SOCIOBIBLOG: A DECENTRALIZED PLATFORM FOR SHARING BIBLIOGRAPHIC INFORMATION

Aman Shakya, Hideaki Takeda
National Institute of Informatics
2-1-2 Hitotsubashi, Chiyoda-ku, Tokyo, 101-8430, Japan
shakya_aman@nii.ac.jp

Vilas Wuwongse
Asian Institute of Technology
Klong Luang, Pathumthani 12120, Thailand
vw@cs.ait.ac.th

Ikki Ohmukai
National Institute of Informatics
2-1-2 Hitotsubashi, Chiyoda-ku, Tokyo, 101-8430, Japan
shakya_aman@nii.ac.jp

ABSTRACT

Sharing of bibliographic information is very important in a research community. SocioBiblog is a semantic blogging system that provides a decentralized environment to share bibliographic information. It uses the SWRC ontology for adding metadata about publications in blogs and aggregates publications from the social network neighborhood and co-authors of the researcher. RSS aggregation has been extended to handle embedded publication metadata in BuRST feeds. The FOAF network of the researchers is crawled to gather FOAF profiles for discovering aggregation feeds. Interoperability with other systems has been maintained by adopting standard formats. The aggregated collections may be searched and filtered flexibly by defining metadata criteria. The aggregated and filtered results can be redistributed as new feeds which can further be used by other systems. Thus, a decentralized ecosystem can be formed where each unit can publish, aggregate and redistribute information.

KEYWORDS

Decentralized information sharing, semantic blog, bibliographic metadata, RSS aggregation, BuRST.

1. INTRODUCTION

We live in highly interlinked communities and exchange information with each other. Communities of researchers are good examples of online communities. Researchers from different parts of the world know each other by working on common area, sharing ideas, co-authoring papers and reading each other's publications. Researchers need to publish information about their publications, share bookmarks about interesting publications and write their views on them. They should be able comment about each other's publications. Moreover, researchers always need to be up-to-date about latest publications in their fields. Useful information about publications can be aggregated from the community and we should be able to filter such aggregated collections and search information according to our requirements. Integrating various information sources and filtering them based on information needs can help in creating more useful tailored information sources.

Research activities are highly autonomous with numerous research organizations spread worldwide using different information systems. Thus, we need a decentralized mechanism for information sharing in research communities. Interoperability and compatibility are crucial in such a decentralized scenario.

Blogging is a popular mechanism for easy publishing on the web. Blogs provide commenting mechanisms and often provide RSS feeds which can be used for aggregation of posts from different sources. However, bibliographic information needs to be structured in a standard way. The Semantic Web (Berners-Lee et al. 2001) offers a decentralized platform to publish structured information. Semantic blogging combines the features of both blogging and the semantic web (Cayzer 2004a, b).

This paper describes SocioBiblog[1], a prototype semantic blogging system enabling researchers to share bibliographic metadata. The system can aggregate information about publications from different sources and filter such aggregated collections. We will start with a brief introduction to semantic blogging in Section 2 and discuss a decentralized scheme for aggregation. Our contributions will be discussed in Section 3 as requirements. System design and implementation will be detailed in Section 4. We will discuss related works and their limitations to meet our requirements in Section 5. Finally we will conclude in Section 6 pointing out to some future works.

2. SEMANTIC BLOGGING AND AGGREGATION

2.1 The Concept of Semantic Blogging

A blog is a publicly accessible web-based periodic publication, usually in reverse chronological order that often serves as a personal journal. Blogging makes publishing information on the web very easy. Blogs also foster communities with methods like blogrolling and enable commenting on each other's posts. However, filtering, organizing and navigating through the blogosphere are still difficult. Semantic blogging is a technology that builds upon blogging and enriches blog items with metadata (Cayzer 2004a, b). For publishing information such as research publications, there is need of some structure and semantic blogging provides this. Ontologies such as SWRC exist for the research community which can be utilized for semantic blogging. SWRC (Semantic Web for Research Communities) is an ontology for modeling entities of research communities such as persons, organizations, publications and their relationships (Sure et al. 2005).

2.2 Decentralized Aggregation

RSS aggregators are usually centralized online services or standalone desktop applications. A centralized system has to process huge volumes of information and risks being a single point of failure. Moreover, this may not address personal information needs. On the other hand, standalone aggregators cannot share aggregated information online. We propose aggregating feeds on personal blogs. This would be a decentralized solution offering personalized aggregation and facilitate online sharing of aggregated contents.

2.2.1 Social Network based Aggregation

Researchers working in common area may be connected by social network links or co-authorship of publications. Closely related people have similar interests and are more eager to communicate and share resources. We may expect to aggregate highly relevant resources from directly linked people and possibly up to the next level of links. Collecting information from the network neighborhood and redistributing through one's blog facilitates flow of information in the linked community.

2.2.2 Information Source Integration and Metadata-based Filtering

We need to aggregate information from multiple distributed sources and integrate these in a homogenous way. The information sources may be feeds from socially linked people or other useful systems. Integration of information from different systems becomes possible if a common semantic standard is followed. Furthermore, such an aggregated collection can be filtered by metadata to meet our information requirements. Semantic structure provides better fine grained control over information than using simple text.

[1] An online demo is available at http://dutar.ex.nii.ac.jp/

Selection of appropriate information sources and filtering aggregated information can be done to suit personal needs and new customized information sources can be constructed. These information sources can further be integrated with other information sources. Integration and interaction of components in the network can result in the emergence of new powerful systems delivering useful streams of information.

2.3 An Example Scenario

Fig.1 illustrates an example scenario. A researcher, 'A', publishes information about his publications on his semantic blog. He may enter metadata about his publication and upload the text file. Another researcher 'B' has some comments about the publication 'X' and writes them on his own blog. The metadata of publication 'X' is quoted in B's entry which points to the original entry by 'A'. A trackback ping is also sent which appears as a link on A's blog entry. The researcher may also bookmark interesting publications from popular bibliographic sites and write his views on them. The BibTeX metadata would be scraped from the original site and quoted in the blog entry.

The researcher 'A' can list his friends and other researchers he knows in his blogroll. In the example, 'A' knows 'B', 'C' and 'D'. SocioBiblog aggregates RSS or BuRST feeds from the sources in his blogroll. BuRST (Bibliography Management using RSS Technology) is a lightweight specification for publishing bibliographic information using RSS 1.0 and bibliography-related metadata standards (Mika et al. 2005; Mika 2005). Further, feeds from friends of a friend are also aggregated. For instance, 'C' knows 'E', so feeds from 'E' are also aggregated in A's blog. 'A' may obtain interesting information from 'E' even if he doesn't know him directly. When any publication entry is opened, latest posts by the co-authors of the publication are also listed. The publication 'X' has authors 'A', 'P' and 'Q'. While viewing 'X', other posts by the co-authors 'P' and 'Q' are also listed. The co-authors of a publication are usually highly related people working on the same field. Their publications would most likely be related to the particular publication or topic and thus could be highly relevant to the reader.

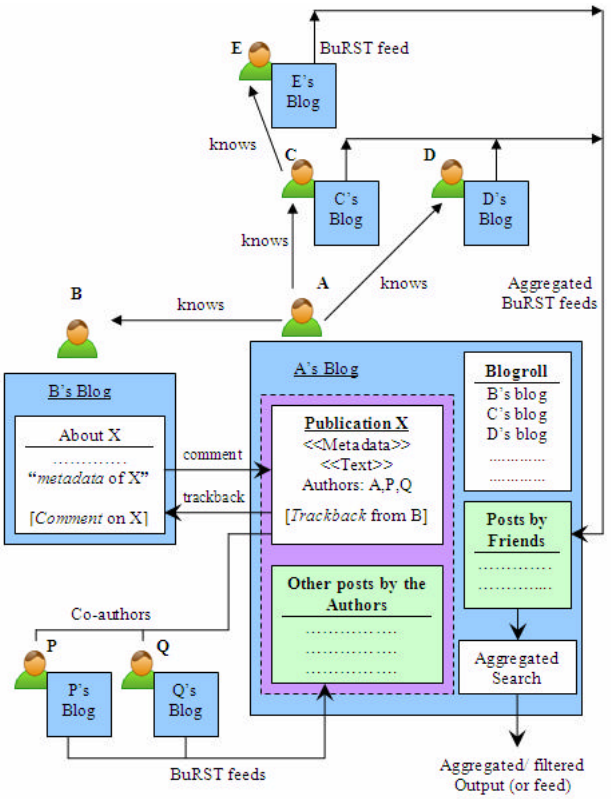

Figure 1. Example scenario.

The user may also aggregate information from other information sources that support BuRST format. Then, he/she may search and filter the aggregated collection. For instance, the user may only be interested in the articles from a particular journal and containing a specific keyword. The result may be sorted by various fields. The aggregated and filtered output thus obtained is again exposed as a new BuRST feed. The user may subscribe this feed on his RSS reader and get desired notifications. The feed can also be used by others and further integrated with other information sources. For instance, a user may integrate the feed with articles from other related journals. Thus, information can be seamlessly integrated, filtered and reused.

3. REQUIREMENTS

Based on the discussions and scenario presented in the above sections we can summarize our main contributing requirements as follows.

Decentralized Information Sharing. The system should be fully decentralized as ordinary blogging systems. The platform should not only be able to publish metadata but also aggregate information from distributed sources. Publishing, aggregation and redistribution of information should serve decentralized information exchange.

Extended RSS Aggregation. The system should be able to aggregate extended RSS to process the embedded bibliographic metadata in RSS feeds.

Social Network based Aggregation. We should be able to aggregate bibliographic metadata and posts on each blog through socially connected sources.

Interoperability. It should be possible to aggregate structured information from different systems and interoperate with them. Standard formats to ensure compatibility and interoperability. We should also be able to quote metadata from different bibliography management systems and comment on them.

Integration and Filtering. We should be able to integrate information from different sources, filter the collection by metadata and redistribute the results. The system should serve as a platform for flexible mixing of information sources resulting in new sources which can further be mixed with other sources.

4. DESIGN AND IMPLEMENTATION

4.1 System Architecture

Fig. 2 shows the designed architecture of the system. It consists of two sub-systems.

The *publishing system* facilitates publishing blog entries and metadata about publications. The publishing system is built over an existing blogging infrastructure. Structured blog entries contain metadata about publications based on the SWRC ontology. BibTeX scrapers extract bibliographic metadata about quoted publications from other blogs and bibliographic sites. The metadata is stored in an RDF metadata store. The blog contents are published in machine readable BuRST feeds. The publication system also publishes FOAF (Friend of a Friend) profile of the blog-owner. FOAF is an ontology for describing people and the relations between them (http://www.foaf-project.org/).

The *aggregation system* utilizes RSS technology to aggregate publications from multiple sources. RSS or BuRST feeds to be aggregated may be retrieved from the linked FOAF profiles of researchers in the community. The FOAF crawler is used to gather FOAF profiles from the FOAF network of the researchers. The aggregated posts are output on the blog. The aggregated search allows the user to filter aggregated publications by defining various metadata criteria. The aggregated and filtered BuRST feed thus obtained can be exported again as a new BuRST feed.

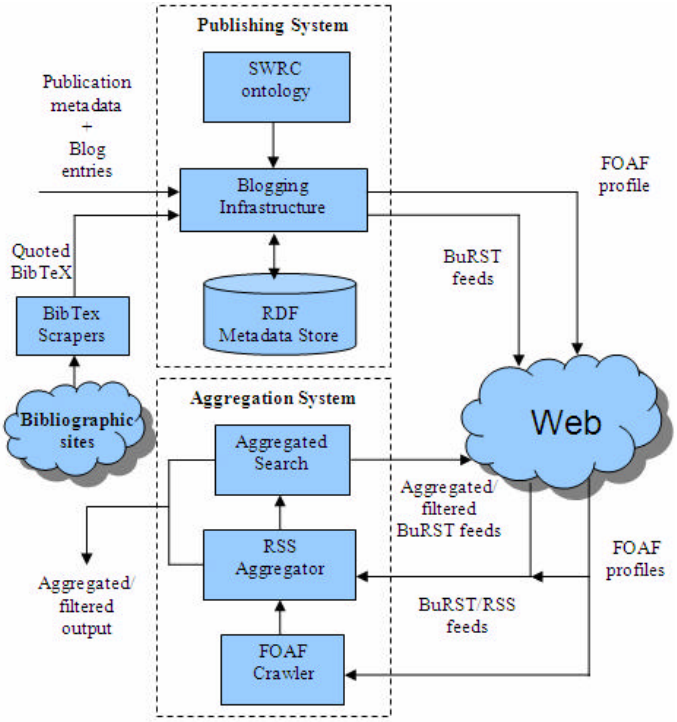

Figure 2. System architecture of SocioBiblog

Fig. 3 illustrates how the system can be introduced to co-exist and interoperate with existing systems. Existing blogging and publishing systems can usually generate RSS feeds (some publishing systems can generate BuRST feeds too). Aggregation systems exist separately. Our design integrates both publishing and aggregation parts. The publishing system extends existing blogging infrastructure and embeds metadata in RSS to produce BuRST feeds. The aggregation system can handle BuRST feeds as well as plain RSS (being compatible with BuRST). On the other hand, existing aggregation systems can also consume the RSS part (shown by the solid arrow) from a BuRST feed (shown by the dashed arrow) discarding the metadata.

For demonstration, SocioBiblog can aggregate BuRST feeds from BibSonomy (http://www.bibsonomy.org/) and BuRST feeds generated by the web services made available by Mika (http://www.openacademia.org/doc/services.html). These services can be used to generate BuRST feeds from any page containing list of BibTeX records or from a Sesame repository with SWRC publication instances.

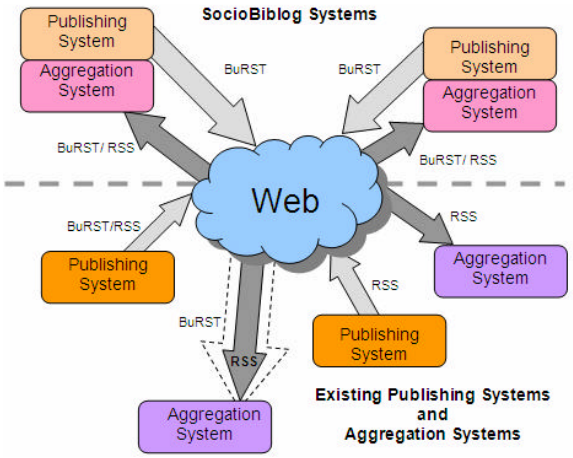

Figure 3. Publishing and aggregation on the current web with SocioBiblog.

4.2 Publication

4.2.1 Publication of Blog Entry and Metadata

Fig. 4 shows the SocioBiblog interface with some publication metadata. The semantic blog provides metadata entry forms for different SWRC publications. BibTeX snippets can also be imported directly to populate the entry form. Normal text blog entries can also quote and comment on other publications. Fig. 4 also shows a blog entry quoting a publication. Publication metadata entry is exported in SWRC, BibTeX formats and BuRST feeds. BibTeX (http://www.bibtex.org/) is currently the most widely used format by different systems and sites. Thus, export and import in this format would be important for interoperability. Blojsom (http://blojsom.sourceforge.net) has been used as the blogging platform. Metadata about publications are stored in RDF format in a MySQL database using the Jena Semantic Web framework (http://www.hpl.hp.com/semweb/jena.htm).

Metadata Search. The system allows searching bibliographic metadata published on the blog by specifying various metadata fields. It also searches into metadata quoted from other sources. The commented posts are marked to distinguish from the original publications. The interface is similar to the aggregated search discussed in section 4.3.2. The result of the metadata search is also exported as BuRST feed.

Blogroll and FOAF Profile. A web-based interface to maintain the blogroll of the blog-owner has been provided. Values from the XFN profile (http://gmpg.org/xfn/) are used to define relations with people in the blogroll which are mapped into FOAF one-to-one.

Figure 4. The blog interface.

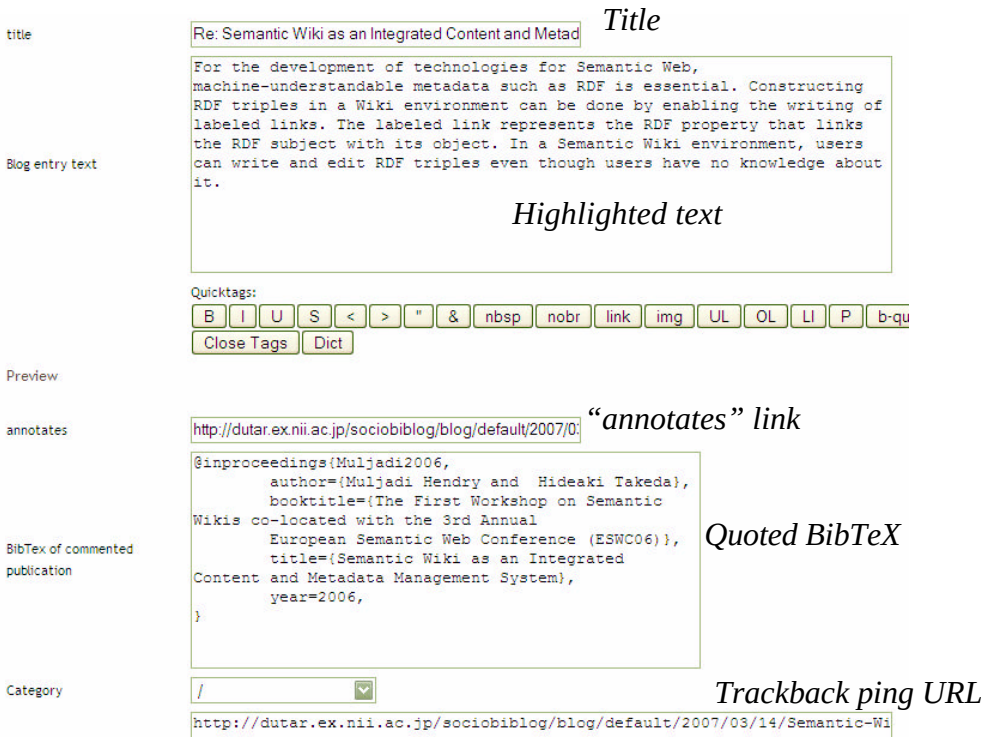

Figure 5. *Blog this* interface.

4.2.2 Commenting Mechanism

***Blog this* bookmarklet.** Commenting has been made convenient by employing javascript bookmarklet. The bookmarklet captures the title, URL, trackback ping URL of the blog entry being annotated and any highlighted text. The entry form is then automatically populated as shown in Fig. 5. The "annotates" element models the forward annotation link manifested as shown in Fig. 4. The link is also added in the BuRST feed to distinguish between quoted entries and the original publication.

BibTeX Scraping. When a publication is commented using *Blog this*, the system tries to scrape out BibTeX information if available. SocioBiblog currently provides scrapers for SocioBiblog instances, the ACM digital library (http://portal.acm.org/) and a generic BibTeX scraper which works for any web page that contains BibTeX snippet on the same page (this is applicable for many sites like Citeseer, DBLP, BibSonomy, CiteULike, etc). If the commented page contains multiple BibTeX snippets, the system selects the entry highlighted by the user or the first entry.

4.3 Aggregation

4.3.1 BuRST/RSS Aggregation

SocioBiblog generates BuRST feeds with embedded SWRC publication elements. The system aggregates BuRST/RSS feeds from friends listed in the blogroll and connected people in the social network neighborhood. Feeds from other systems and repositories can also be added to the blogroll. The latest publications and posts aggregated are displayed alongside in the blog as shown in Fig. 4. The system separates publications and non-publications. Further, when a blog entry for a publication is opened, BuRST/RSS feeds of the coauthors of the publication are downloaded and shown alongside. The feed URLs are determined from FOAF profiles of the co-authors. Flock RSS aggregator (http://flock.sourceforge.net) has been extended to handle BuRST.

Social Network based Aggregation. SocioBiblog aggregates feeds from the friendship neighborhood up to two levels deep – directly linked friends and friends of the friends. The aggregator first subscribes feeds from

sources listed in the blogroll. The blogroll includes FOAF links and optionally BuRST/RSS feed URLs of the friends. If the system cannot find the feed URL of a person in this list, it looks for the URL in that person's FOAF profile. Then, the system goes one level deeper into the FOAF social network to find friends of people listed in the blogroll and adds their feeds to the subscription list. The BuRST/RSS feed URLs are retrieved from their FOAF profiles. The second level of linked friends is traced and aggregation feeds are collected whenever the blogroll is updated. Discovery of linked sources and aggregation are done in background without affecting responsiveness of the system.

FOAF Crawler. The Elmo scutter (http://www.openrdf.org/doc/elmo/users/index.html) has been adapted as a FOAF crawler to find out FOAF links of authors. Elmo provides the interface and options to manage crawling. The crawler traces rdfs:seeAlso elements for FOAF links and gathers FOAF profiles in a database. The profiles can contain RSS/BuRST feed URLs. While uploading publications, the user may enter FOAF links of authors. The user can search the crawled FOAF database for FOAF links.

4.3.2 Aggregated Search and Filtering

The BuRST feeds aggregated from multiple sources can be searched and sorted by SWRC fields like title, author, type, etc as shown in Fig. 6 The system also searches the metadata quoted in comments on publications. The user can specify values for different metadata fields using the online form. The results can be sorted by clicking on the field headers. The process can be considered as filtering because only those entries satisfying the specified metadata criteria are included in the result. The metadata filtering parameters can also be directly specified in the request URL.

The aggregated and filtered results obtained are exported again as new BuRST feed. The user can subscribe this BuRST feed URL in his/her RSS reader and instantly get notified of desired updates. The feed can further be used by others and combined with other sources to construct their own aggregated information collections. Thus, the system serves to integrate various distributed information sources, filter them and construct new information sources.

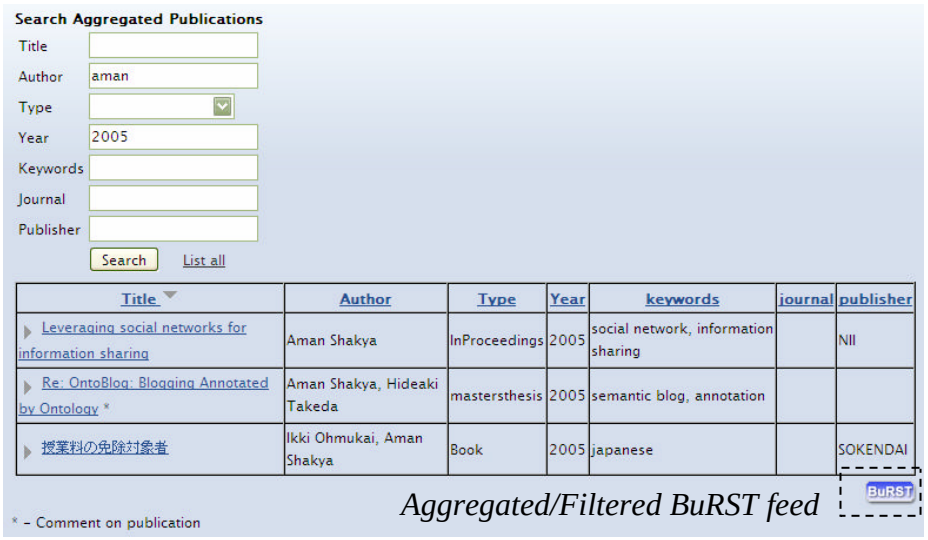

Figure 6. Searching aggregated publications.

5. RELATED WORK

A semantics based publication management system using RSS and FOAF has been described by Mika et al. (2005). The application harvests BibTeX from several locations and crawls FOAF files. All the information is stored in RDF format in a centralized store. Bibster (Haase et al. 2004) is a peer-to-peer application for sharing bibliographic metadata. It uses SWRC as application ontology and ACM topic hierarchy as domain

ontology. Blogging makes a wide online community instead of being limited to a peer-to-peer network and supports commenting.

BibSonomy (Hontho et al. 2006) is a social bookmarking system for sharing bookmarks and publication references. The folksonomy used has been extended by introducing binary relation between tags. The system is a centralized collaborative site. It does not provide aggregation and metadata search features. It allows posting remarks but it is one way. Trackbacks make it bidirectional in SocioBiblog. BibSonomy does not distinguish between original publication and quoted data; rather it employs duplicate detection heuristics. SocioBiblog maintains the distinction by providing a link to the original source. CiteULike (http://www.citeulike.org/) is popular social bookmarking site. Citeseer, Google scholar, etc serve as bibliographic libraries. These sites are centralized and do not provide semantic web based formats. However, SocioBiblog can interoperate with these sites because they provide data in BibTeX format which can be easily scraped, quoted and commented in blog entries. The BibTeX listings provided by such legacy systems can also be converted to BuRST feeds using web services made available by Mika (http://www.openacademia.org/doc/services.html) which in turn can be aggregated by SocioBiblog.

The Semantic Blogging Demonstrator uses the bibliographic domain (Cayzer 2004a, b). Annotation has been modeled as "blog entries *contain* bibliographic items". The demonstrator offers semantic navigation, view and query capabilities. The RDF store has been implemented as a single file. We use a relational database for scalable RDF storage. The system doesn't offer cross-blog aggregation. Karger and Quan (2005) extend Haystack to enable blogging. Blog entries are viewed as cross-blog reply graphs. RSS subscription facility is also provided. Haystack is complex for a lightweight application like blogging. Möller et al. (2006) identify two kinds of metadata in blogging. Structural metadata deals with parts of a blog and relations between them. Content metadata describes the post. Different small ontologies like FOAF, vCard, BibTeX/SWRC, iCalendar, etc. have been used. semiBlog (Möller et al. 2006, 2005a, b) emphasizes utilizing data available on the user's desktop. The semantic blogging systems serve as decentralized publishing systems but aggregation and interoperation are not possible.

Online systems like Yahoo Pipes (http://pipes.yahoo.com/pipes/) and Dapper (http://www.dapper.net/) enable the users to aggregate information from various sources, filter and process the information flexibly and share them with others. However, these are limited to traditional RSS feeds and web pages and cannot take advantage of semantically rich information. SocioBiblog offers a decentralized environment where different sources can interoperate and structured information can be aggregated from multiple sources, filtered and utilized flexibly.

6. CONCLUSION AND FUTURE WORK

In this paper, we proposed SocioBiblog as a semantic blogging platform for decentralized sharing of bibliographic information in a community. Traditionally, web sites and repositories are data providers and users are passive consumers. SocioBiblog combines the capability of both publishing and consuming information into a single unit for each user. The unit can aggregate information from linked sources in the community and traditional data providers. It can actively publish original contents as well as aggregated contents. It can act as an information broker to aggregate, filter and redistribute information. In this decentralized information ecosystem, SocioBiblog can be considered as an active unit which can push and pull contents and facilitate flow on information in the community. The network of such units would allow flexible handling of information channels and can offer the potential of evolution. The semantic web offers the basis for the units to share structured information, process them based on semantics and interoperate with each other and existing systems.

We still have good scope for enhancements and future research. The domain ontology, such as a topic hierarchy, could be incorporated in the blogging system to categorize posts. Full features of the SWRC ontology, including concepts like organization, person and event, could be utilized. Semantic capabilities like semantic search and navigation would make the system more useful. We can cooperate with a semantic blogging client like semiBlog (Möller et al. 2006, 2005a, b) to utilize data from the user's desktop. Currently, RSS aggregation is limited by fixed number of latest entries. Effective archiving mechanisms may be explored in the future. Further, indexing and ranked search mechanisms could be incorporated.

REFERENCES

Berners-Lee et al, 2001. The Semantic Web. *In Scientific American*, Vol. 284, No. 5, pp. 34-43.

Cayzer, S. 2004a. Semantic Blogging and Decentralized Knowledge Management. *In Communications of the ACM*, Vol. 47, pp. 48-52.

Cayzer, S., 2004b. Semantic Blogging: Spreading the Semantic Web Meme. *In XML Europe*. http://www.idealliance.org/papers/dx_xmle04/papers/03-05-03/03-05-03.html

Haase, P. et al, 2004. Bibster - A Semantics-Based Bibliographic Peer-to-Peer System. *Proceedings of the International Semantic Web Conference (ISWC2004)*, Hiroshima, Japan.

Hotho, A. et al, 2006. BibSonomy: A Social Bookmark and Publication Sharing System. *Proceedings of the First Conceptual Structures Tool Interoperability Workshop at the 14th International Conference on Conceptual Structures*, Aalborg, Denmark, pp. 87-102.

Karger, D. R. and Quan, D., 2005. What Would It Mean to Blog on the Semantic Web? *In Journal of Web Semantics*, Vol. 3, No. 2, pp. 147-157.

Mika, P. et al, 2005. Semantics-based Publication Management using RSS and FOAF. *Proceedings of the 1st Workshop on the Semantic Desktop (SD 2005), 4th International Semantic Web Conference*, Galway, Ireland.

Mika, P., 2005. Bibliography Management using RSS Technology (BuRST). http://www.cs.vu.nl/~pmika/research/burst/BuRST.html

Möller, K. et al, 2006. Using Semantics to Enhance the Blogging Experience. *Proceedings of the 3rd European Semantic Web Conference (ESWC2006)*, Budva, Montenegro, pp. 679-696.

Möller, K. and Decker, S., 2005a. Harvesting Desktop Data for Semantic Blogging. *Proceedings of the 1st Workshop on The Semantic Desktop - Next Generation Personal Information Management and Collaboration Infrastructure at ISWC2005*, Galway, Ireland, pp. 79-91.

Möller, K. et al, 2005b. semiBlog - Semantic Publishing of Desktop Data. *Proceedings of the 14th Conference on Information Systems Development (ISD2005)*, Karlstad, Sweden.

Sure, Y. et al, 2005. The SWRC Ontology - Semantic Web for Research Communities. *Proceedings of the 12th Portuguese Conference on Artificial Intelligence - Progress in Artificial Intelligence (EPIA 2005)*, Covilha, Portugal, pp. 218 – 231.