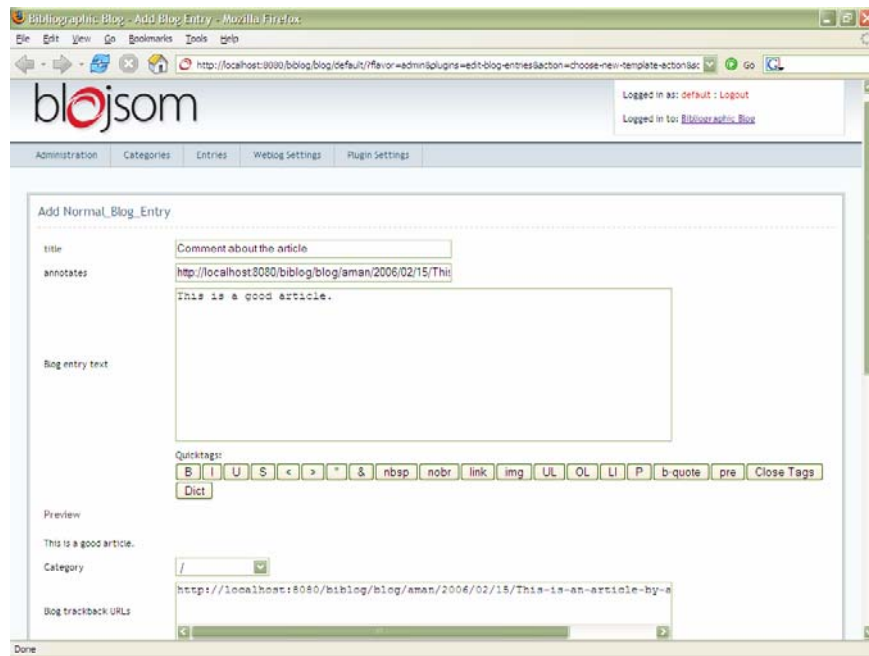


Fig. 3. Normal blog entry interface

5.2 Blogs

Blojsom² has been used as the blogging platform. It is a Java based open source software which can host multiple blogs for different users. However, the blogs in the community may be based on any system provided that they produce BuRST feeds [8]. But BuRST feeds are needed only for publication metadata. In fact, the system can aggregate information from any RSS feed.

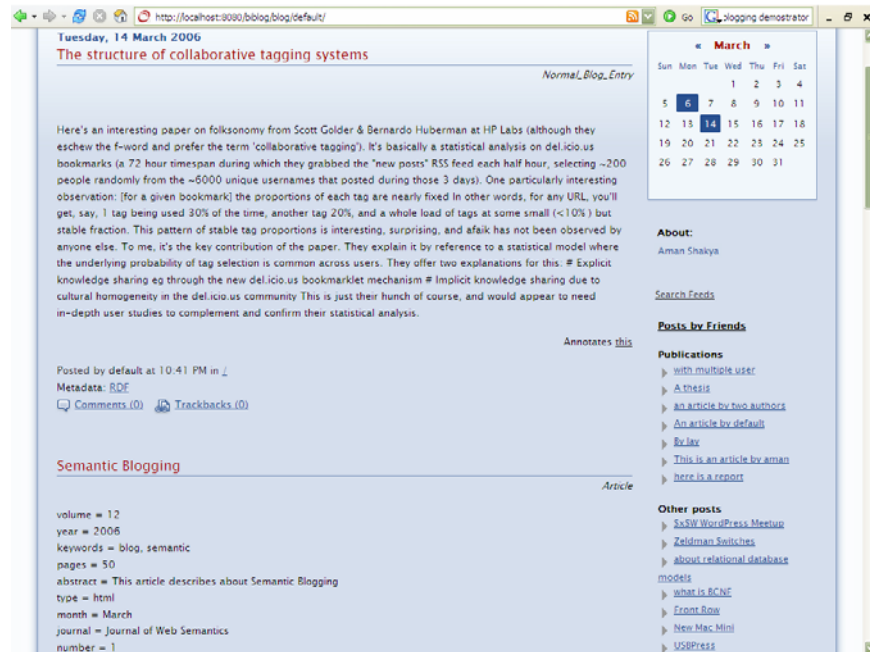


Fig. 4. The blog interface

5.3 RDF Metadata Store

The metadata about the publications are stored in RDF format in a MySQL database with the help of Jena³. Jena is a semantic web framework which can be used to manipulate RDF and OWL. The use of database for RDF storage can address the problem of scalability which may occur in systems using a single text file to store all RDF metadata as in [2,3]. The data is stored in the form of RDF triplets. The RDF metadata points to the concerned blog entry using the permalink of the entry. The

² <http://blojsom.sourceforge.net>

³ <http://www.hpl.hp.com/semweb/jena.htm>

RDF metadata is modeled as description about the permalink. The RDF snippet shown below is an example metadata for the blog entry at <http://localhost:8080/biblog/blog/default/?permalink=Semantic-Blogging.html>

An RDF snippet showing the metadata storage

```
<rdf:Description
rdf:about="http://localhost:8080/biblog/blog/default/?permalink=
Semantic-Blogging.html">
  <swrc:volume>12</swrc:volume>
  <swrc:keywords>blog, semantic</swrc:keywords>
  <swrc:journal>Journal of Web Semantics</swrc:journal>
  <swrc:author>Eric Miller </swrc:author>
  <swrc:pages>50</swrc:pages>
  <swrc:number>1</swrc:number>
  <swrc:abstract>
    This article describes about Semantic Blogging
  </swrc:abstract>
  <semblog:FOAF_link_of_author>
    http://purl.org/net/eric/webwho.xrdf(mailto:em@w3.org)
  </semblog:FOAF_link_of_author>
  <swrc:type>html</swrc:type>
  <swrc:year>2006</swrc:year>
  <swrc:month>March</swrc:month>
</rdf:Description>
```

The blog text is handled by the blogging infrastructure. Blojsom saves the blog text in text files and assigns a permalink to the blog entry which uniquely identifies the blog entry.

5.4 FOAF Profiles

A web-based interface to maintain the FOAF profile of the blog-owner has been provided in the blog itself. The information about the blog-owner and his friends is maintained in a simple readable text file which can be easily understood and updated by the user. Values from the XFN⁴ profile are used to define the relation of the blog-owner with the persons listed. These relations are mapped into FOAF vocabulary one-to-one. The profile includes the FOAF links of the friends of the blog owner and optionally their RSS feed URLs. The system publishes the profile on the web in FOAF syntax. The entries in the FOAF profile are utilized to aggregate RSS feeds as described later in section 5.6. On the other hand, the list of friends maintained in the profile also serves as a blogroll for the blog.

5.5 FOAF Crawler

The Elmo scutter⁵ has been used as a FOAF crawler to find out the FOAF link of authors of publications in the research community. The Elmo scutter is a generic RDF

⁴ <http://gmpg.org/xfn/>

⁵ <http://www.openrdf.org/doc/elmo/users/index.html>

crawler that follows `rdfs:seeAlso` links in RDF documents. RDF(S) `seeAlso` is also the mechanism used to connect FOAF profiles and thus the scutter allows to collect FOAF profiles from the Web. The crawler starts crawling from the FOAF profile of the blog owner and traces the `rdfs:seeAlso` links for FOAF links. The `rdfs:seeAlso` link is assumed to be a FOAF link if it is an immediate child to a `foaf:Person` element. The scutter creates a database of FOAF links of the interlinked community.

While uploading information about publications, the user enters the FOAF links of the authors also. This can be seen in the input form of fig. 2. To support the user for data entry, the crawled FOAF database may be searched to find out FOAF links of authors in the community. A person is identified by his email address or the `mboxsha1sum` (hash of the email address) along with the FOAF link. Elmo also provides a *smusher* that merges multiple FOAF profiles of the same person.

5.6 RSS Aggregator

The blogs generate RSS feeds according to the BuRST specification [8]. Fig. 5 shows a part of the BuRST feed generated. Metadata about publications are expressed as SWRC publication elements in the feed. The authors are embedded as FOAF elements. These feeds are aggregated by other users on their blogs.

The latest publications and posts by friends of the blog-owner are downloaded through the feeds and displayed alongside in the blog as shown in fig. 4. Generally, RSS aggregators are separate standalone applications. However, the system integrates the RSS aggregator in the blog interface itself. The blog readers have the advantage of knowing about latest publications of other authors also, relevant to the blog entries. Further, when a blog entry for a publication is opened, the RSS feeds of the authors of the publication are downloaded and shown alongside. Thus, relevant posts by the authors are shown alongside each publication entry. For each aggregated entry, the title followed by a collapsible view of the description is shown. For publications, the description is generally the abstract. The title has a hyperlink to the concerned blog entry.

The RSS aggregator makes subscriptions for RSS feeds from the friends' list of the blog. The FOAF profile, as mentioned in section 5.4, includes FOAF links and RSS feed URLs of the friends. The RSS feed URL is included in the `rss:channel` element in the weblog description as shown in the FOAF snippet below. The `rdf:seeAlso` element shows the FOAF link of the person.

An example FOAF snippet

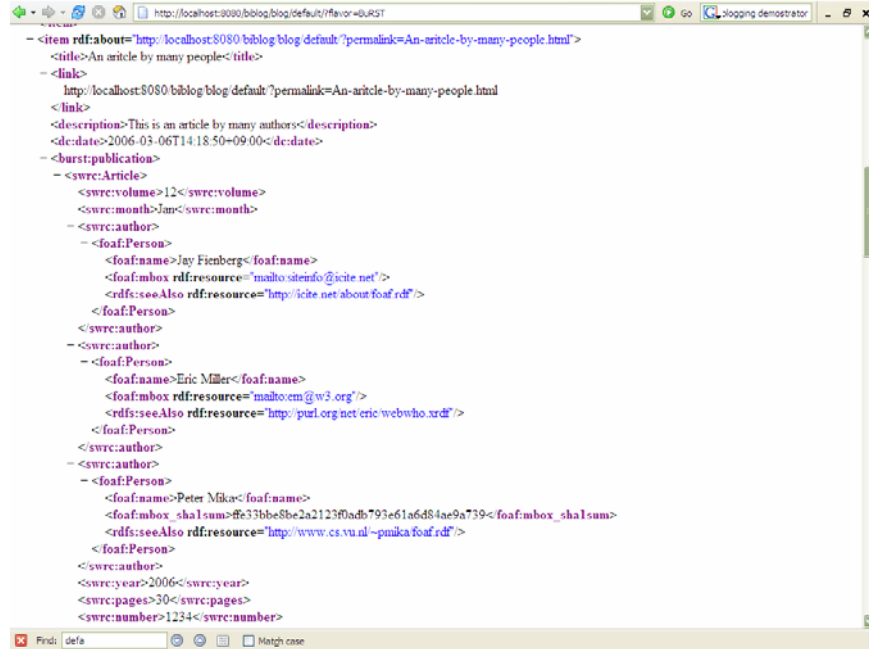
```
<Person rdf:nodeID="n5">
<name>Morten Frederiksen</name>
<weblog>
  <Document rdf:about="http://wasab.dk/morten/blog/">
    <dc:title>Binary Relations</dc:title>
    <rdfs:seeAlso>
      <rss:channel
rdf:about="http://www.wasab.dk/morten/blog/feed/rdf"/>
        </rdfs:seeAlso>
      </Document>
    </weblog>
  <rdfs:seeAlso
rdf:resource="http://www.wasab.dk/morten/blog/archives/au
thor/morten/foaf.rdf"/>
</Person>
```

When a single publication entry is opened, the RSS feeds of the authors are only aggregated. The RSS feed of the author is determined through the FOAF link of the author. The FOAF links of authors are entered while posting the publication on the blog as described in section 5.5.

Flock⁶ RSS aggregator, a java based open source, has been used as the RSS aggregator. Flock uses subscriptions represented in OPML⁷ syntax. The appropriate OPML is produced as the subscription list by our system. Flock uses this to aggregate feeds and displays the result in the blog. The aggregator downloads the feeds at regular intervals to keep contents up-to-date. The refresh interval can be configured as a setting. Hence, the latest posts are shown in the aggregated list. The Flock aggregator has been extended to handle BuRST feeds. If the RSS feeds are in BuRST format, publications can be distinguished from other posts and SWRC metadata can be extracted. The system separates the list of publications and non-publications. Besides BuRST, the system also supports RSS 1.0, RSS 0.91 and RSS 2.0.

⁶ <http://flock.sourceforge.net>

⁷ <http://www.opml.org/>



```
- <item rdf:about="http://localhost:8080/bblog/blog/default/?permalink=An-article-by-many-people.html">
  <title>An article by many people</title>
  <link>
    http://localhost:8080/bblog/blog/default/?permalink=An-article-by-many-people.html
  </link>
  <description>This is an article by many authors</description>
  <dc:date>2006-03-06T14:18:50+09:00</dc:date>
  <burst:publication>
  <swrc:Article>
    <swrc:volume>12</swrc:volume>
    <swrc:month>Jan</swrc:month>
  <swrc:author>
    <foaf:Person>
      <foaf:name>Jay Fienberg</foaf:name>
      <foaf:mbox rdf:resource="mailto:siteinfo@jicite.net"/>
      <rdfs:seeAlso rdf:resource="http://jicite.net/about/foaf.rdf"/>
    </foaf:Person>
    <swrc:author>
    <foaf:Person>
      <foaf:name>Eric Miller</foaf:name>
      <foaf:mbox rdf:resource="mailto:em@w3.org"/>
      <rdfs:seeAlso rdf:resource="http://purl.org/net/eric/webwho.xrdf"/>
    </foaf:Person>
    <swrc:author>
  <swrc:author>
    <foaf:Person>
      <foaf:name>Peter Mika</foaf:name>
      <foaf:mbox_sha1sum>#e33bbe8be2a2123f0adb793e61a6d84ae9a739</foaf:mbox_sha1sum>
      <rdfs:seeAlso rdf:resource="http://www.cs.vu.nl/~pmika/foaf.rdf"/>
    </foaf:Person>
    <swrc:author>
  <swrc:year>2006</swrc:year>
  <swrc:pages>30</swrc:pages>
  <swrc:number>1234</swrc:number>
```

Fig. 5. BuRST feed

5.7 Aggregate Search

The aggregated BuRST feeds can be searched and sorted by SWRC fields like title, author, type, etc as shown in fig. 6. A web based interface has been provided for the search. The user can specify the search criteria by entering values in different metadata fields. If values for multiple fields are specified, logical And operation is done on the conditions. Only a limited number of attributes have been implemented for demonstration purpose. However, it would be quite straightforward to extend the search interface into a more complete one. The search results can be sorted in ascending or descending order by clicking on the headers of different fields. Search is done by filtering the aggregated RSS feeds according to the search criteria provided through the web-based form.

The aggregate search is different from the search provided by the blogging system. The normal search provided by the blogging system does a text search over the entries of the particular blog. The aggregate search does a metadata based search over multiple blogs. The feature would be useful to search out desired publications from a large aggregated collection of RSS feeds.

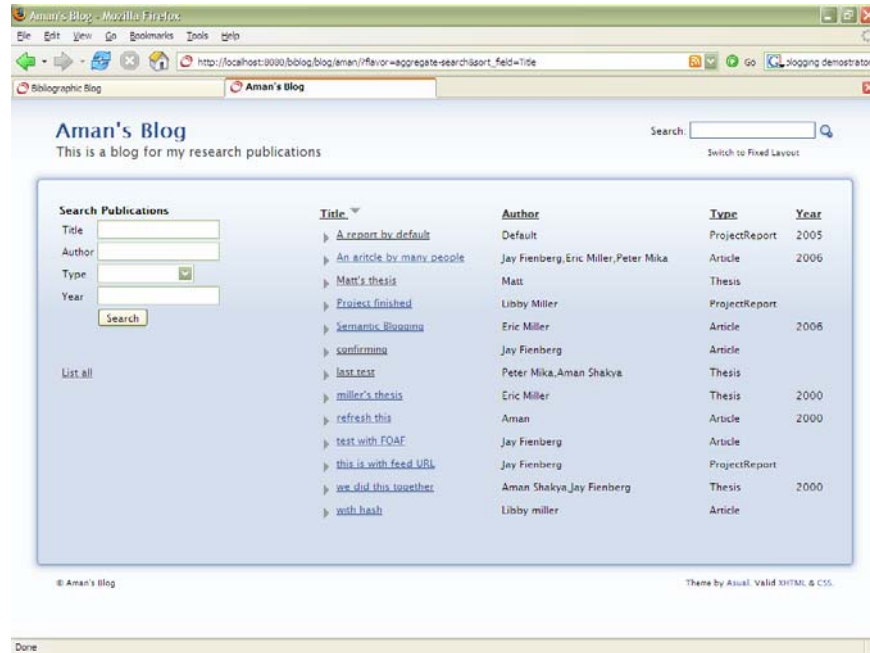


Fig. 6. Aggregate search of publications

6 Discussion

The work serves the need of research communities to share information about publications and comment on these. Current systems like the publication management system [4] and the semantic blogging demonstrator [2,3] are centralized. Peer-to-peer systems like Bibster [5] are limited to a peer network and moreover don't provide the facility of commenting. The presented work tries to combine the functionalities of publication management systems and semantic blogging systems. Semantic blogging is suitable for the need of posting both structured bibliographic metadata and unstructured comments or annotations. Further, RSS aggregation mechanism helps to pool information from relevant sources and present them on the blog. The presented work is an attempt to develop a decentralized aggregation system based on semantic blogging. The system tries to aggregate relevant posts and publications from the social network by tracing the FOAF links of co-authors and friends of researchers. Thus, the system also strengthens the sense of community.

7 Future work

The domain ontology, such as a topic hierarchy, could be incorporated in the blogging system as a future work. The ontology would help to categorize the blog entries and publications. Basically, only the publication class has been used from the SWRC. The full features of the ontology could be utilized in the future. Further, inference could be employed based on the ontology to deduce useful semantic relations. Another future work could be collecting information by harvesting of widely existing BibTex files.

Acknowledgement

The authors would like to thank the Asian Institute of Technology, Thailand and the National Institute of Informatics, Japan to arrange the opportunity for the research. The support and guidance from professors and colleagues in both the institutes were highly valuable for conducting the research.

References

1. Sure, Y., Bloehdorn, S., Haase, P., Hartmann, J., Oberle, D.: The SWRC Ontology - Semantic Web for Research Communities. In: Bento, C., Cardoso, A., Dias, G. (eds.): Proceedings of the 12th Portuguese Conference on Artificial Intelligence - Progress in Artificial Intelligence (EPIA 2005), Vol. 3803. Springer, Covilha Portugal (2005) 218 – 231
2. Cayzer, S.: Semantic Blogging and Decentralized Knowledge Management. Communications of the ACM, Vol. 47 (2004) 48-52.
<http://portal.acm.org/citation.cfm?id=1035164&coll=GUIDE&dl=ACM&CFID=49229987&CFTOKEN=13649580&ret=1#Fulltext>
3. Cayzer, S.: Semantic Blogging: Spreading the Semantic Web Meme.
<http://citeseer.ist.psu.edu/698724.html>
4. Mika, P., Klein, M., Serban, R.: Semantics-based Publication Management using RSS and FOAF. In: Proceedings of the 1st Workshop on the Semantic Desktop (SD 2005), 4th International Semantic Web Conference, Galway, Ireland (2005)
5. Haase, P., Schnizler, B., Broekstra, J., Ehrig, M., Harmelen, F., Menken, M., Mika, P., Plechawski, M., Pyszlak, P., Siebes, R., Staab, S., Tempich, C.: Bibster - A Semantics-Based Bibliographic Peer-to-Peer System. In: Proceedings of the International Semantic Web Conference (ISWC2004), Hiroshima, Japan (2004)
<http://bibster.semanticweb.org/publications/publications.htm>
6. Ohmukai, I., Numa, K., Takeda, H.: Egocentric Search Method for Authoring Support in Semantic Weblog (2003). <http://citeseer.ist.psu.edu/ohmukai03egocentric.html>
7. Ohmukai, I., Takeda, H.: Semblog: Personal Knowledge Publishing Suite. In: Proceedings of WWW 2004 Workshop on the Weblogging Ecosystem: Aggregation, Analysis and Dynamics, New York, USA (2004)
<http://www-kasm.nii.ac.jp/papers/takeda/04/www2004ohmukai.pdf>
8. Mika, P.: Bibliography Management using RSS Technology (BuRST) (2005)
<http://www.cs.vu.nl/~pmika/research/burst/BuRST.html>
9. Six Apart : Developer Documentation : TrackBack Technical Specification.
http://www.sixapart.com/pronet/docs/trackback_spec