Semblog Project

Hideaki Takeda
National Institute of Informatics
2-1-2, Hitotsubashi, Chiyoda-ku
Tokyo, Japan
takeda@nii.ac.jp

Ikki Ohmukai
National Institute of Informatics
2-1-2, Hitotsubashi, Chiyoda-ku
Tokyo, Japan
i2k@nii.ac.jp

ABSTRACT

Semblog Project aims to provide personal knowledge publishing tools with Semantic Web technlogies, in particluar, metadata technologies. Semblog tools provide an integrated environment for gathering, authoring, publishing, and making human relationship seamlessly to enable people to exchange information and knowledge with easy and casual fashion. We use lightweight metadata format like RSS to activate the information flow and its activities. We define three level of interest of information gathering and publishing i.e."check","clip" and "post" and provide suitable ways to distribute information according to the interest level. We currently provide two types of extended content aggregator and information retrieval recommendation applications. We also design new metadata module to define personal ontology that realizes semantic relations among people and Weblog sites.

1. INTRODUCTION

Semblog Project aims to provide personal knowledge publishing tools with Semantic Web technlogies. Semblog tools provide an integrated environment for gathering, authoring, publishing, and making human relationship seamlessly to enable people to exchange information and knowledge with easy and casual fashion. We overview our project in this paper.

2. SIX TYPES OF HUMAN INFORMATION ACTIVITIES

We aim to build systems that support not only human information activites, i.e, activities in which people handle and process information, but also human communication activities, i.e., those in which people exchange information with other people. We show the scheme of human information and communication activities in Figure **??**.

It is an extension of "Activities and Relationships Table" proposed by Shneiderman[**?**]. The first layer called "information layer" has three elements that concern information handling, i.e. *collect*, *create* and *donate* information. Three activities form cycle, i.e.they are invoked repeatedly. The second layer called "communication layer" has also three elements that concerns communication handling, i.e. *relate*, *collaborate* and *present* people. They also form cycle. The two layers are not isolated rather closely related to each other. For example, human relationship can contribute to collect information, while collecting information can result in forming new human relationship. It implies that support for information activities requires support for communication activities, and vice versa.

Web itself supports only *donate* (publishing) activity. *Collect* activity is supported by independent service, i.e, search engines. *Create* activity is also supported independetly like HTML editors. We can say that web covers information activities but each activity is supported independently thus not integrated to each other. Furthermore no support for activities on the communication layer.

Weblog is better than web from this viewpoint, because Weblog covers more activities in an integrated manner. Weblog directly supports authoring and publishing in an integrated way, since Weblog tools usually support both. Furthermore it commits communication activities. Each weblog can be seen as an identifier of indivisual. We regard a weblog as persnality of an author or a group of authors. Such weblog users tend to refer to each other and form so-called "weblog communities". Weblog tools support such communication like TrackBack and ping. But the support is partially, indirectly, and not integrated.

Semblog tools extend weblogs by adding flexible but uniform operations for weblog sites and entries like aggregation and clipping, and facilities for searching and contacting to other weblog sites. It means that Semblog suite supports communication activities as well as information activities.

3. SEMANTIC WEB AND WEBLOG

In this section, we overview the current situation of information distribution on Web, Semantic Web and Weblog. Web lacks functionality of information distribution and Semantic Web aims to fulfill the functionality by metadata. We support Semantic Web approach but it has difficulty of metadata annotation. So we focus on Weblog that also use metadata for information distribution.

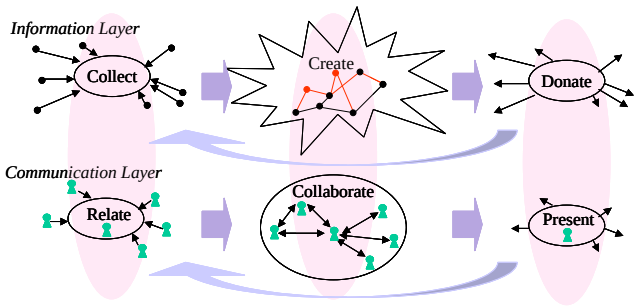


Figure 1: Information and Communication Activities

3.1 Information Processing with Semantic Web Techniques

We propose content distribution support system for individuals with Semantic Web techniques. Information distribution process does not mean just publishing but an integrated process containing information gathering and authoring. In the current web, however, there is no framework to support the whole process of information distribution despite the fact that Berners-Lee specified the first world wide web to support both authoring and publishing process equally[?]. Protrusion of information publishing just accelerates so-called information overload.

There is great hope that the Semantic Web technologies will resolve information overload. According to the manifesto[?], Semantic Web is an environment, which consists of the contents with machine-readable (semantic) tags and the software agents, to realize autonomous information distribution and syndication. Resource Description Framework (RDF)[?] and ontology definition languages like OWL[?] are recommended by W3C as elemental technologies of the Semantic Web and these are now in practical use.

However it is difficult to produce contents with semantic tags because of their complicated syntax and vocabulary. Ordinary people hardly find merit of semantic annotation although it is time-consuming task. It is also impossible to annotate the semantic tags to existing enormous information on the Internet. There are some researches about automatic annotation with AI techniques and natural language processing[?] however their effects are still unclear.

In our approach, we use lightweight metadata formats, i.e.RSS 1.0 (RDF Site Summary[?])to activate the information flow and its activities. RSS is one of the metadata to describe summary of web site. It contains general attributes of the site i.e. title and publisher ' s name of the web site, and excerpt and updated date of its contents. Number of web sites already publish RSS metadata, and several applications and services called RSS aggregator are provided based on this trend.

An aggregator collects these RSS from various web sites and reform them to organize this large amount of contents to show at glance. There are two types of aggregator; one is standalone application that is executed on client PCs. The other is an aggregation service that runs on the internet server and the user access via her/his web browser. The former applications provide rapid browsing of RSS by their flexible user interface and the latter enables the user to access their information wherever she/he is.

Use of RSS and the aggregator decreases information gathering, but it is just part of information distribution process as we mentioned. Gathering should be related to information authoring and publishing otherwise it will be another search engines without any selection or extraction.

3.2 Information Creation with Weblog

Weblog is one of the most advanced systems that use metadata in gathering, creating and publishing.

Recently weblog has come into the spotlight in the Web[?]. There is no strict definition about Weblog but it is recognized as web site that consists of miscellaneous notes updated daily[?]. In such sites the authors do not make efforts to knit up these contents because weblog tools align them in chronological order with well-designed HTML format. We call these frequently posted contents as small contents in this paper. Small contents include various subjects including journal, expertise and critique. One of most popular topics is the introductions and comments of the web sites ranging from news sites to the other small contents. Some weblog sites attract the attention with their own editorial policy. The authors of weblog sites reedit the existing web contents by quoting them. Moreover there are new types of Weblogs that criticize the other weblogs so that these weblogs are regarded to organize the "weblog community". Now there are millions of weblog sites in the World. It is surprising number because these people are now active information creators and distributors as well as information receivers thanks to weblog.

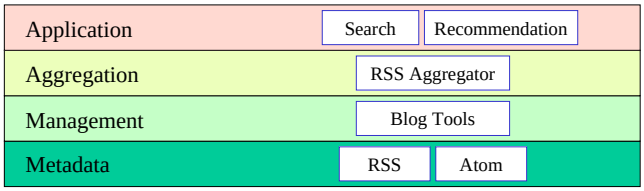
Most of weblog sites use so-called weblog tools that are kind of content management systems (CMS)Weblog tools enable the author to describe and edit the small contents via web browser and transform the contents form text format to HTML files. These tools are implemented based on MVC (Model View Controller) model which is the fundamental concept of web applications. The author defines view template once then do not have to decorate the contents with various HTML tags. This model decreases the cost of publication remarkably in comparison with traditional style that requires local text editor and FTP. This feature contributes abundant production of the small contents.

Weblog tools usually generate RSS automatically. General attributes such as publisher ' s name are set as profile by the user. Excerpt and updated date of each content are generated by the tools. Most of distributed RSS are generated by these Weblog tools. Main purpose of RSS aggregator is also to browse the Weblog site. Currently the number of news site feeding RSS is expanding.

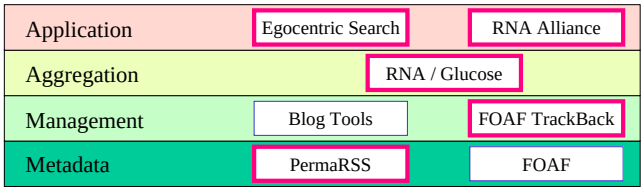
4. SEMBLOG ARCHITECTURE

4.1 The basic model

We propose personal publishing system using Semantic Web techniques and weblog tools. The system supports the whole process of information distribution which includes gathering,



(a) weblog



(b) Semblog

Figure 2: Four basic layers for weblog and Semblog

authoring and publishing. Furthermore it connects information distribution process of different people seamlessly, i.e., it supports finding collaborating and advertising people for information distribution.

We divide system architecture into the following four levels, i.e., metadata format, metadata management, metadata aggregator, and metadata application. In weblog, they correspond to RSS, Weblog tools, RSS aggregator, and various applications on Weblog respectively (see Figure ??(a)). In Semblog, Metadata on person and interpersonal relations is added in order to include activities on the communication levels. We adopt FOAF (Friend-Of-A-Friend) as person metadata. Content (RSS) and person metadata should be processed seamlessly in every level.

We build our systems by integrating new technologies and existing technologies shown in Figure ??(b). A box with think line indicates our proposal and a box with thin line an existing one.

4.2 RNA: Web-based RSS Aggregator

RNA is a Web-based RSS aggregator written with Perl. A user can operate RNA through her/his Web server. Figure ?? shows a snapshot of RNA.

Firstly the user register URIs of RSS in configuration page of RNA shown in Figure ??. The user can categorise these RSSs. List of sites checked by the user are converted into an RSS that can be read by other RSS-based applications again. RNA can also import and export OPML that is a standard metadata set for Web bookmark.

RNA produces site/entry list ordered by updated time of each element. After getting RSS files from various sources, RNA parses these RSSs and merges into single a "global" RSS tree. RNA converts this global tree to several forms by ordering chronologically. These partial trees are published as RSS and rendered into HTML. Figure ?? shows a site list with HTML. User can browse description element of RSS in each channel (site).

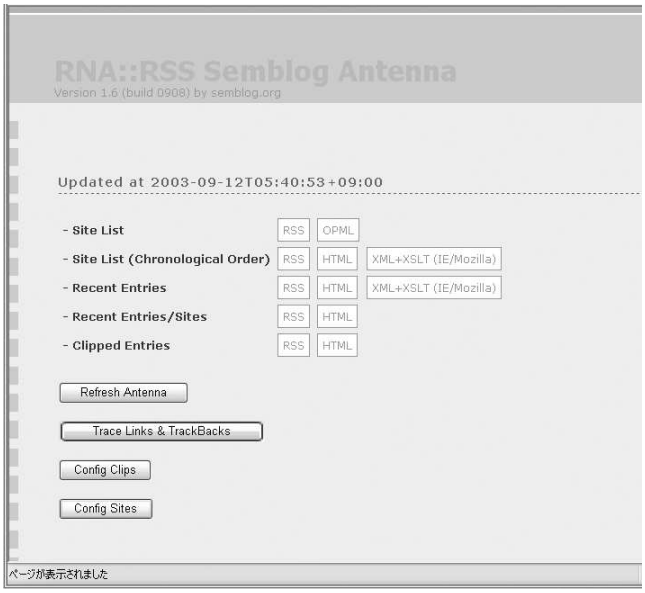


Figure 3: Snapshot of RNA

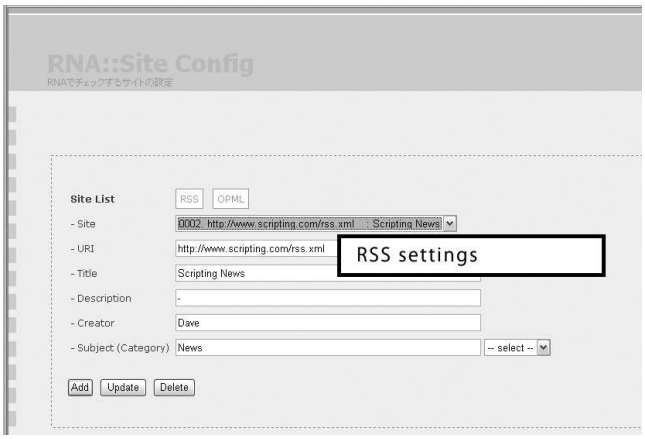


Figure 4: RSS Registration

RNA supports various output formats such like HTML, RSS and other forms i.e., JavaScript with server/client-side XSLT engines and original template engine. The user can create customised partial tree using plug-in and template script.

The user can save a favourite content in RNA to a clip list with one click. Clipped contents are stored in the "clipped" RSS tree and it is published like other RSSs. RNA can post clips to social bookmark service such like del.isio.us[1]. RNA extracts TrackBack links from each entry in registered sites, and embeds TrackBack metadata in RSS and renders it.

The current version of RNA cooperates with RSS-based search engine such as Technorati[2] and Bulkfeeds[3]. By set-

[1] http://del.isio.us/
[2] http://www.technorati.com/
[3] http://bulkfeeds.net/

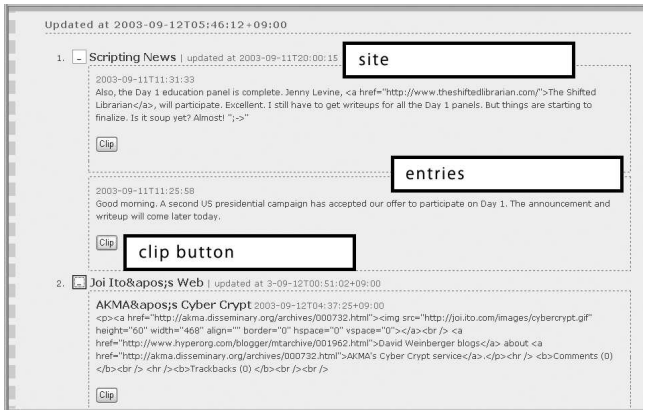


Figure 5: Site List in HTML

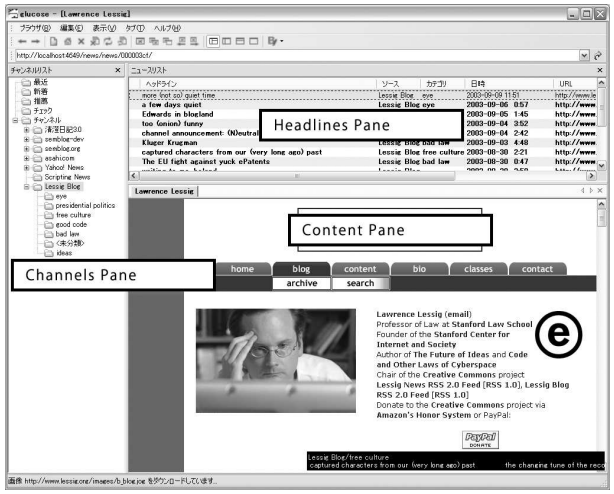


Figure 6: Snapshot of Glucose

ting some keywords, the user can obtain new contents from non-registered sites.

Most of RSS generated by Weblog and news sites include categorical information with `<dc:subject>` vocabulary. RNA can aggregate selected contents by specifying certain categories, and re-distribute such categorical RSS too.

It is necessary to get RSS and build trees periodically since those contents may change according to update of information sources. RNA can update periodically by cron interface of the server. Update interface can be called both manually and remotely by XML-RPC message that is generated automatically by Weblog tools.

RNA checks syntax of acquired RSS and corrects them if they are not valid. RNA converts all versions of RSS into 1.0, which is based on RDF model.

4.3 Glucose: Stand-alone RSS Aggregator

Glucose is also an extended RSS aggregator for Windows. Figure ?? shows a snapshot of Glucose. Unlike orthodox aggregators, Glucose is developed to support information distribution process in cooperation with coordinating with RNA. Main functions and interfaces of Glucose are shown below.

Like in RNA, the user registers URIs of RSS or OPML site list. Glucose can access several news sites without RSS by "sensor" script which extracts articles and converts them into RSS.

Glucose has three panes interface. The left pane shows "RSS Channels" that are subscribed RSSs by the user. The upper right pane indicates the headline list of contents including title, updated time, source and category. The lower right pane is the main pane that shows the content of the entry specified by the headline pane.

Glucose can extract TrackBack links from each content. Obtained links are shown below the corresponding entry in headline pane with "Re:..." like a mailer.

The user can post an entry to her/his Weblog if she/he has specific interest for content, since Glucose is equiped With Weblog editor interface which communicates Weblog tools with XML-RPC protocol (Figure ??). The user can also clip contents and publish the clip list, since Glucose can communicate RNA with XML-RPC.

4.4 Suuport for communication activities with Semblog tools

In our systems, we provide support of information activities in the various level. People simply read RSS-based contents from various information sources with our aggregators. These functions can make benefit to individual users in reading and writing Weblog contents.

As we mentioned earlier, the real value of Semblog is support of communication activities as well as information activities. Our Semblog system can be used as an information sharing platform. It is based on simple metadata so that it can be extended easily.

4.4.1 FOAF TrackBack

RNA alliance is a content recommendation system based on cooperation of multiple RNAs. We use FOAF metadata to identify each RNA. FOAF is an RDF-based metadata format for describing human relationship. Besides the basic elements such as name, email and URI of the user, FOAF provides a statement that user X knows user Y. The current version of RNA can generate FOAF data.

RNA also has an interface for FOAF management to extend social network easily. We call this method as "FOAF TrackBack". First the user X enters an RNA URI of the user Y in her/his own FOAF manager. The manager X asks the manager Y to acquire the FOAF data of Y, and writes "X knows Y" link in its FOAF. The manager Y records "Y isKnown by X" link in its FOAF and notifies to the user Y. If the user Y agrees, her/his manager registers "Y knows X" link. Repeating this process, a personal network of the user is constructed. Following recommendation methods are performed in the network.

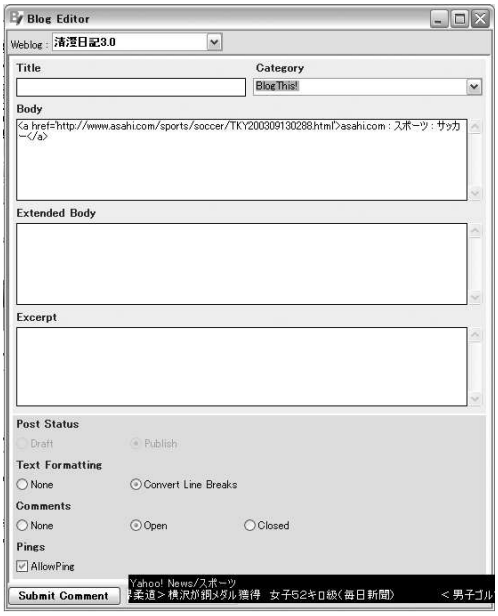


Figure 7: Weblog Editor

4.4.2 RNA Alliance

RNA Alliance is collaborative recommendation based on difference of registered sites or clips among multiple RNAs. At first it calculates similarities: S_i between the user's RNA: R_0 and a RNA on the personal network: R_i $(1 < i < n)$. Each RNA has a list of URIs: $R_i = \{u_0, \ldots, u_k\}$.

$$S_i = \frac{|R_0 \cap R_i|}{|R_0| + |R_i|}$$

The system gives recommendation score: $V(u)$ to each URI by the following formula:

$$V_i(u) = \begin{cases} S_i & \text{if } u \in R_i \\ 0 & \text{if } u \notin R_i \ (i = 1, \ldots, n) \end{cases}$$

$$V(u) = \frac{\sum_{i=i}^{n} V_i(u)}{n}$$

This score is used for recommendation to R_0's user if URI u is not included in R_0.

The system shows the list of recommended URIs sorted by the score. The user can add these URI to her/his own "check" list.

4.4.3 Personal Ontology

We propose a bottom-up personal ontology framework using RSS and FOAF metadata. To process small contents in various forms, we have to annotate a semantic markup with an ontology language to those contents. It is difficult to organize practical ontology hierarchy with top-down approach because building and maintaining such well-organized large ontology takes a lot of efforts. We aim to develop loose and bottom-up ontology system by combining personal classification, because we consider that personal knowledge will be represented with a routine work such as categorization and arrangement of information. Figure ?? indicates a conceptual architecture of the personal ontology system.

At first we define a personal ontology as a hierarchical system of categories. Everyone has those categories, and they routinely classify described and collected contents to the category. A label of a category can be named arbitrarily by user. Unlike the conventional ontology, the personal ontology has to be related to the person who produces it. Therefore we apply FOAF metadata to link between the ontology and the person.

Personal ontology metadata consists of FOAF, RDFS Ontology and Contents RSS. The FOAF describes personal information, and the RDFS ontology shows a structure of the categories, and the contents RSS shows written and collected contents by the user. We add two elements to basic FOAF model shown in Figure ?? (a). One is `<foaf:interest>` which is to point the contents RSS, and the other is `<rs:personalonto` that is originally defined by our Rough Semantics project [4] to indicate the RDFS ontology. The RDFS ontology is described with the form of Open Directory RDFS format shown in Figure ?? (b). Each node has a fragment ID.

The content RSS is similar to a conventional RSS. Our RSS uses `<foaf:topic>` to point a category on the RDFS ontology, while the conventional model applies `<dc:subject>` to express a thesis of a content. This makes our RSS to have backward compatibility. Example of this RSS is shown in Figure ?? (c). It should be noted that topics pointed by this tag are not restricted to those in their own ontology, rather any topics in others' and some global ontology. Separating ontology and instances enables such flexible management.

FOAF, RDFS ontology, and RSS are described in separate files so that we can keep compatibility with existing applications on these formats. This is a great benefit that our system can cope with such existing applications via these files.

Our framework enables applications and services to produce new types of search or recommendation. For example, mapping methods between two directories or bookmarks are applicable to the personal ontology. Egocentric search[?] is also able to realize easily by building a social network with `<foaf:knows>` in the users' FOAF.

Unlike these peer-to-peer model, we can calculate a similarity among a personal ontology and the global ontologies such like WordNet and ODP in advance. Multiple personal ontology can be matched with each other via the global ontology and this method needs less computation cost as shown in Figure ??. In addition, it is not necessary to modify that algorithm in P2P model and personal-global model because both ontology has the same structure.

[4] http://www.roughsemantics.org/

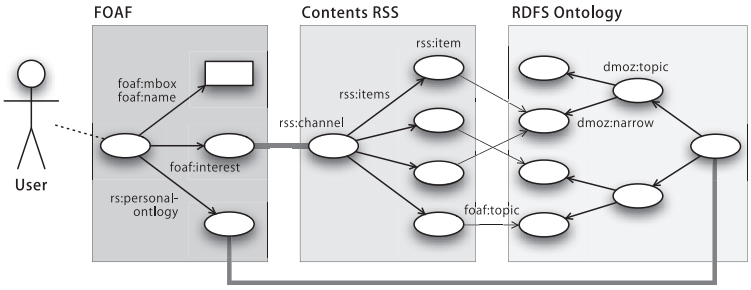


Figure 8: Personal Ontology Framework

```
<rdf:RDF
 xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
 xmlns:foaf="http://xmlns.com/foaf/0.1/"
 xmlns:rdfs="http://www.w3.org/2000/01/rdf-schema#"
 xmlns:rs="http://www.roughsemantics.org/rs/0.1/"
>

<foaf:Person>
<foaf:name>Ikki Ohmukai</foaf:name>
<foaf:nick>i2k</foaf:nick>
<foaf:mbox rdf:resource="mailto:i2k@grad.nii.ac.jp" />
<foaf:weblog rdf:resource="http://www.semblog.org/i2k/" />
<rdfs:seeAlso rdf:resource="http://www-kasm.nii.ac.jp/~i2k/foaf.rdf" />
<foaf:interest rdf:resource="http://www-kasm.nii.ac.jp/~i2k/index.rdf" />
<rs:personalontology
            rdf:resource="http://www-kasm.nii.ac.jp/~i2k/ontology.rdf" />
....

</foaf:Person>
</rdf:RDF>
```

(a) Extended FOAF

```
<RDF xmlns:rdf="http://www.w3.org/TR/RDF/"
     xmlns:dc="http://purl.org/dc/elements/1.0/"
     xmlns="http://directory.mozilla.org/rdf">

<Topic rdf:id="Top">
  <tag catid="1"/>
  <dc:Title>Top</dc:Title>
  <narrow rdf:resource="Top/Arts"/>
  <narrow rdf:resource="Top/Business"/>
  <narrow rdf:resource="Top/Economy"/>
  <narrow rdf:resource="Top/Tech"/>
  ....
</Topic>

<Topic rdf:id="Top/Arts">
  <tag catid="2"/>
  <dc:Title>Top/Arts</dc:Title>
  <narrow rdf:resource="Top/Arts/Fine"/>
  ....
</Topic>
```

(b) RDFS Ontology

```
<item rdf:about="http://www.semblog.org/i2k/archives/000304.html">
<title>Blog Hacks</title>
<link>http://www.semblog.org/i2k/archives/000304.html</link>
<description>
Monday's child is fair of face, Tuesday's child is full of grace,
Wednesday's child is full of woe, Thursday's child has far to go,
Friday's child is loving and giving, Saturday's child works hard for his living,
And the child that is born on the Sabbath day is bonny and blithe, and good and gay. ...
</description>
<dc:subject>trivia</dc:subject>
<foaf:topic rdf:resource="http://www-kasm.nii.ac.jp/~i2k/ontology.rdf#Top/Arts">
<dc:creator>i2k</dc:creator>
<dc:date>2004-04-09T01:24:16+09:00</dc:date>
</item>
```

(c) Contents RSS

Figure 9: Personal Ontology Metadata

5. USE CASE

We applied our system to some communities.

One is for an academic conference called JSAI2004 (Figure ??). Participants registered URI of her/his Weblog to RNA so that other attendees can browse various opinion for the conference and papers. Unlike conventional closed system, RNA provides that an author of an opinion keeps her/his authorship permanently.

Other example is education support. Senshu University developed class support system based on RNA. In this system, all students and teaching staff should have Weblog and all contents will be aggregated with each class or project respectively. The user post her/his content using original editor interface which communicates multiple Weblog tools and RNA. RNA aggregates and shows recently updated contents of member so that the user can access newest topics in the class and project in the university.

RNA is used as person-based contents management system. Research institute of economy, trade and industry (RIETI)[5] publishes Weblog of its research associates with RNA. Official contents should be managed in single policy but it may restrict their contribution since it is so messy to follow. On the other hand, it does not seem to official contents when each member just publishes her/his Weblog freely. Thus RIETI introduces RNA to aggregate all contents from their Weblog and embed composite contents into official Web site. This model may decreases management cost.

We distribute RNA and Glucose in our web site[6]. About 3,000 users downloaded RNA and over 150,000 users downloaded Glucose from September 2003.

[5] http://www.rieti.go.jp/en/index.html
[6] http://www.semblog.org/wiki/?en

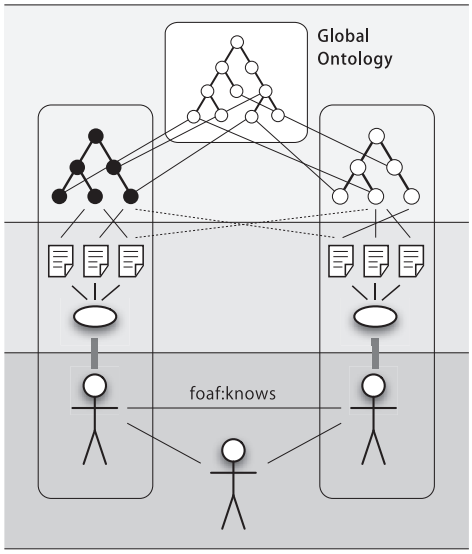


Figure 10: Bottom-up Ontology

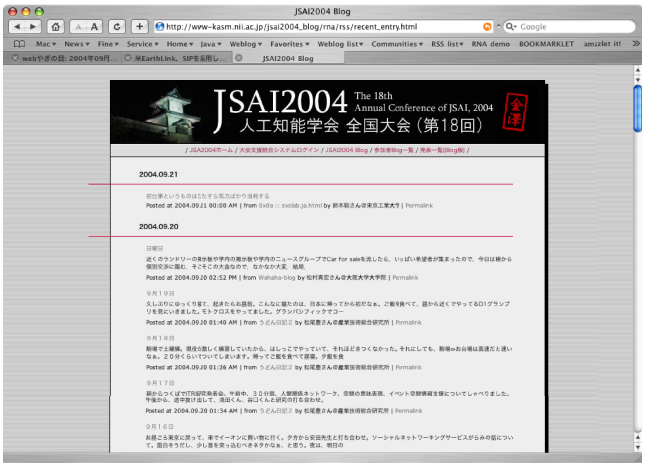


Figure 11: RNA in an academic conference

6. CONCLUSION

In this paper we propose personal publishing system with Semantic Web techniques and Weblog tools. We use lightweight metadata format like RSS to activate the information flow and its activities. We define three level of interest of information gathering and publishing, i.e., "check", "clip" and "post", and provide suitable ways to distribute information depending on the interest level.

Through these techniques and systems, we have shown that metadata can be used to realize more efferent and more personalized information distribution. Metadata design should be careful because it should be acceptable by many people and systems. Our approach, i.e., extending and integrating RSS and FOAF, is successful in this aspect since existing tools like Weblog tools are ready to use them. We hope that our approach will be bridge between emerging Semantic Web technologies and other growing technologies on the Internet.

7. REFERENCES

[1] E. Aimeur, G. Brassard, and S. Paquet. Using personal knowledge publishing to facilitate sharing across communities. In M. Gurstein, editor, *Workshop on (Virtual) Community Informatics, Held in conjunction with the Twelfth International World Wide Web Conference (WWW2003)*, 2003.

[2] T. Berners-Lee. Roadmap to the semantic web. http://www.w3.org/DesignIssues/Semantic.html, 1998.

[3] T. Berners-Lee. *Weaving the Web: The Original Design and Ultimate Destiny of the World Wide Web.* HarperBusiness, 2000.

[4] R. Blood, editor. *We've Got Blog: How Weblogs Are Changing Our Culture.* Perseus Books, 2002.

[5] S. Dill, N. Eiron, D. Gibson, D. Gruhl, R. Guha, A. Jhingran, T. Kanungo, S. Rajagopalan, A. Tomkins, J. A. Tomlin, and J. Y. Zien. Semtag and seeker: bootstrapping the semantic web via automated semantic annotation. In *WWW '03: Proceedings of the twelfth international conference on World Wide Web*, pages 178–186, New York, NY, USA, 2003. ACM Press.

[6] R. S. S. . S. W. Group. Rdf site summary (rss) 1.0. http://web.resource.org/rss/1.0/spec, 2001.

[7] F. Manola and E. Miller. RDF Primer. W3C Recommendation 10 February 2004. http://www.w3.org/TR/rdf-primer/.

[8] D. L. McGuinness and F. van Harmelen. OWL Web Ontology Language Overview. W3C Recommendation 10 February 2004. http://www.w3.org/TR/owl-features/.

[9] K. Numa, I. Ohmukai, M. Hamasaki, and H. Takeda. Egocentric search based on rss. In *Poster Proceedings of Third International Semantic Web Conference (ISWC2004)*, 2004.

[10] B. Shneiderman. *Leonardo's Laptop: Human Needs and the New Computing Technologies.* MIT Press, 2002.