Creating and Sharing Structured Semantic Web Contents through the Social Web

Aman SHAKYA

DOCTOR OF
PHILOSOPHY

Department of Informatics,
School of Multidisciplinary Sciences,
The Graduate University for Advanced Studies (SOKENDAI)

2009 (School Year)

September 2009

A dissertation submitted to

The Department of Informatics,

School of Multidisciplinary Sciences,

The Graduate University for Advanced Studies (SOKENDAI)

In partial fulfillment of the requirements for

The degree of Doctor of Philosophy

PhD Committee:

Hideaki Takeda	National Institute of Informatics, SOKENDAI
Nigel Collier	National Institute of Informatics, SOKENDAI
Kenro Aihara	National Institute of Informatics, SOKENDAI
Asanobu Kitamoto	National Institute of Informatics, SOKENDAI
Takahira Yamaguchi	Keio University

Acknowledgements

I would like to acknowledge my advisor Prof. Hideaki Takeda for his constant guidance and support throughout my study, research activities and in producing this thesis. I would also like to convey special thanks to my sub-advisors and all members in the committee for their help in improving and enhancing the thesis by providing constructive suggestions. I would like to convey special thanks to Prof. Vilas Wuwongse for constantly supporting my study by providing valuable suggestions and co-authoring several papers with me. I am also grateful to Dr. Ikki Ohmukai for his academic and technical assistance, server machine set-ups and co-authoring papers with me. I am especially thankful for his major role in the SocioBiblog project. I should also acknowledge Dr. Hendry Muljadi for the useful discussions and guidance, especially in increasing my knowledge about semantic wikis. I also thank Dr. Hideyuki Tan for his technical assistance to my work. Special thanks to assoc. Prof Ryutaro Ichise for his constructive advices to improve the research work. My sincere thanks also go to assoc. Prof. Yutaka Matsuo from the University of Tokyo for his interest in our work and providing the opportunity to present our work at the Biz-model conference.

I would like to express sincere gratitude to the Semantic Web Company, Vienna and CEO Andreas Blumauer for providing recognition to our work by awarding it in the Linked Data Vision competition and offering helpful advices. I am also grateful to IADIS for honoring our work with the best paper award in the area of Web 2.0. I also acknowledge Harry Halpin from the University of Edinburgh for showing special interest in our work and helping us with constructive discussions.

I should acknowledge the efforts of Dr. Kei Kurakawa for bringing my research work into real practical application for Japanese universities. I would like to thank Mr. Sanjil Shrestha from the Asian Institute of Technology, Thailand for using our work for significant real world project. I also thank Dr. Yessy Arvelyna and friends from the Tokyo International Exchange Center for accepting our experimental system for their real purpose and providing us useful feedback. I would like to thank all friends in different parts of the world who used our online systems and provided us vital feedback and suggestions.

I would like to convey special thanks to all the participants to my experiments for contributing their precious time and providing valuable feedback. I would like to thank Karina Shakya for her special help in my experiments. I am also grateful to all my colleagues and friends for their continuous support and fruitful discussions on my research and related areas. I would also like to express my sincere gratitude to the National Institute of Informatics for providing the environment, resources and funding indispensable for carrying out research studies, disseminating our research results and connecting us with researchers worldwide. Finally, I would like to thank my family and friends for their understanding, patience and support for my years of study abroad.

Preface

Sharing of information is important for its utilization to full potential. Information should be published with understandable semantics so that it can be used by others. It should also be accessible and properly disseminated. The Semantic Web provides structure and semantics to data making it machine understandable. The social web has made it easy for people to publish information online. It also enables collaboration and facilitates information dissemination by connecting people. These two areas complement each other to form a social Semantic Web. This is a highly promising direction but poses some major challenges.

The first challenge is to have people publish structured data on the social Semantic Web. Some specific problems for this are as follows. Systems for publishing structured data on the Semantic Web are complex and have considerable learning curve for people. It is also difficult for people to contribute due to strict constraints imposed by such systems. The second challenge is to form the models, so called ontologies, required to structure data with understandable semantics. People have a wide variety of data to share but there are limited ontologies and creating ontologies is difficult. Some specific problems for this dealt by the thesis are as follows. It is difficult to create perfect concept definitions to model things. It is not easy to cover the evolving requirements of all people. Moreover, different people may have multiple conceptualizations for the same thing due to different perspectives and contexts. It is not always possible to have consensus over conceptualizations and the collaborative process is itself difficult. Finally, proper dissemination of structured data on the web is also challenging. Information dissemination is mostly happening in a centralized and static way. There is a lack of flow of relevant structured information among people.

The thesis proposes some solutions to the specific problems. It proposes enabling people to contribute structured data by providing an easy-to-use social platform. It proposes allowing users to define their own concepts and freely contribute various types of data through a flexible and relaxed interface. Concepts contributed by people are partial definitions from their own perspective and multiple conceptualizations are allowed. These can be consolidated to form a rich unified conceptualization. This is possible by semi-automatic techniques for data integration and schema alignment supported by the community. A formalization of concept consolidation is also presented in the thesis. This serves as a loose collaborative approach that does not enforce consensus and direct interaction. Further, concepts can be semi-automatically grouped and organized by similarity. As a result of consolidation and grouping, informal lightweight ontologies gradually emerge in a bottom-up way. A system called StYLiD has been implemented to realize the proposed approach.

The thesis also proposes a decentralized approach for disseminating structured data in communities. Relevant information can be aggregated through socially linked sources. This has been demonstrated experimentally. By combining the capabilities of publishing and aggregating, proper flow of information can be maintained in the community. A semantic blogging system called SocioBiblog has been implemented to demonstrate this for the bibliographic domain.

Experimental evaluations have been done to test the usability of StYLiD. Experimental studies have also been done to observe the multiple conceptualizations done by people and to verify that such conceptualizations can be consolidated. Methods used for concept consolidation and grouping have also been experimentally tested with some real data. The applicability and significance of the proposed approach has also been demonstrated by some real practical applications.

List of Figures

List of Tables

Table of Contents

1. Introduction

1.1 Background

Information has become very valuable as the world is moving towards globalization. Today, information is power. However, information has to be shared to be utilized to its full potential. Information cannot be truly utilized when it is hoarded or locked up at a place. People should be able to obtain and use the information they are seeking for. On the other hand, people should be able to disseminate the information they can provide. To share information, we should we able to express it properly so that it can be understood and make it available to people who need it or who can utilize it. We should be able to have right information at the right place. When information pieces collected from different sources fit together it can form valuable knowledge. The significance of information sharing has been deeply realized by the United States after the 9/11 attacks. The United States Intelligence Community, Information Sharing Strategy (2008, February 22) established thereafter states that,

"*The need to share information became an imperative to protect our Nation in the aftermath of the 9/11 attacks on our homeland … Each intelligence agency has its own networks and data repositories that make it very difficult to piece together facts and suppositions that, in the aggregate, could provide warning of the intentions of our adversaries. The inability or unwillingness to share information was recognized as an Intelligence Community weakness by both the 9/11 Commission and the Weapons of Mass Destruction (WMD) Commission…*".

Information sharing comprises the following three main aspects.

1. *Information Publishing*. People should be able to express, represent and publish the information they have and want to provide. Proper mechanisms and medium should be provided to enable to people to publish information.
2. *Information Semantics*. For successful information sharing, it is also very important that the semantics, or meaning, of the published information is understandable to the consumers of the information. The semantics intended by the publisher should correspond to the semantics perceived by the consumer. The representation of the information should be well-defined and usable for necessary operations.
3. *Information Dissemination and Access*. It is also important to make relevant information available to people or parties who need it. Information sharing may be desired between different people or organizations or different systems, located globally or within communities. Proper mechanisms should be in place which allows people to disseminate information to desired targets and obtain desired information from desired sources.

Information Sharing on the Web

Worldwide communication is possible today due to communication networks and the Internet making us globally connected. Taking advantage of this, the web has established itself as the most powerful global medium for information sharing via the Internet. The web provides a global platform for people to publish information they want to share. People can publish textual or multimedia contents in web pages and these can be easily understood and used by other people around the world. The web has become a huge global repository, one common place for people to publish and find any type of information. It caters a worldwide audience, across boundaries of organizations and countries. Unlike other applications on the internet like email, which can only serve limited targeted group of people at a time, information on the web can persist and continue to serve all people. Information shared online may be used by others in unexpected ways for useful applications. Moreover, the power of the web is in the fact that web pages are interlinked to form a global network which makes all information reachable simply by following the links.

However, the traditional web still does not completely solve all the problems of information sharing. Firstly, it was not easy for all to publish on the web. Publishing on the web required access to server infrastructure and technical knowledge. So the web became a one way medium with few publishers providing information and rest of the world simply using the information as consumers.

Secondly, people are getting overwhelmed by the huge volume of information available in the web. Humans cannot consume or process all the available information. Such huge volumes of data should be processed by machines to provide useful results for the people. It is challenging to retrieve exact desired data from the web. Although current search engines technologies have proved to be very useful, they are mainly based on text search returning a ranked list of relevant results. It still needs considerable human effort to sort out the desired information from these results. Furthermore, people need to look for information pro-actively knowing what they need or what would be useful to them. Relevant information does not come by itself.

Finally, it is difficult to express and publish all our knowledge as web documents such that it is understandable and usable. If we want the information to be processed by machines it should be published in formats understandable by machines. It is difficult to ensure that the intended meaning of the represented information is correctly understood or interpreted even by the humans. Everyone may have different ways of representing and perceiving information and knowledge.

Structured data and the Semantic Web

Different types of data can be modeled by structuring them systematically, representing different parts and the relations between them. The Semantic Web (Berner-Lee et al., 2001) envisions creating a web of such structured data and providing well-defined meaning to the pieces of structured data. In the Semantic Web, knowledge is modeled using *ontologies* which explicitly represent conceptualizations of things in the real world. Information can be structured and shared using such ontologies. The Semantic Web with structured data offers solutions for information sharing overcoming some limitations of the traditional web.

- It provides the mechanisms to model different types of information systematically and publish them over the existing web infrastructure.
- Structuring makes it easy to define the semantics of data so that it can be machine understandable and hence processing can be automated.
- Information with its intended meaning can be communicated among different parties by following standard formats or mapping different formats.

- Structured data from various sources can be easily integrated and mixed.
- Search and browsing can be more effective with structure and semantics.

However, there are some major challenges due to which the Semantic Web remains largely unrealized (Siorpaes and Hepp, 2007a; Van Damme et al., 2007; Hepp 2007).

- Semantic Web technologies are too complicated for ordinary people and it is difficult to have people publish structured data for the Semantic Web.
- Ontology building is a difficult process and, hence, there are not many ontologies needed to cover all the data people may want to share.
- Ontologies are difficult to understand and use.

The Social Web

The social web is the recent generation of online applications and services that allow people to participate, interact and contribute freely on the web. The social web has leaded us into the new generation of web often called Web 2.0 (O'Reilly, 2005). It has advanced the web along the following aspects for information sharing.

Easy Publishing. Publishing on the web has become very easy and dynamic due to social platforms like blogs and wikis. Today anyone one can publish on the web unlike the traditional web scenario. Now people have more freedom to express their information in their own way. Thus, publishing has become more democratic with the social web.

Connecting People. The social web has provided technologies, like online social networks, that effectively connect people for information sharing. Information can be disseminated to desired parties and relevant information can be obtained from social circles. Online communities facilitate a new way of communication.

Collaboration. Social web applications, like wikis and online communities, enable collaboration among people. Collaboration can help in establishing consensus or common understanding required for meaningful information sharing.

Social web applications are easy to understand and use for ordinary people. People can socialize and enjoy on the social web. Therefore, the social web has proven to be very successful in drawing mass participation and it is exploding with user-generated contents. However, social web also faces some major challenges and still leaves many problems of the traditional web unsolved.

- Social web data is usually unstructured and the semantics is not defined for machines. So it cannot be processed automatically.
- It is difficult for different systems to share information and interoperate due to the lack of standard formats.
- It is still difficult to search and browse desired contents due to lack of semantic structure.

1.2 Current Limitations and Needs

Social web technologies have become a part of today's life and modern culture. Web applications are no longer just for the IT-experts. Easy and interactive interfaces are now successfully entertaining ordinary people from any background. Web has become a democratic publishing platform for information sharing among many-to-many. Information exchange on the web has become a social activity for everyone. Businesses are also utilizing this new trend of web 2.0 applications to enable better communication, collaboration and outreach. However, people and organizations still have many requirements that are not being addressed by the current technologies. The explosive growth of contents on the social web has further increased the necessity to addresses these issues. The current trend of new web applications has introduced new possibilities as well as new challenges. Some of these rising needs and challenges are as follows.

1. *Effective processing and retrieval*. Huge volumes of data can be obtained through mass contribution. But it becomes very difficult to process and analyze the data because the data is mostly in the form of unstructured text or multimedia. Even personal information collections become too big in the course of time. Mechanisms like tagging, keyword search and natural language processing can help to some extent to retrieve relevant information. However, when we need to do some more complex processing or analysis, for e.g., if we need to sort, filter or aggregate data by different dimensions or analyze the data from different views, it cannot be done directly. A lot of tedious manual work would be needed to handle such unstructured data from the web although we have excellent search engines and tagged data. Providing some structure to the data can help in overcoming this challenge. With the structure, people would have tables of data which can be sliced and diced as necessary for desired purposes. Desired information can be filtered and retrieved by various criteria along different dimensions. Analysis of the data would become convenient and this capability would definitely be valuable for many.

2. *Automation and useful applications*. If the semantics of the structure is defined, various automated operations over the data would become possible. Semantically structured data can prove to be very useful for people and organizations. People have always wanted computers to do useful things for them. They need applications that can solve their problems. When using any new system, people are most easily convinced by some instant visible benefit. Social web applications have been quite successful in this and many times people just want some fun with web applications. However, there are greater possibilities that people are not aware of and do not demand explicitly. New web applications should show people the additional unforeseen possibilities and enhance their experience. Applications have to prove the value of semantically structured data to the people.

Semantic Web technologies have already demonstrated the potential in targeted domains like life sciences and biology. In the future, Semantic Web technologies may even help to solve big problems like finding cures to diseases because such problems can be tackled effectively with the analysis of volumes of various types of complex data. Recently, big players in the web industry like Google and Yahoo are also getting in to utilize these technologies and provide useful services to the public. Yahoo's Search Monkey platform[1] enhances the Yahoo search results presentation by utilizing

[1] http://developer.yahoo.com/searchmonkey/

embedded structured data. It encourages developers to build applications to exploit the structured data and also encourages the information providers to embed structure to realize the full potential of their data. Rich Snippets[2] introduced by Google also provides similar capability to enhance search results. The Google Squared [3] application provides structured data in a table layout that can be manipulated flexibly.

3. Interoperation. One major difficulty all people are facing today, regarding social web applications, is that of interoperability. Social web applications collect a lot of data from people and keep them entertained within the application. But these become like walled data gardens or isolated data islands. People cannot move their data from one application to another. If a new social networking service is introduced people cannot move their friends list and profile to it. Also people cannot reuse the same data across multiple applications without duplication. This problem is distinctly being realized by both the users and online service providers. Some proprietary formats and APIs like OpenSocial[4] and Facebook Connect[5] are also coming up in the bid to become the standard for social networking data. However, we need more open and widely acceptable solutions covering wider range of contents.

4. Integration. As pointed out earlier in the background, when pieces of data from multiple sources are integrated, valuable knowledge can emerge. Integration of data provides greater value to people than when the data are kept separate. Currently, we cannot easily integrate data from various online sources. Similarly, we should be able to search data across different sources though a single interface. Data integration is an old problem and solutions have also been proposed, mainly for databases. However, the problem still remains, especially in the decentralized scenario of the World Wide Web. Currently, there are no straight forward mechanisms to combine data from multiple social web applications.

All the above requirements and challenges can be addressed by effective introduction of semantically structured data. However, while doing so, the advantages of simplistic social web applications should also not be undermined. Online applications should continue to be easy to use and require minimum learning. Also the freedom offered by social applications to the people should be maintained to ensure mass contribution. Powerful technologies tend to be more complex and constraining. Hence, it is challenging to introduce powerful Semantic Web technologies while maintaining the popular characteristics of current social web applications.

1.3 The Social Semantic Web

A promising direction to address the challenges discussed above is the social Semantic Web. The Semantic Web and the social web can complement each other because the weakness of one can be addressed by the strength of the other (Ankolekar et al., 2007; Gruber, 2008; Schaffert, 2006b). Social web applications provide easy-to-use platforms for ordinary people motivating them to share data in the community. The social web also enables collaboration and harvests collective intelligence which is necessary for establishing common understanding and shared models needed for the

[2] http://googlewebmastercentral.blogspot.com/2009/05/introducing-rich-snippets.html

[3] http://www.google.com/squared

[4] http://code.google.com/apis/opensocial/

[5] http://developers.facebook.com/connect.php

Semantic Web. On the other hand, the Semantic Web can provide semantic structure to social data and enable interoperation and information sharing among social web applications. The wide range of both Semantic Web and social web technologies results into wider range of possibilities for their combinations. The combination of these two trends can form a social Semantic Web which has emerged as a promising area for research and applications. The Semantic Web is heading towards practical realization along with real world social applications. This is leading us to the next generation of the web and people have even started calling it Web 3.0 (Hendler, 2008; Breslin et al., in press).

1.3.1 Some open problems

Although the integration of the social web and the Semantic Web offers great potential, it also poses several important challenges. While social web applications can provide easy interfaces, data contributed freely by the users may be imperfect for the Semantic Web meant for machines. We need more tolerant mechanisms to handle inconsistencies and inaccuracies that result from the informal approach of the social web (Schaffert, 2006b). On the other hand, while the Semantic Web can provide well-structured data, the complexity and structural constraints can degrade the usability of the social web application. Some general challenges for the social Semantic Web combination are as follows.

1. Obtaining structured data from the people. Ordinary people can only understand simple interfaces as offered by social web applications. They are only used to posting simple data and contribute data freely as they like. They may not be able to contribute complex structured data. Therefore, it is challenging to keep the interface easy for ordinary users and have them contribute structured data. Ontologies are needed to structure and organize data and provide well-defined meaning. However, we cannot expect ordinary people to understand about Semantic Web technologies and ontologies. On the other hand, if we allow people to freely contribute unstructured data it would be difficult to derive proper structure and semantics.

2. Collaborative ontology creation. Different people need to share different types of data. We would need various ontologies to model the different types of data. If we cannot find appropriate ontologies, new ones have to be created. To have common ontologies for information sharing, they should satisfy the requirements of different people. To ensure this, ontology engineering should be a highly collaborative process (Siorpaes and Hepp, 2007a). Social web platforms can facilitate collaboration among people. However, ontology creation is known to be a very difficult process. It would be challenging to keep the process simple and gain participation from the people and on the other hand ensure the creation of useful ontologies. Moreover, people have different perspectives. So building consensus among people may be difficult.

3. Motivation and useful applications. Another important challenge for having social participation to produce structured contents or build ontologies collaboratively is how to motivate the people. How do we ensure that the people will contribute or participate? We need to provide benefits to the people in return, especially today when web users are becoming more impatient and selfish (Nielsen, 2008). We need to prove the value of structured data and Semantic Web technologies through useful practical applications. It is important to provide services to search and utilize the structured data effectively. Search and browsing become powerful with structured data providing exact answers. Besides the end users, we should also motivate developers and business entrepreneurs and convince them to introduce the power of Semantic Web technologies into their applications for the public. The significance of the combination of social and Semantic Web technologies needs to be demonstrated to the industry.

4. Structured information dissemination. Besides producing structured data, it is also important to facilitate proper dissemination of the structured data in online communities. Usually social applications are only designed for exchanging unstructured information or information with limited structure. Therefore, we need additional mechanisms to transport structured data. Furthermore, it would be desirable to have a decentralized mechanism for such information sharing because the web is a decentralized platform with many different systems distributed worldwide.

5. Interoperable standards. For information sharing among distributed systems, interoperability is crucial. Usually existing social websites and information systems are closed confining the data within themselves. Every organization or information source maintains its own information models and formats, own ways of organizing the information and own systems. Interoperability is necessary for exchange and integration of information from different sources. Semantic Web technologies can help in establishing standards and the basis for interoperability. However, bringing different parties to common understanding, establishing interoperable standards and having different systems and organizations follow these is challenging.

6. Reuse of existing contents. There is already a huge amount of data in the existing web and it is growing rapidly with user-generated contents. A lot of digital contents are also available off the web and in users' desktops. It would be wise to utilize reuse these existing contents, add meaningful structure to them and bring them to the Semantic Web. This may be more effective than producing all new structured data from scratch. Hence, a potential direction is how to create structured data for the Semantic Web from the existing social web contents.

7. Compatibility. Although new semantic technologies are introduced, the existing web technologies, social applications, database-driven systems should be retained. People will not be willing to replace well-established popular technologies with nascent Semantic Web technologies. Moreover, it is better to reuse and build upon the existing technologies rather than reinventing the wheel. A major challenge is how to introduce the new semantic capabilities into existing systems and technologies without replacing them or destroying their usual aspects. It is important to be compatible with the existing technologies to coexist and cooperate with them. Therefore, reusing existing technologies and having compatibility among existing social systems, web technologies and new Semantic technologies is an important issue.

1.4 Scope of the Thesis

As described above, the area of information sharing in the social Semantic Web poses many challenging research problems. The thesis mainly focuses on and contributes to some of these problems as follows. However, the other issues are also considered while proposing solutions to these problems.

1. Obtaining structured data from the people. A major focus of the thesis is to obtain structured data for the Semantic Web from the ordinary people with the help of social web applications. The thesis aims to enable ordinary people to produce new structured data. However, some ways to reuse existing contents are also pointed out.

2. Collaborative ontology creation. The aim of the thesis is to enable people to share a wide variety of structured data. To model the structure of different types of data, we need to facilitate collaborative creation of new concepts, the building blocks for ontologies. Ontologies also serve to organize the data and concepts. Hence, collaborative creation of ontology for information sharing is considered.

3. Structured information dissemination. Finally, the thesis also explores ways to disseminate structured information in the community. Interoperability and compatibility with existing systems are important issues to be considered for this.

The thesis also considers aspects for motivation while proposing the solutions and attempts to demonstrate useful practical applications. The question of interoperability also arises while creating ontologies and producing structured data. Practical issues like reusing existing technologies and maintaining compatibility with existing systems are also considered while proposing new solutions and implementations.

1.5 Objectives

In order to address the above mentioned agenda, the main objectives of the thesis have been set as follows.

1. To study the ways of combining social web and Semantic Web technologies for structured information sharing, identify specific issues and propose new solutions for the following.
    a. To enable ordinary people to produce structured data.
    b. To enable formation of ontologies by collaborative effort of people.
    c. To enable dissemination of information in communities.
2. To implement working systems to realize the proposed solutions.
3. To demonstrate practical applications of the implemented systems.
4. To evaluate the proposed solutions and implementations.

1.6 Thesis Outline

The remainder of the thesis has been organized into the following chapters.

Chapter 2. The Social Semantic Web. In this chapter, necessary background knowledge and literature is presented. This includes details about the Semantic Web, structured data, ontologies, different types of ontologies and existing Semantic Web technologies. The social web is also discussed in some details. Some available ways for information dissemination in communities are also mentioned. Then, the social Semantic Web is presented along with some challenges in combining the two worlds. A detailed literature review about works on sharing structured data on the social Semantic Web is presented. Finally, some specific limitations of the state-of-art in structured data sharing in the social Semantic Web are summarized.

Chapter 3. Sharing concepts and structured data. In this chapter, first, the notion of concepts and their nature are explained. It is pointed out that concepts are essentially vague and cannot be defined uniquely. Cognitive theories about concepts are also discussed to support this. Hence, multiple conceptualizations may exist for the same thing. It is also pointed out that ways for integrating and mapping such conceptualizations exist. Based on these, an approach for authoring structured data and collaborative ontology creation is proposed. It enables people to create concepts freely and share different types of structured data. It proposes consolidation and grouping of concepts facilitating emergence of lightweight ontologies. A system called StYLiD implementing this approach is described in detail.

Chapter 4. Structured information dissemination in communities. This chapter discusses some ways of disseminating structured data in communities. The significance of sharing information through social links is demonstrated through an experimental study. An approach for decentralized sharing of structured data though social networks is proposed. An implemented system called SocioBiblog, for sharing of bibliographic information in communities, is described in detail.

Chapter 5. Evaluation and applications. This chapter shows some experimental evaluation of the approach proposed in chapter 3. Multiple experiments have been conducted and various observations have been made regarding different aspects of the proposed approach. Some real applications of the implemented system are also described. This includes a project about integrating research staff directories among different Japanese universities. Other social information sharing applications are also mentioned. An implemented system, called OntoBlog, is also described to show further possible applications of structured semantic data. Then, the proposed approach is compared with some existing approaches for collaborative creation of ontology and structured resources in the social Semantic Web. A discussion about the strengths and limitations of the proposed approach is also presented.

Chapter 6. Conclusions and future directions. Finally, conclusions are drawn from the entire study. Then, the future directions open for investigation are pointed out.

1.7 Contributions

The main original contributions of the thesis are as summarized below.

Social platform for structured data sharing. The thesis proposes to enable ordinary users to publish structured Semantic Web data through simple social software interface. StYLiD has been implemented as an online social platform that enables people to share a wide variety of data in the community. Users may freely define their own concept schemas and share different types of structured data on the Semantic Web. Other semantic blogging platforms, SocioBiblog and OntoBlog, have also been implemented which enable structured data publication through blogs.

Multiple conceptualizations. The thesis proposes allowing different people to have multiple conceptualizations over the same thing, rather than attempting to build consensus over a single common conceptualization. It is proposed to allow multiple conceptualizations to co-exist and still enable information sharing across them.

Concept consolidation. The thesis proposes an approach for consolidating multiple conceptualizations by mapping and linking concept schemas. A theoretical formalization of concept consolidation is presented. Concept consolidation is proposed as a new approach for building up conceptualizations from the community. This is a loose collaborative approach requiring minimum understanding and allowing different parties to maintain individual perspectives.

Emergence of lightweight ontologies. Besides community-based formation of conceptualizations by consolidation, in the proposed approach, concepts can evolve and gradually emerge with popularity. Further, similar concept schemas can be grouped and organized semi-automatically. Together these processes enable the emergence of informal lightweight ontologies.

Structured information dissemination in decentralized social networks. An approach for sharing of structured information though social networks in a decentralized environment is proposed and implemented as the SocioBiblog system.

2. The Social Semantic Web

2.1 The Semantic Web and Structured Data

The Semantic Web was originally envisioned by Sir Tim Berners-Lee, the inventor of the Web. A popular definition of the Semantic Web states that "*The Semantic Web is an extension of the current web in which information is given well-defined meaning, better enabling computers and people to work in cooperation*" (Berners-Lee et al., 2001).

The following aspects are important to understand the Semantic Web and the above definition.

A web of data. The current World Wide Web is a web of documents interlinked by hyperlinks. These web documents can only be understood by human. The actual data in the documents cannot be understood by machines as such. The Semantic Web aims to provide well-defined structure and meaning to the data so that even machines would be able to understand the data, process them and provide useful applications. The Semantic Web is a *"Web of data"*. Pieces of well-defined data are interlinked to form a global web, as an extension to the current web of documents, using the same basic technologies and infrastructure. Berners-Lee, in his blog post[6], has even proposed calling this global graph of data as the *Giant Global Graph* (GGG, in the same fashion as WWW).

Data modeling and knowledge representation. The Semantic Web provides the languages for modeling and representing data about real world objects, in formats suitable for computers. Modeling data with well-defined structure provides the basis for assigning machine understandable meaning or *semantics* to the data. A specification called an *ontology* is usually created in a particular domain (area of interest) to model data for the Semantic Web.

Consensus and common formats. An ontology is usually created through consensus among different users. When common specifications are followed, data drawn from diverse sources can be integrated and processed homogeneously. Information exchange and interoperation between systems become possible. Consensual specifications can be widely adopted and useful applications would be developed over the structured data following these common formats. Thus, the Semantic Web also aims to provide common formats for data.

2.1.1 Ontologies

"*An ontology is an explicit specification of a conceptualization*" - Gruber (1993).

This is one of the most commonly cited definitions of an ontology. Here, conceptualization means the modeling of the objects, concepts, and entities that exist in the area of interest and the relationships that hold among them. Gruber's notion of conceptualization is basically extensional as it depends on the state of objects in the real world. Guarino (1998) has refined this definition of ontology, emphasizing the *intension* of conceptualization, as follows.

[6] http://dig.csail.mit.edu/breadcrumbs/node/215

"*An ontology is a logical theory accounting for the* ***intended meaning*** *of a formal vocabulary, i.e. its* ***ontological commitment*** *to a particular* ***conceptualization*** *of the world. The intended models of a logical language using such a vocabulary are constrained by its ontological commitment. An ontology indirectly reflects this commitment (and the underlying conceptualization) by approximating these intended models.*"

According to Guarino, ontologies are only approximate specifications of conceptualizations. Guarino stresses that an *intensional* account of the notion of conceptualization has to be introduced, which gives the intended meaning of the conceptualization independent of any particular state of affairs.

Classification of ontologies

There are various types of ontologies differing in multiple aspects. Schaffert et al.(2005) have classified ontologies along three dimensions - model scope, level of expressiveness and model acceptance. The model scope refers to the area or coverage that is of interest. The acceptance dimension deals with the target communities of the application and its knowledge model and various methods of building consensus within a specific community. The level of expressiveness is particularly significant and is briefly described below.

Level of expressiveness (Light-weight and Heavy-weight ontologies)

The spectrum of expressiveness of ontologies as defined by Corcho et al. is illustrated in the Figure 1 below (as cited in Schaffert et al., 2005, p. 7)

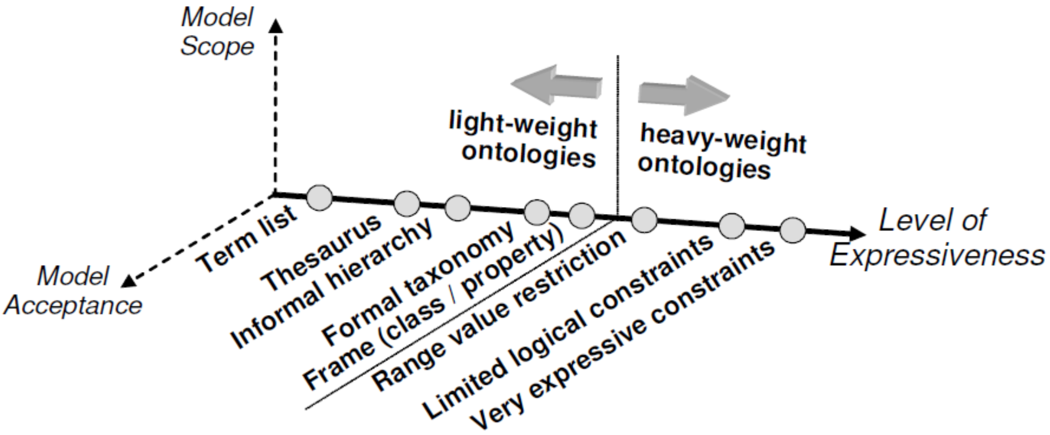


Figure 1. Level of expressiveness of ontologies.

(source: Schaffert et al., 2005, p. 7)

Corcho et al. distinguish between the two main groups – **light-weight ontologies** and **heavy-weight ontologies** – and define eight sub categories based on their level of expressiveness.

1. A term list or controlled vocabulary contains a list of keywords. Such lists are typically used to restrict possible values for properties of some kind of instance data in the domain.
2. A thesaurus also defines relations between terms, e.g. proximity of terms.

3. An informal taxonomy defines an explicit hierarchy of generalization and specialization, but there is no strict inheritance, i.e. an instance of a sub-class is not necessarily also an instance of the super-class.
4. A formal taxonomy defines a strict inheritance hierarchy.
5. A frame or class/property based ontology is similar to object-oriented models. A class is defined by its position in the subclass hierarchy and its properties. Properties are inherited by sub-classes and realized in instances.
6. A range value restriction defines, in addition, restrictions for the defined properties. The restrictions may be data type or domain restrictions.
7. By using logic constraints, property values may be further restricted.
8. Very expressive ontology languages often use first-order logic constraints. These constraints may include disjoint classes, disjoint coverings, inverse relationships, part-whole relationships, etc.

Significance of lightweight ontologies

With heavy semantics, powerful reasoning can be done and successful applications have been demonstrated in enterprise scales. However, such systems cannot tolerate any inconsistency. On the other hand, with lightweight ontologies not much reasoning can be done. However, there is far less risk of inconsistencies because only little ontological agreements are in place. With little semantics, applications can scale very well. This is a significant aspect when we consider the huge scale of the web which is important for the practical realization of the Semantic Web vision. Therefore, lightweight ontologies have become more popular and widespread. A popular quote by Jim Hendler[7] puts it as "*A little semantics goes a long way*".

2.1.2 Benefits of structured data and semantics

As already pointed out in the introduction, structured data and semantics have significant advantages. Some are listed below (Bergman, 2007; Iskold, 2007).

- Semantics of data can be well-defined so that processing can be automated. The Semantic Web would provide a vast amount of openly available interlinked data that can be processed automatically by machines. A wide range of intelligent applications would be possible using well-defined data and standards.
- Information exchange becomes effective following common formats.
- Data from various sources can be easily integrated.
- Interoperability between systems becomes possible with standard formats or mapping different formats.
- Online information search and browsing would become more effective and precise with well-defined semantics and powerful Semantic Web technologies.

The global knowledge base represented using ontologies may be utilized to realize unprecedented powerful applications. The potential of Semantic Web technologies

[7] http://www.cs.rpi.edu/~hendler/LittleSemanticsWeb.html

and having structured data on the web is being realized by organizations and enterprises and these technologies are gradually being embraced by the industry for large scale applications (Provost, 2008).

2.1.3 Challenges for structured data and the Semantic Web

However, the Semantic Web is facing some big challenges for its widespread practical adoption. Firstly, Semantic Web technologies have been difficult to understand for ordinary people. Most of the tools and services have been complex and difficult to use. A detailed discussion about the usability issues of semantic web technologies can be found in (Di Maio, 2008). The paper also proposes some possible ways of incorporating usability factors in development of semantic web applications. It further points out that usability also lies in usefulness, which is the ability to satisfy real user needs. While showcasing very powerful capabilities Semantic Web technologies have been lagging behind in addressing the real user needs. Secondly, ontology creation is a very difficult process and it is not easy for people to arrive to consensus. Finally, the Semantic Web is lacking enough data and applications. People will not be motivated to contribute Semantic Web data unless there are useful and interesting applications. However, it is difficult to demonstrate interesting applications without enough data in the first place. This deadlocking problem is also being known as the "c*hicken and egg problem*" of the Semantic Web (Hendler, 2008).

Hepp (2007) has identified 4 bottlenecks for ontology creation as follows.

- *Conceptual dynamics*. The real world is a dynamic place with changes occurring all the time. Conceptualizations should be updated rapidly to match these changes. Otherwise the ontologies built upon the conceptualizations will become invalid. Moreover, the understanding of people of the real world and the accuracy of conceptualization improve along with time. Ontology engineering is faced with the challenge to meet this dynamics.

- *Economic incentive*. Building ontologies require significant effort and resources. If the benefit of this investment is not apparent or not enough, people will not be motivated to create ontologies. Moreover, an immediate tangible benefit is necessary to convince people of the potential of these new technologies.

- *Ontology perspicuity*. There if often a gap between the creators and users of an ontology. The individuals have to use the ontology may not easily grasp the meaning of all the elements as intended by the ontology creators. Keeping the ontology understandable for the non-technical people or stakeholders is a major challenge for ontology development.

- *Intellectual property rights*. Various copyright and patent issues also hinder the re-use and hence rapid development of ontologies from existing works and resources.

2.1.4 Semantic Web technologies

The Semantic Web is a big vision as a whole. Nevertheless, it is gradually being realized. The Semantic Web technologies have been represented in a stack often called the “Semantic Web cake” or “Semantic Web stack” as shown in Figure 2 and Figure 3. The stack shows the layers of technologies required to realize the full Semantic Web vision. The bottom layers of the stack have been fully realized (at least upto RDF + rdfschema). The Ontology vocabulary layer has been partly realized and is actively being developed. The upper layers are still not quite mature at the web scale though these have been deployed within local or enterprise levels. However, the Semantic Web stack is itself evolving frequently along with new technologies, research and practical challenges coming to the scene.

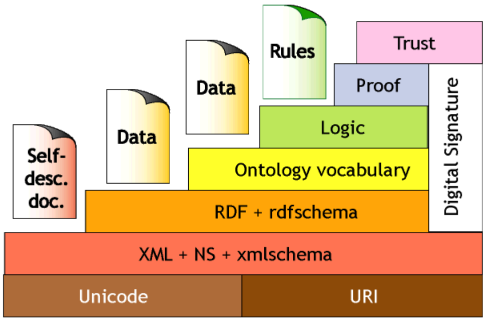


Figure 2. The Semantic Web stack.

(Source: Berners-Lee, 2000)

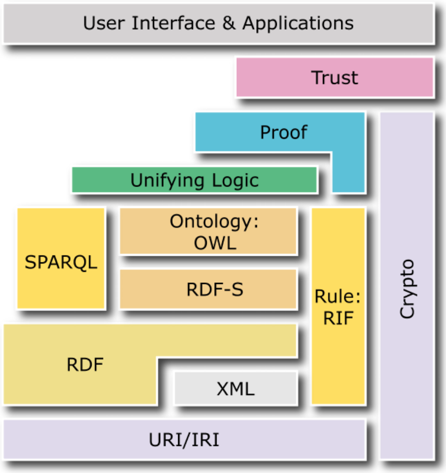


Figure 3. A more recent version of the Semantic Web stack.

(Source: Bratt, 2007)

Existing technologies. The basic Semantic Web technologies and frameworks are quite well-established by now. Just as web documents are identified and interlinked by URLs, data resources are identified and interlinked by URIs (Uniform Resource Identifiers) in the Semantic Web. RDF (Resource Description Format)[8] has become the standard language used to describe data for the Semantic Web. With the RDF model, all infromation is represented as (subject, predicate, object) triples, also known as RDF triples. There are various syntactic formats to represent RDF for e.g., RDF/XML, N3, N-triples, turtle, etc. Formats like Microformats[9] and RDFa[10] have

[8] http://www.w3.org/RDF/

[9] http://microformats.org/

also become popular because they can be embedded within existing web pages. RDFa is a set of extensions that enable us to express RDF inside XHTML elements. OWL (Web Ontology Language)[11] has become the standard for representing Semantic Web ontologies. Similarly, SPARQL [12] has become the standard for querying in the Semantic Web. Many ontologies have been created for different information domains. Semantic Web techonologies are successfully being used in many industrial applications (Provost, 2008).

2.1.5 Linked Data

The term Linked Data was coined by Sir Tim Berners-Lee in his Linked Data Web architecture note (Berners-Lee, 2006). Linked data is a method of exposing, sharing and connecting data on the Semantic Web. It provides the mechanisms for publishing and interlinking structured data into a Web of Data. Linked Data is about using the web to connect related data. The Semantic Web is not just about putting data on the web. It is about making links, so that a person or machine can explore the web of data. Berners-Lee outlined the following four rules of Linked Data in his design issues notes.

1. Use URIs to identify things that you expose to the Web as resources.
2. Use HTTP URIs so that people can locate and look up (dereference) these things.
3. Provide useful information about the resource when its URI is dereferenced.
4. Include links to other, related URIs in the exposed data as a means of improving information discovery on the web.

The practice emphasizes web access to data using existing web technologies such as URIs and HTTP so that we can inherit all the HTTP mechanisms already in place. The standard web transfer protocol, HTTP, should be used to be “on the web” while being “in the web” of data.

In the web of documents, people mostly publish unstructured documents and interlink those using hyperlinks. Linked data shifts the paradigm from document publishing to *data publishing* and from hyperlinking to *data-linking*.

Linked data is about making data available in standard ways so that others can use and link to. This is essential to connect the data we have into a global web. Due to network effect, usefulness of data increases the more it is linked with other data. This forms a data commons where people and organizations can post and consume data about anything. This common data network is often called the Web of Data. The unexpected re-use of information is the value added by the web. Organizations benefit by being in this global data network, accessible to both people and machines. Organizations can achieve more through sharing their data and collaborating than being closed and isolated in islands. The linked data web offers unbound global commercial opportunities for enterprises and entrepreneurs. Interestingly, linked data can be fully realized with existing technologies maintaining compatibility with legacy applications while exposing data from them. Thus, linked data is a significant

[10] http://www.w3.org/TR/xhtml-rdfa-primer/
[11] http://www.w3.org/TR/owl-features/
[12] http://www.w3.org/TR/rdf-sparql-query/

practical movement towards the vision of the Semantic Web (Berners-Lee, 2006; Bizer et al., 2007a).

Bizer et al. (2007a) have provided some guidelines about how to publish linked data on the web. Some guidelines have also been set up to create good URIs, so called Cool URIs[13], regarding simplicity, stability and manageability of the URIs. These guidelines are easy to implement and provide a well-defined way to expose data to the open linked data web.

RDF description for a URI

It is recommended that the following information be returned when the URI for a resource is dereferenced.

1. *The description*: all triples from the dataset that have the resource's URI as the subject.
2. *Backlinks*: all triples from the dataset that have the resource's URI as the object.
3. *Related descriptions*: additional information about related resources that may be of interest in typical usage scenarios.
4. *Metadata*: such as a URI identifying the author and licensing information.

The data source should at least provide RDF descriptions as RDF/XML. Links among structured data elements are made using RDF links or predicates. Usually, the application domain will determine which RDF properties are used as predicates. Terms from well-known vocabularies should be re-used wherever possible. An extensive list of well-known vocabularies is maintained by the W3C SWEO Linking Open Data community project[14].

Many organizations are information sources are actually opening up their data online as linked data forming the so called Linked Open Data cloud[15] which is growing rapidly. However, currently data is mostly being linked at the instance level only. Jhingran (2008) rightly pointed out that data should also be linked at the schema level and through communities of people. The thesis explores these directions for producing linked data.

[13] http://www.w3.org/TR/2007/WD-cooluris-20071217/

[14] http://esw.w3.org/topic/SweoIG/TaskForces/CommunityProjects/LinkingOpenData

[15] http://linkeddata.org/

2.2 Social Web and Web 2.0

The social web is a recent phenomenon observed in the current web. It constitutes all the online applications and services that allow people to participate and contribute to the web. It serves as a platform for people to connect with each other, socialize and interact. This helps in bringing together related people or people sharing similar interests. The social web facilitates people to share information easily and freely. It is driven by user-generated contents. Contributions from millions of worldwide users results into vast up-to-date collection of resources although little effort is needed in the part of each individual contributor. This phenomenon driven by the *power of people*, or *collective intelligence*, is also known as *wisdom of the crowds*. The social web is the most distinguishing aspect of the new generation web being called Web 2.0 (O'Reilly, 2005). Some common social web applications are

- Blogs (e.g., Wordpress[16], Blogger[17], etc.)
- Wikis (e.g., Wikipedia[18])
- Social bookmarking (e.g., Delicious[19], Digg[20], etc.)
- Multimedia sharing sites (e.g., YouTube[21], Flickr[22], Last.fm[23], etc.)
- Tagging (incorporated in most social websites)
- Social networking (e.g., Facebook[24], MySpace[25], etc.)

2.2.1 Information dissemination in the social web

The social web is an excellent network for dissemination of information. Almost all people in the world are connected by social links forming a global human network. In such a network, everyone is connected within few links. It is believed that any person in the world is acquainted with any other person in the world by not more than six steps. This is popularly known as the "six degrees of separation" and has been supported by empirical studies too. Hence, the huge human network actually shrinks into a small world where everyone can be reached easily within few steps. This is known as the "small worlds" phenomenon in the science of networks (Watts, 1999; Watts, 2003; Barabási, 2003). Therefore, the social web can serve very well in disseminating information online by establishing the human network online.

Information is usually disseminated in the social web by following 3 ways.

Centralized online publishing. The information can be published on social websites. This is a centralized solution. Everyone can access the information through the centralized site where the information is published.

[16] http://wordpress.com/
[17] https://www.blogger.com/
[18] http://www.wikipedia.org/
[19] http://delicious.com/
[20] http://digg.com/
[21] http://www.youtube.com/
[22] http://www.flickr.com/
[23] http://www.last.fm/
[24] http://www.facebook.com/
[25] http://www.myspace.com/

Information feeds. Social platforms like blogs also provide information feeds (in formats like RSS or ATOM). These feeds can be aggregated by other systems. Feed readers or new readers are popularly used to aggregate feeds from different sources. Therefore, this can serve as a decentralized way of sharing information among distributed sources. Further, special online services like Yahoo Pipes[26] and Dapper[27] enable users to aggregate, mix, filter and process such information feeds from various sources and share them with others. Yahoo Pipes provides an online visual editor to remix feeds and create data mashups easily.

Targeted information dissemination. Finally, social applications can disseminate information to targeted people, groups or parties. Social recommendation has become a common feature provided by many sites. The system automatically suggests related information items to the user which he/she may be interested in. Collaborative filtering (Goldberg, 1992) can help in producing such recommendations. Besides this, social web applications also help people to explicitly distribute information to targeted people, or simply notify of available information.

2.2.2 Benefits of the social web and challenges

The social web has enjoyed huge success because the technologies are easy to understand and use for any ordinary user. This results in wide participation. Further, people are enjoying these applications as they can identify themselves and their friends to the applications. However, these simple applications have some severe limitations. The data is usually in the form of unstructured text or multimedia files. These contents are understood by human but cannot be understood or processed automatically by machines. The semantics of the contents is unclear, even for human sometimes. The same tag may have different meanings for different people or the same thing may be tagged in different ways by different people. Hence, information processing and retrieval becomes difficult. Also, there is a lack of interoperability among the social websites. To enable exchange of information and interoperation among systems, the meaning of the information should be well-defined and understood by all the systems. Currently, the social websites are hoarding lot of data within themselves as closed *data silos*. These are *walled data gardens* that may appear beautiful to the users at first but confine the users within. To overcome this limitation we need well-defined standards for different types of data which can be adopted by online systems worldwide.

[26] http://pipes.yahoo.com/pipes/

[27] http://www.dapper.net/

2.3 The Social Semantic Web

The Semantic Web has been slow to realize its potentials due to the lack of mass participation and motivating applications, usable and understandable by ordinary people. The social web may complement the Semantic Web and bring it to practical realization. The social web provides platforms that can be easily understood and used by ordinary people. By facilitating interaction and collaboration among people on the social web, consensus can be achieved and standards can emerge. The social web phenomenon has motivated millions of users with simple applications and has successfully collected huge volumes of data from the users. The same phenomenon may bootstrap the Semantic Web with enough data and applications breaking the notorious "chicken or egg" cycle. The first step for Semantic Web is to have lots of data. Rationalization of data can be done later (Huynh, et al., 2007a).

On the other hand, the social web is running into several problems due to its unstructured nature and lack of semantic standards. The tons of user generated data cannot be understood and processed effectively by machines. Moreover, different systems cannot interoperate with each other because the semantics of data is not clear. The Semantic Web can provide well-defined structure to the data on the social web so that they can be processed by machines. The Semantic Web can also provide the standards needed for interoperability among online applications of the social web.

In this way, the social web and the Semantic Web and can complement each other to address the challenges both worlds are facing. The combination of social software with Semantic Web technologies has been gaining significant attention recently (Ankolekar et al., 2007; Gruber, 2008; Schaffert, 2006), entire books about it are coming up (Blumauer & Pellegrini, 2008; Breslin et al., in press) and there are a large number of works that try to enable ordinary users to produce Semantic Web contents by using social software. The resulting social Semantic Web can help in collaborative knowledge creation by facilitating mass participation and interaction. However, combining these two different cultures is not so easy. One is an unstructured world mainly for people and another is a structured world mainly for machines.

2.3.1 The structure chasm

The vast amount of data being sharing in the social web is basically unstructured whereas the Semantic Web requires data to be properly structured. Crossing this structure chasm is a major challenge for combining these two worlds. Halevy et al. (2003) have discussed some merits of the unstructured world (U-world) that the structured world (S-world) of data lacks. Following are some important differences which form a "chasm" between these worlds.

Authoring. Authoring data is straightforward in the U-world enabling rapid content creation. However, authoring is difficult in the S-world. The data needs to be modeled conceptually into a schema and entered conforming to the schema. Designing the schema is itself a difficult job. Therefore, the average user usually produces unstructured data rather than structured data.

Searching. In the U-world, search is simply done by keywords and these need not be exact. So the answers are also not exact. Multiple answers are returned ranked according to relevance. In the S-world, search is done through queries and formulating queries is more difficult. However, the results are exact. But no other answer except the exact answers is returned. Also precise knowledge of the schema is usually needed to be able to formulate queries.

Data sharing. Data sharing is easy in the U-world as all documents can be shared and searched in a uniform way. In the S-world, due to difficulties in authoring and querying and differences between schemas of different sources, sharing data also becomes challenging.

This chasm can be bridged to some extent by introducing the attractive properties of the U-world into the S-world. Halevy et al. (2003) have proposed the REVERE platform offering several mechanisms for crossing this structure chasm on the web. This includes the following 3 major components.

1. *The MANGROVE data structuring component*: MANGROVE (McDowell et al., 2003) provides a convenient tool for easily annotating existing unstructured data in web pages. The tool displays a rendered version of the HTML document alongside a tree view of the schema being used for annotation.

Instant gratification. MANGROVE provides a set of applications that entices people to author structured data by instant gratification in the form of some immediate visible personal benefit for the contributor. Some example services provided are - an online department schedule created dynamically based on the annotations made on different pages, etc., a departmental paper database, a “Who’s Who,” and an annotation-enabled search engine.

Deferral of integrity constraints. Users are not required to follow integrity constraints while annotating. This simplifies the process of annotating. The responsibility of cleaning up the data and enforcing constraints is passed to the application using the data as different applications have varying requirements for data integrity. This relaxation is necessary for the practical realization of large-scale distributed authoring.

2. *The Piazza peer-data management system*: Piazza is a data sharing environment, based on a peer data management system (PDMS). It tries to bridge the structure chasm by combining the ad hoc extensibility and distributed nature of the unstructured web with the rich semantics of database systems. It serves an ad hoc environment with peers maintaining different schemas to structure their own information. Semantic mappings between different schemas are made locally between peers. Using these semantic mappings transitively, peers can query data from the entire connected but distributed system. The number of semantic mappings needed is linear in the number of data sources. With this scheme, users can formulate queries over their own schema or preferred existing schema without having to learn a new global schema. The transitive closure of the mappings is used to query all the sources and the result is structured back in the user’s schema.

3. *Tools using statistics over structures*: Importing the information retrieval techniques of the U-world based on computing statistics over text corpora, Halevy et al. propose computing some useful statistics over the corpora of schemata and structured data. The corpora contain the schema information, queries, mappings between schemas, actual data and relevant metadata. Some basic statistics like the term usage in structured data, co-occurring schema elements and similar names are computed. Some composite statistics on frequently appearing partial structures may also be computed. Using the corpora and computed statistics, automatic assistance can

be provided for authoring, querying and sharing of structured data. Halevy et al. have proposed some tools for this as follows.

DesignAdvisor – This tool can assist the user while authoring data with MANGROVE. It can suggest more complete schemas to a user by returning a ranked list of similar schemas. This can also be used to propose extensions to a schema.

MatchingAdvisor – This is a semi-automatic tool for schema matching which assists in creation of semantic mappings between different schemas.

A user may not always know about the schema of the information sources to be queried. The user would prefer to pose the query in his own terms. In this case, the corpus can be useful to reformulate the user's query in terms of the existing schemas.

2.3.2 User motivation and incentives

As already indicated in the previous section, motivating the users through some incentives is necessary to cross the structure chasm. One the major bottlenecks for collaborative creation of structured data and ontologies is how to motivate the users (Hepp, 2007). Social software has been successful in motivating users by providing visible personal benefits. People should have some instant gratification. People have fun with social websites. Social applications feel like personal, about friends and social activities. There is active ongoing research on incentives for the Semantic Web. Siorpaes and Hepp (2007b) have tried to provide incentives, as enjoyable experience through online games. Hasen and Jameson (2008) have identified some factors that can affect user motivation including automatic algorithms, user interface, user input, affordances of situations and use of external resources. Often, a favorable combination of these factors achieves good results.

Contribution inequality

No matter how good motivation the system provides, it has been observed for online social systems that most of the contents are contributed by a very small percentage of users (Nielsen, 2006). 90% users are passive consumers, 9% contribute from time to time and only 1% are heavy contributors. Conversely, 90% of postings come from the 1% users, 10% come from the 9% users and there are no postings from 90% users. In spite of this inequality, the success of social applications has shown that it is even enough to motivate this small percentage of contributors to sustain the system.

A different approach for ensuring user participation is to provide solutions for targeted users who have specific requirements that can be met by the system. Then they would be self-motivated. Further, as mentioned earlier, the fact that the value of one's data increases when it is combined with others' in the network can motivate optimistic stakeholders. Jhingran (2008) highlighted that there needs to be a virtuous cycle of linked data and value creation which in turn produces more linked data and more value out of it.

2.4 Structured Data Production in the Social Semantic Web

In spite of the challenges, the combination of social and Semantic Web technologies is definitely promising and a lot of work has been done and are being done in this area. The combination is significant for the production of structured contents required for the practical realization of the Semantic Web. Hence, creation of structured contents in the social Semantic Web is a major focus of the thesis.

In the following text, the state-of-art will be described with significant works done in this area. There are several ways in which current approaches can be distinguished, as shown in Figure 4. It is not necessary that a system belongs to a particular category in the classification. Actually many works use multiple approaches together. The sources of structured data may be different. For e.g., data may come from the users, existing web pages, user's desktop, unstructured text, databases, etc. In some systems users actively contribute structured data. In other approaches, users continue to use the existing systems and semantic contents are derived from these indirectly without involving the users. Some systems only produce structured instance data while some produce concepts and ontologies too. Users may participate independently or collaboratively for content creation.

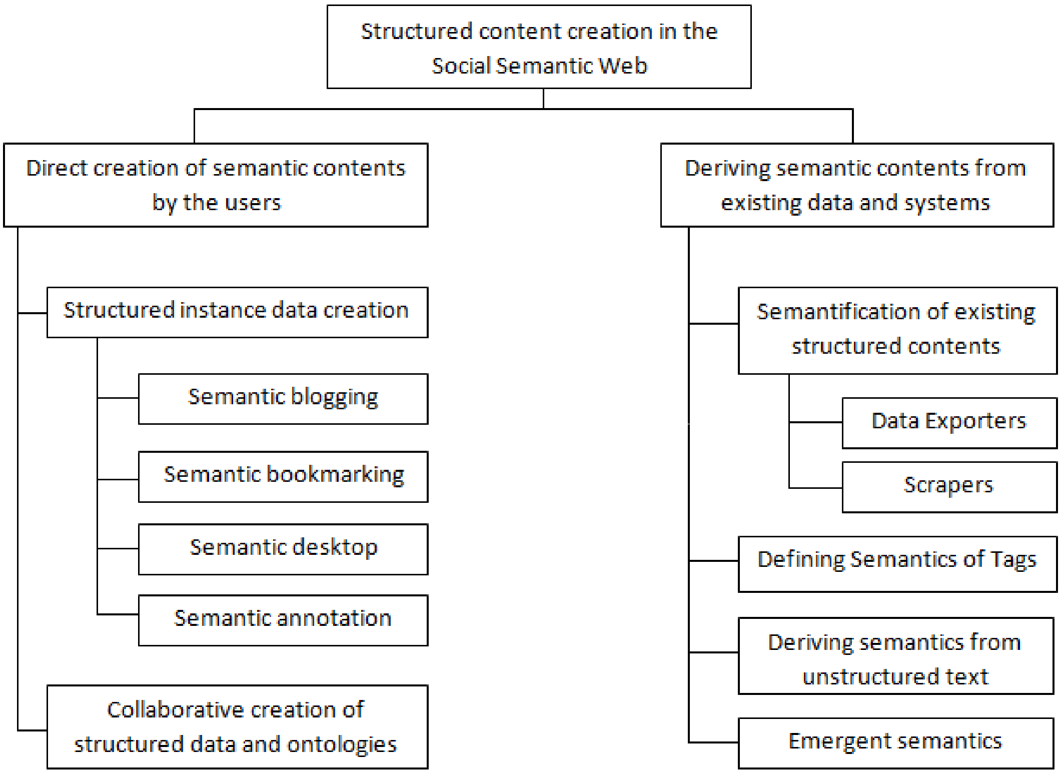

Structured content creation in the Social Semantic Web
Direct creation of semantic contents by the users
Structured instance data creation
Semantic blogging
Semantic bookmarking
Semantic desktop
Semantic annotation
Collaborative creation of structured data and ontologies
Deriving semantic contents from existing data and systems
Semantification of existing structured contents
Data Exporters
Scrapers
Defining Semantics of Tags
Deriving semantics from unstructured text
Emergent semantics

Figure 4. Classification of works on structured content creation in the social Semantic Web.

2.4.1 Direct creation of semantic contents by the users

In this broad category, the users explicitly create the semantically structured contents. This may further be classified into two groups based on the type of content created.

a. Structured instance data creation

In this category, the users directly contribute structured instance data only based on some existing ontology or concept schema. Usually the users contribute data independently without any collaborative effort with others in the community. However, the entire community benefits from the collection of individual personal contributions. There are several works in this category some of which are as follows.

i) Semantic Blogging

Blogs have made publishing information on the web very easy. Blogs serve as dynamic media showing the latest posted information. Blogs can effectively capture informal knowledge from several users and cater to the entire community. Conventional database driven information systems are rigid and do not cover all types of information that people may want to share within an organization or community. Informal snippets in blogs can cover a wide variety of information. Cayzer (2004a; 2004b) discusses elaborately why blogs are suitable for managing information snippets. However, traditionally blog entries do not have much structure and organization and cannot be processed effectively. Semantic blogging is a technology that builds upon blogging and enriches blog items with metadata (Cayzer, 2004a, b). Semantic blogging exploits the easy publishing paradigm of blogs and enhance them with semantic structure. It combines desirable features of both blogging and the Semantic Web. Blogging provides an easy platform for online publishing along with mechanisms likes RSS, comments and trackbacks. The semantic web can provide well-defined structure to information based on ontologies so that it can be processed by machines. This also enables interoperability between different systems and facilitates information exchange. Pieces of structured data in semantic blogs can be interlinked with semantic relations. This enables meaningful navigation and organization of related contents in blogs. Semantic blogging can extend blogging for decentralized informal knowledge management. Some works done in semantic blogging are as follows.

The *Semantic Blogging Demonstrator*[28] (Cayzer, 2004a, b) is a semantic blog for the bibliographic domain. Blog entries *contain* bibliographic items as metadata. Cayzer emphasizes the distinction between blog entries and information items. The demonstrator organizes blog entries within a category tree based on 'broader than/narrower than' relations, using the SKOS[29] vocabulary, to categorize blog entries. The demonstrator provides a category chooser functionality which works based on simple language processing. The demonstrator offers 3 main capabilities - semantic view, navigation and query.

Karger and Quan (2005) extended *Haystack*[30] (Quan et al., 2003) to enable users to view cross-blog reply graphs and track conversation in multiple blogs. RSS subscription facility is also provided. However, Haystack is too complex for a lightweight application like blogging.

[28] http://www.semanticblogging.org/semblog/blog/default/

[29] http://www.w3.org/2004/02/skos/

[30] http://haystack.lcs.mit.edu

Möller et al. have developed the *semiBlog* system (Möller & Decker, 2005; Möller et al., 2005, 2006). They identify two types of metadata, structural and content-related metadata, in blogging. *Structural metadata* deals with parts of a blog and relations between them. The SIOC ontology (Breslin et al., 2005) has been used for structural metadata. A WordPress SIOC plugin is used to expose SIOC metadata from the blog engine. *Content metadata* describes the posted content. FOAF, vCard, BibTex/SWRC, iCalendar, etc. have been used for content metadata. semiBlog emphasizes generating metadata by utilizing data on the user's desktop. It uploads content metadata, derived from the desktop, to an external service for publishing RDF. The structural and content metadata are integrated by providing the URLs of content metadata in *rdfs:seeAlso* statements in the structural metadata.

The *Semblog* platform (Ohmukai & Takeda, 2004) allows users to annotate content using their personal ontologies, using FOAF [31] (Friend of a Friend) metadata, syndicating this metadata over extended RSS. Both topic and social network information are thus available for information retrieval and recommendation. Structured blogging[32] also embeds machine readable information in blog entries using Microformats. We have developed the following two semantic blogging prototypes.

SocioBiblog (Shakya et al., 2007a, 2008b) facilitates sharing of bibliographic information in a social network. The SWRC (Semantic Web for Research Communities) ontology (Sure et al., 2005) is used for the bibliographic metadata. SocioBiblog aggregates publications from the socially linked sources by extending RSS to embed publication metadata. The system is described in detail in Section 4.3.

OntoBlog (Shakya et al., 2007b, 2008a) is a semantic blogging prototype which links blog entries to an existing ontology and instances. OntoBlog attempts to provide an integrated platform to facilitate publication, semantic annotation and information utilization. More descriptions of the system are provided in Section 2.4.1, sub-section for semantic annotation and in Section 5.6.5.

After observing the many works about semantic blogging by many researchers, Cayzer (2006) reviewed the history of semantic blogging, and discussed some promising future directions along with two experimental projects based on semantic blogging – BlogAccord for music blogging, and the Snippet Manager information integration portal. The snippet manager can merge disparate information sources demonstrating the potential of semantic blogs for enterprise information management. Cayzer points out that semantic blogs can further be extended for *concept mapping* whereby blog entries are associated with *ideas about an information item*. Semantic blogs can prove to be more useful when combined with social networks and folksonomies. This can enable the system to infer recommendations like related resources or related authors for the blog entries. On the other hand, data in the semantic blogs can also be mined for analyzing the social networks.

Recently, a semantic microblogging service has also been proposed by Passant et al. (2008). They have implemented SMOB as a distributed microblogging system for publishing and aggregating posts structured with Semantic Web ontologies, mainly SIOC and FOAF.

[31] http://www.foaf-project.org/
[32] http://structuredblogging.org/

ii) Semantic Bookmarking

Revyu (Heath & Motta, 2007) is a reviewing and rating site that allows people to share a wide variety of data by reviewing and rating anything. The system generates dereferenceable URIs for things, reviews, people and tags. Data items can easily be linked with other items using URIs to produce linked data. Revyu produces RDF output and provides a SPARQL endpoint for query. It also exposes reviews using hReview microformat embedded in XHTML. However, most concepts are modeled simply as things. The detailed structure of the information is not modeled and different things are not differentiated.

BibSonomy (Hotho et al., 2006) is a social bookmarking system for sharing bookmarks and publication references. Bibliographic metadata is provided in several structured formats including SWRC.

Twine[33] is a commercial online social application built upon Semantic Web technologies. It is a social site where users can bookmark contents from the web, keep track of their interests and connect to related people. The system provides automatic personalized recommendations about relevant online resources. Resources can be shared within various communities. It serves as a platform for leveraging and contributing to the collective intelligence of communities (Hendler, 2008).

iii) Semantic Desktop

A Semantic Desktop is a set of technologies that enables data in the user's desktop to be easily shared across different applications and different desktops. It brings the functionalities of Semantic Web technologies to the user's desktop and allows users to structure and organize their data, as semantic resources, according to their own preferences and contexts. Conversely, it exposes the data and personal models, locked in users' desktops, to the Semantic Web enabling them to be shared with the help of common ontologies. A semantic desktop can produce a lot of data for the Semantic Web by directly tapping in to the desktop which is the user's primary interface.

The term "Semantic Desktop" was coined by Stefan Decker and used by Sauermann (2003) in the Gnowsis Semantic Desktop research project (Sauermann, 2003; Sauermann et al., 2005). The goal of the Gnowsis project was to complement desktop applications and operating system with Semantic Web features. Sauermann et al. proposed identifying and representing desktop resources with URIs and integrating desktop data sources in a unified RDF graph. The primary focus was on Personal Information Management (PIM), enabling people to use their desktop computers like a personal semantic web. Documents on a user's desktop are related to their background, context and personal interests. This can be expressed as personal mental models through a semantic desktop. A framework called the Personal Information Model (PIMO) (Sauermann et al., 2007) has also been proposed which is used to represent and categorize users' concepts, such as projects, tasks, contacts, organizations, e-mails, etc. At the same time common background knowledge shared in the community can be represented by common ontologies.

The Haystack project (Quan et al., 2003) provides a range of desktop applications like word-processors, email clients, image manipulation, instant messaging, etc. It attempts to remove the information interoperability barriers in existing applications by

[33] http://www.twine.com/

replacing them with these new solutions. The integrated approach also allows individuals to manage their information according to their preferences. Haystack offers a complete semantic programming environment enabling creation of dynamic user interfaces. However, Haystack suffers from performance problems with heavy resource requirements. Moreover, the user is faced with a new environment, in place of established desktop applications, which is much complex and needs a long training time. Also the project did not establish any open standards for a semantic desktop.

The Social Semantic Desktop

Decker and Frank (2004) first stated the need for a "Networked Semantic Desktop". They presented a vision of how the Semantic Web, P2P computing, and online social networking will evolve into a networked semantic desktop. In a first phase, Semantic Web, P2P, and social networking technologies become mature and widespread. In a second phase, Semantic Web technologies are brought to the desktop leading to the development of Semantic Desktop and Semantic Web and P2P are integrated to form Semantic P2P. Social networking and Semantic Web lead to ontology driven social networking. In a third phase, the social, desktop and P2P technology integrate leading to the vision of a Social Semantic Desktop.

The NEPOMUK[34] project (Groza et al., 2007) aims to realize the vision of the Social Semantic Desktop. It aims to extend the personal desktop into a collaboration environment which supports both personal information management and also social and organizational information sharing. The project aims to create a standard for the Social Semantic Desktop, independent of the operating platform. A reference implementation has also been provided as an open source. Structured resources can either be manually added to the NEPOMUK desktop or extracted from desktop applications. A Data Wrapper would extract meta-data form structured data sources (e.g., email headers, calendar entries, etc.) and a Text Analysis service would extract data from unformatted text. Social information sharing is enabled though a peer-to-peer file sharing system. The NEPOMUK middleware also proposes a Mapping Service to map between many ontologies in overlapping domains (e.g., FOAF and vCard for contact data). A number of case studies have used NEPOMUK's solutions in various knowledge-work scenarios.

Lightweight Semantic Desktop applications. The ambitious Semantic Desktop frameworks mentioned above face challenges regarding the complexity, performance, usability and acceptance. This has been learnt from experiences as already mentioned. On the other hand, significant interest has also been drawn by rather lightweight but useful semantic desktop applications.

semiBlog (now renamed as Shift because it needs not be limited to blogging only) enables the reuse of data from desktop applications for the annotation of blog posts (Möller & Decker, 2005; Möller et al., 2005, 2006). The data can be pushed online to different blogging platforms through their APIs. It utilizes data in the users' desktop applications like the addressbook, calendar and bibliographic databases. Additional plugins may be developed for other desktop data sources. It provides a simple way for semantic annotation of blog posts, using techniques such as drag&drop and autocompletion.

[34] http://nepomuk.semanticdesktop.org/xwiki/bin/view/Main1/

The Semantic Clipboard. The desktop operating system provides a clipboard to copy and paste data between applications. However, the semantic structure of the data is lost in the transfer. The Semantic Clipboard (Reif et al., 2006) is a solution to preserve semantics of the data while being transferred across applications by using the Semantic Web as a clipboard. A prototype implementation has been demonstrated that can be used to copy and paste RDF meta-data between desktop applications. The Semantic Clipboard can also be used to copy and paste the meta-data from semantically annotated Web pages to a user's desktop application. The Microsoft's Live ClipBoard[35] also adopts the same idea and allows copy/paste of structured information between web pages and applications.

Möller et al. (2007) later proposed combining the above mentioned tools as they complement each other. While semiBlog allows the user to export data from various desktop applications to online blogs, Semantic Clipboard can import such data back into the applications from the web. Both semiBlog and the Semantic Clipboard use RDFa and Microformats to export and import annotations.

iv) Semantic annotation

Annotations are comments, notes, explanations, or remarks attached to any document or a selected part of the document. Annotation that references an ontology has been termed semantic annotation (Uren et al., 2006). Semantic annotation can enhance information retrieval and improve interoperability. Annotation metadata can be used not only for describing content, but also to organize and classify it (Kahan et al., 2001; Kiryakov et al., 2003; Popov et al., 2003; Koivunen, 2005). Automatic or semi-automatic annotation with pre-existing information can reduce the users' burden of creating annotations. Annotation may involve - i) both authoring and annotation together, or ii) only annotation with some existing data. Uren et al. (2006) present a detailed survey of annotation frameworks and semantic annotation tools and analyze them on the basis of a number of requirements. A large body of research on semi-automatic semantic annotation exists including significant works like Annotea (Koivunen, 2005; Kahan et al., 2001), S-CREAM (Handschuh et al., 2002), extraction ontologies (Ding et al., 2006), etc.

Semantic annotation in blogs

Semantic blogging may also be viewed as annotation to blog entries. OntoBlog (Shakya et al., 2007b, 2008a) demonstrates the application of semantic annotation to blogs. It proposes linking blog entries to ontology and instances, as illustrated in Figure 5. Blog entries are unstructured and scattered without explicit links among them. On the other hand, an ontology is well structured with semantic links. Blog entries can be linked to ontology using semantic annotation. Blog entries are self-contained snippets (Cayzer, 2004a) of information or small contents (Ohmukai & Takeda, 2004). So a blog entry may be considered as a single discrete unit of information. Thus, annotations can be applied to the blog entry as a whole.

OntoBlog acts as an integrated platform providing a single point of entry interface for publication and annotation. Such an integrated environment has been pointed out as a requirement for semantic annotation systems by Uren et al. (2006). Moreover,

[35] http://www.liveclipboard.org/

automation makes the process of annotation fast and easy for the blogger. The system automatically discovers related instances when blog entries are added and provides suggestions to the author.

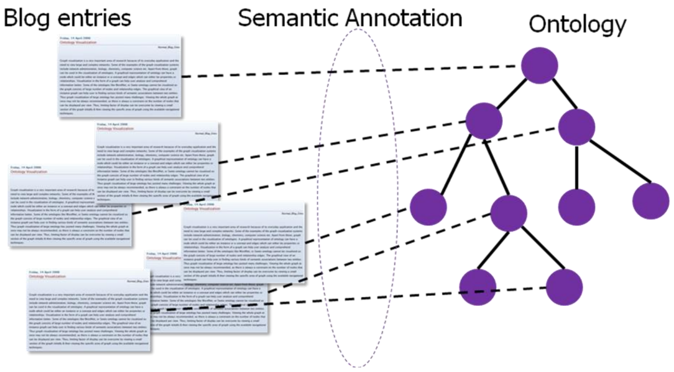


Figure 5. Linking blog posts and ontology by semantic annotation.

The MOAT (Meaning of a Tag)[36] framework (Passant & Laublet, 2008) facilitates people to define the meanings of tags explicitly by annotating them with URIs from existing semantic data resources. A MOAT server has been implemented to assist the process of meaningful tagging with URIs. LODr (Passant, 2008) is a service providing features to semantically annotate existing tagged content from various Web 2.0 services, based on MOAT and Linked Data principles.

Structured tagging techniques, like the Flickr machine tags[37], geo-tagging, triple-tags[38] or dc-tagging[39] try to inject structured information in existing social tagging platforms. This can also be considered as structured annotation with the tags.

Actually most of the works about structured instance data creation demonstrate semantic annotation if they are annotating some existing text or resources.

b. Collaborative creation of structured data and ontologies

All the works discussed in the above category can only produced limited types of instance data. The concepts do not evolve and new concepts or ontologies are not created. This category describes more powerful approaches which can have users produce both structured data and concepts or ontologies. Creating shared ontologies usually requires some form of consensus. Hence, these approaches are collaborative in nature. Some of the prominent works and approaches are as follows.

Semantic Wikis

Semantic wikis facilitate collaborative creation of resources by defining properties as wiki links with well-defined semantics. Semantic wikis enhance wikis to make the collaborative knowledge contributed by users more explicit and formal. Usually, the

[36] http://www.moat-project.org/
[37] http://www.flickr.com/groups/api/discuss/72157594497877875/
[38] http://geobloggers.com/archives/2006/01/11/advancedtagging-and-tripletags/
[39] http://efoundations.typepad.com/efoundations/2006/10/dctagged.html

relations between resource pages are encoded by semantically annotating navigational links using simple syntax. Although semantic wikis vary in their degree of formalization and semantic capabilities, frequently found features are typing/annotating of links, context-aware presentation, enhanced navigation, semantic search and reasoning support (Schaffert, 2006a).

Buffa et al. (2008) have reported on the current state-of-art of semantic wikis. They have broadly categorized the approaches used for semantic wikis into two categories - "the use of wikis for ontologies" and "the use of ontologies for wikis". Most of the current semantic wikis fall into the first category in which the wiki acts as the front-end of the collaborative ontology maintenance system. The Semantic MediaWiki (Krötzsch et al., 2006), which is one of the most popular semantic wikis, falls in this category. It has already been deployed in large scale applications. Some other semantic wikis in this category are Platypus, SHAWN, Rise, Rhizome, Semantic Media Wiki, WikSar, AceWiki, etc.

Semantic MediaWiki (SMW) is an extension to MediaWiki, that allows encoding of semantic data within wiki pages. This is done using an extended wiki-syntax within the wiki-text. SMW converts these into a formal description. Every article corresponds to exactly one ontological element (class or property). Every annotation in the article makes statements about this element. Relations are expressed as links from a page to another page. Attributes of a resource page are specified as data values in annotations for the page. The data types like integer numbers, strings, and dates have to be explicitly stated in the annotation. This is necessary for the proper processing of attributes. The Semantic Forms[40] extension for the SMW, developed recently, allows users to create forms for adding and editing pages that use templates to store semantic data. The forms are defined using editable text files, written in a custom markup language.

A factbox at the bottom of the page enables users to view all the annotation metadata. Users can also create dynamic pages by embedding queries into the wiki-text. An external SPARQL query service synchronized with the semantic content is also provided. Most of the annotations are simple ABox statements. The schematic information (TBox) in SMW is kept shallow. It is also possible to import data from OWL ontologies and to map wiki-annotations to existing vocabularies such as FOAF. But such powerful features are restricted to the administrator only.

The second category of semantic wikis consider the use of ontologies for enhancing wikis. Semantic wikis like IkeWiki, SWIM and SweetWiki fall into this category. IkeWiki (Schaffert, 2006a) supports WYSIWYG editing of page content and metadata aided by interactive features like auto-completion. It requires an existing ontology to be loaded. However, some support for ontology editing is also provided. IkeWiki provides support for different levels of formalization ranging from informal texts to formal ontologies. IkeWiki is being used for the EU-funded KiWi (knowledge in a wiki) project[41] which aims at collaborative knowledge management that combines the wiki philosophy with the intelligence and methods of the Semantic Web.

Although semantic wikis can be potentially used to create ontologies from the community, most semantic wikis usually focus on collaborative creation of instance

[40] http://www.mediawiki.org/wiki/Extension:Semantic_Forms
[41] http://www.kiwi-project.eu/

data resources. There are many other works about collaborative creation of semantic resources and ontologies. Some of them are as follows.

Freebase. Freebase (Bollacker et al., 2007), similar to Google Base[42], allows users to define their own schemas to model different types of data and maintain online collections of structured data (organized as *bases*). Freebase is a large collaborative knowledge base. However, it may be difficult for casual users to create their own types because of strict constraint requirements and the elaborate interface. All the attributes must have strict types and the range should be within the types already defined in the system. It may also be difficult to enter instance data in Freebase because of strict schema constraints. If an attribute takes as value a resource of some type, the resource must be entered first. Although Freebase has made a lot of instance data available by scraping data from vast sources like Wikipedia and MusicBrainz, a non-existing instance must be modeled and entered by the user. It is also difficult to link to external resources from within Freebase. Freebase interlinks instance data to each other as attribute values. However, it cannot link to external resources at the data level and it is difficult for other systems to link to Freebase data resources.

Exhibit (Huynh et al., 2007a) is a lightweight framework which enables casual users to publish web pages with different types of structured data based on their own schema. Exhibit attempts to empower the ordinary users to publish structured information on the Web for effective browsing, visualization and mash-ups. However, authoring such structured data pages manually would be cumbersome to the users.

The myOntology project (Siorpaes & Hepp, 2007a) also uses wikis for community-driven horizontal lightweight ontology building by enabling general users to contribute. The myOntology project proposes to use the infrastructure and culture of wikis to enable collaborative and community-driven ontology building. It intends to enable general users with little expertise in ontology engineering to contribute. It is mainly targeted at building horizontal lightweight ontologies by tapping the wisdom of the community. But when the direct goal is ontology construction, it may be difficult to motivate people to participate.

There are also other community-driven lightweight ontology construction applications like ImageNotion and SOBOLEO (Braun et al., 2007), where users collaboratively build a SKOS taxonomy from tags while using them to annotate resources. These applications are based on their model of ontology maturing in which tags are gradually structured into emerging ontologies in the form of hierarchies with the help of the community. In the first phase, emergence of ideas, the community freely contributes structured tags. In the second phase, consolidation in communities, people share tags and these evolve to represent a common vocabulary with common understanding. In the third phase, formalization, the tags are organized into hierarchies collaboratively. The final phase of axiomatization adds logical formalizations with the help of knowledge engineers to derive heavyweight ontologies.

2.4.2 Deriving semantic contents from existing data and systems

This broad category includes approaches in which the user continues to use and contribute to the existing social web applications but semantic contents are derived from the data without involving the end users directly.

[42] http://base.google.com/

a) Semantification of existing structured contents

Most of the online social applications are database driven and already contain contents structured to some extent. Some systems have more structure in the contents and others may have less structure. However, most of the systems do not follow any semantic standard or ontology. Hence, the data in the application, though already structured, cannot be shared with other systems. As a most common example, today people are using many different social networking services and other social websites. These are isolated islands of data or walled data gardens because the user cannot take his data from one site to another. However, the structured data in these silos can be easily released by exporting, translating or mapping them into open semantic standards. SIOC is such a semantic standard for online communities.

The **SIOC** (Semantically Interlinked Online Communities) (Breslin et al., 2005) initiative[43] aims to enable the integration of information shared in online communities. The SIOC ontology can be used to represent data from the Social Web in RDF format. It defines classes and properties that describe conversation media like discussion forums and posts in online community sites. The main concepts include *Site*, *Forum*, *Post*, *Event*, *Group* and *User* and posts are connected by relations like *has_reply*, *related_to*, *topic*, *has_sibling*, *has_creator*, etc. SIOC is commonly used in conjunction with the FOAF vocabulary for expressing personal profile and social networking information. Hence, it includes mappings to existing vocabularies such as FOAF and RSS. SIOC has been achieving significant adoption through its usage in a variety of commercial and open-source software applications.

i) Data Exporters

Implementing exporters for SIOC and FOAF can help in portability of data from social web applications (Bojārs et al, 2008a, b). The *DataPortability initiative*[44] was launched to address these portability issues of social data. Data portability is the ability for people to reuse their data across multiple applications. The project advocates that users should have the right to share their content items with other services and to move this content to other services if needed.

Different SIOC exporters[45] have been written for a number of popular weblogs like WordPress, Dotclear, b2evolution; forums like phpBB and content management systems like Drupal. semiBlog (Möller et al., 2006) also provides a plugin for the WordPress blog engine to export SIOC metadata. A number of other SIOC exporters have been developed, for e.g., the Mailing List Explorer that allows the exploration of mailing lists, Twitter2RDF exporter for Twitter microblogs, IRC2RDF converter for IRC, Sioku Jaiku2RDF converter for the Jaiku microblogging site, etc.

Similarly, Rowe and Ciravegna (2008) have developed a service to export semantic information from the popular Facebook social networking site in FOAF format. The RDF Book Mashup[46] (Bizer et al., 2007b) makes information about books, their authors, reviews, and online bookstores from Web 2.0 data sources available on the Semantic Web. Whenever it gets a lookup call for a book URI, it decodes the ISBN number of the book from the URI and uses the ISBN number to query the Amazon

[43] http://sioc-project.org/

[44] http://dataportability.org/

[45] http://sioc-project.org/applications

[46] http://www4.wiwiss.fu-berlin.de/bizer/bookmashup/

API and the Google Base API for information about the book. The resulting XML responses are turned into an RDF model.

Exposing structured content from relational databases

It would not be wise to try to replace existing well-established database solutions in organizations. Instead we can retain the scalability and stability of the existing systems while importing the wealth of underlying structured data into semantic formats. It would be easier to convince organizations to provide their data in semantic formats and take advantage of semantic technologies without replacing existing database technologies. There are some tools available to expose data in relational databases as linked data for the Semantic Web.

The D2R Server[47] (Bizer & Cyganiak, 2006) is a tool for serving Linked Data views on relational databases. The user only needs to provide the declarative mapping between the schemata of the database and the target RDF terms. D2R Server also provides a SPARQL endpoint for the database.

Triplify [48] (Auer et al., 2009) is a lightweight plug-in which facilitates the semantification of web applications. It exposes the semantic structures encoded in relational databases making the contents available as RDF, JSON or Linked Data. Triplify configurations have been provided for many popular web applications like osCommerce, WordPress, Drupal, Gallery, and phpBB.

The Virtuoso Sponger [49] is a middleware component of OpenLink's Virtuoso platform that generates RDF Linked Data from a variety of data sources. The sponger provides several cartridges, each including data extractors which extract data from one or more data sources, and ontology mappers which map the extracted data to one or more ontologies/schemas. The sponger delivers URI dereferencing functionality over legacy data sources.

RDFizers. A lot of structured data is also available in other formats like CSV, Microsoft Excel, or BibTEX. The RDFizer project[50] maintains a long list of tools for converting various data formats into RDF. ConverterToRdf[51] also provides a series of tools to convert several types of application-specific data into RDF.

ii) Web page scrapers

Often it is not possible to create or install extensions into online systems maintained by others. Many times systems do not have open APIs or extensible architecture. Some systems are not even database driven though the HTML may have visible structure of data. Scrapers may be employed in such cases to extract the structured contents directly from the web pages.

Piggy Bank (Huynh et al., 2007b) is a browser extension which enables people to collect information from existing web pages. It invokes screen-scrapers to collect

[47] http://www4.wiwiss.fu-berlin.de/bizer/d2r-server/
[48] http://triplify.org/
[49] http://virtuoso.openlinksw.com/wiki/main/Main/VirtSponger
[50] http://simile.mit.edu/wiki/RDFizers
[51] http://esw.w3.org/topic/ConverterToRdf

information in web pages and cast them in a well defined structure. The Solvent[52] Firefox extension can help the user in creating screen scrapers for Piggy Bank. Such useful pieces of structured data are gradually accumulated by the user for his own benefit, hence the name ‘Piggy Bank’. The structured data stored in personal piggy banks can be periodically posted to a Semantic Bank which is a central web server application for sharing the collected semantic data online.

Lixto[53] (Baumgartner et al., 2001) is a commercial solution for sophisticated wrapper generation for automated web information extraction. It enables users to semi-automatically create wrappers by providing a fully visual and interactive user interface. The generated wrappers can translate data pieces in HTML pages into structured XML data. The Dapper online service also provides a similar service with an easy-to-use interactive interface. The Intel Mash Maker[54] is a browser extension that enables users to create scrapers interactively to build mash-ups from websites.

DBpedia (Auer et al., 2007) exposes the structured data in Wikipedia on the Semantic Web like a database. It is also based on heuristics to scrap out structured contents from Wikipedia pages en masse. In a similar way, Freebase also maintains and renders Wikipedia data in a machine friendly structured form.

b) Defining the Semantics of Tags

The large volume of tagging data available in social sites may also be made meaningful to machines by modeling the semantics of tags and tagging activity. Various ontologies have been proposed for this.

Gruber (2007) has proposed an ontology of folksonomy to provide well-defined semantics to tags. He formalized core concept of tagging as a quadruple relation *Tagging*(*Object*, *tag*, *tagger*, *source*) where *object* is the object being tagged, *tagger* is the person who does the tagging and *source* is system or space where the tagging action takes place. He further describes the notions of constraints on tagging, negative tagging and tag identity. The Tag ontology[55] (Newman, 2005) also provides a model with *Tag* and *Tagging* classes in order to represent tags and tagging actions. The *Tag* class inherits from *skos:Concept* and it uses FOAF for modeling user aspects.

The MOAT ontology (Passant & Laublet, 2008) extends the tripartite model of folksnomies (Mika, 2007), by adding a *local meaning*, thus defining a tagging action as the quadruple *Tagging*(*User*, *Resource*, *Tag*, *Meaning*). The local meaning is specific to the user. The global meanings of a tag is defined as the list of all different meanings a tag can be assigned to - *Meanings*(*Tag*) = {(*Meaning*, {*User*})}. This is the set of pairs of the local meaning and the set of users who used it. As already described, a framework has also been implemented to semantically annotate tags with the meaning in terms of existing URIs.

The SCOT (an acronym for Social Semantic Cloud of Tags) ontology (Kim et al., 2008) proposes a way to share tags by modeling tag clouds semantically. It also provides various properties like synonymy, case-variation, etc. to link related tags

[52] http://simile.mit.edu/wiki/Solvent
[53] http://www.lixto.com/
[54] http://mashmaker.intel.com/web/
[55] http://www.holygoat.co.uk/projects/tags/

together. The int.ere.st[56] website demonstrates the use of SCOT for bookmarking and sharing social tagging data among different sources.

Common Tag[57] is also an open tagging format providing the capability to reference unique, well-defined concepts, using metadata.

c) Adding semantics to unstructured contents

Most of the contents in social applications are still in the form of unstructured text. Natural language processing (NLP) and information extraction (IE) techniques can help in extracting structure from such unstructured text to some extent. The OpenCalais[58] suite consists of various tools that takes plain-text and automatically embeds metadata into it. The Calais web service enables publishers, bloggers and sites to automatically metatag the people, places, facts and events in their content with the help of NLP technologies. Magpie (Dzbor et al., 2003, 2004) automatically creates a semantic layer over web documents and links instances identified in the document to relevant ontological instance/class. It uses simple lexicon-based parsing and linguistic rules to identify instances. KIM (Kiryakov et al., 2003, Popov et al., 2003) uses IE techniques for the recognition of named entities in documents. It maintains a pre-populated knowledge base of instances. It also introduces indexing and retrieval based on named entities. Ontology learning from text by NLP techniques is a vast area of research. Different elements of ontology like concepts, properties, relations, axioms and instance may be learnt. There are many elaborate works including Text-to-Onto (Maedche & Staab, 2001), DODDLE-II (Kurematsu et al., 2004), Powerset (Pell, 2007), SynDiKATe (Hahn & Romacker, 2001), etc. Shamsfard and Barforoush (2003) have written about the state-of-art in ontology learning from text.

d) Emergent semantics

Ontologies may also be derived from various unstructured or partially structured contents by computations and statistical methods. There are many works on deriving emergent knowledge structures from social data. Lightweight ontologies may be derived from folksonomies (Specia & Motta, 2007; Van Damme et al., 2007) applying basic ideas like grouping similar tags, forming emergent concepts from them, making the semantics more explicit and utilizing external knowledge resources to find semantic relations.

Mika (2007) proposed a unified model of social networks and tagged resources serving emergence of informal lightweight ontologies. He emphasizes that ontologies should come from the community and should reflect the social structures and interests of the community. He adopted the idea of *emergent semantics* as introduced by Aberer et al. (2004) as the global effect derived from individual interactions of a large number of rational agents. Ontologies should be an emergent effect rather than being developed by a small group of engineers. Mika has extended the traditional bipartite model of ontologies, including concepts and instances, with a social dimension forming a tripartite model of actors, concepts and instances. Actors are the users, concepts are represented by tags and instances are the objects being tagged. This

[56] http://int.ere.st
[57] http://commontag.org/
[58] http://www.opencalais.com/

tripartite model with hyperedges can be reduced into 3 bipartite graphs, namely (actors and concepts), (concepts and objects) and (actors and instances), with regular edges which can be handled more conveniently. The affiliation network of actor and concepts can then be folded into two graphs: a lightweight ontology of concepts based on overlapping sets of communities, and a social network of users based on overlapping sets of objects. Similarly, the other two bipartite graphs can also be folded to generate similar networks. For e.g., the concepts and objects graph leads to a semantic network, where the link between two concepts is weighted by the number of instances that are tagged with both. Thus, lightweight ontologies and social networks can emerge through simple graph transformations.

Tijerino et al. (2005) have proposed an approach to generate ontologies from tables semi-automatically. Tables are reverse engineered to create mini-ontologies. These are mapped and merged to form a growing global ontology. Tables are very common in websites and this approach can be applied to any online source with tabular data.

2.4.3 Limitations of the state-of-art

Although the state-of-art in creation of structured contents in the social Semantic Web has advanced a lot with many practical approaches and technologies, there are many limitations to be addressed. Some of major limitations are summarized below.

Limited types of data. Most of the existing works, either direct creation of structured contents by users (through semantic blogging, bookmarking, desktop, annotation, etc) or semantification of existing contents (through exporters, scrapers, etc), only produce limited types of instance data. There are a fixed set of concepts to structure these limited types of data. New concepts are not created and the ontologies do not evolve to accommodate new concepts and relations or update existing ones. However, there is wide variety of data different people are interested in and would like to share in the community. Most of the existing systems and ontologies only cover some popular concepts or types of data that a majority of people are interested in. There are many different concepts which smaller groups of people may be interested in. This wide variety of data with smaller audiences together forms a long tail of information domains (Huynh et al., 2007a) as shown in Figure 6. Together the long tail becomes comparable to the head of popular concepts and can no longer be neglected. Moreover, along with time new concepts may be required or existing ones may have to be changed.

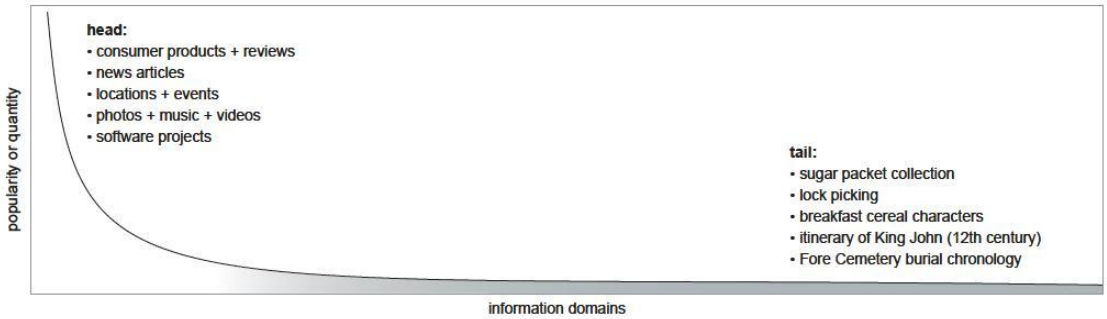


Figure 6. Long tail of information domains.

(Source: Huynh et al., 2007a)

Existence of multiple conceptualizations. In most of the collaborative systems, mainly wiki-based systems like semantic wikis, Freebase, myOntology, etc., each concept or resource has a single prominent model which everyone is assumed to settle with. However, in practice, multiple conceptualizations may exist because people have multiple perspectives and preferences. These multiple conceptualizations have to be taken into account while defining concepts and creating ontologies for people. It may not be possible to satisfy many people and different contexts simultaneously with a single conceptualization. Systems like Freebase do allow people to define their own concepts or types. However, the structured types defined by different users are kept separate in their own spaces and not consolidated or related in any way. Though some instance level reconciliation ("dataserver/reconciliation," 2008) is done in Freebase, schema level consolidation is not done. So the structured schemas defined by different people are not handled together. In spite of multiple conceptualizations, exchange and integration of information from different sources should be possible.

Difficulty of collaboration and consensus. Arriving to consensus among different people may not be easy for everything. Moreover, although some level of consensus may be achieved, collaborative interaction for consensus is itself a difficult and time-consuming process.

Complexity and learning curve. Collaborative platforms may be powerful considering the capability to create rich concepts and complex ontologies. However, more powerful they are more complex they tend to be. Existing platforms still have considerable learning curve for ordinary people and usability issues to be addressed by interface enhancements as for the Semantic MediaWiki (Pfisterer et al., 2008). The Freebase interface was also too complicated initially and was redesigned overall. We should not ignore the fact that these collaborative platforms should be designed for non-technical users. The constraints imposed by the systems also make it difficult for people to share different types of data they have. Systems for producing only structured instance data are relatively easy to use. However, these are limited to few popular types of data.

Difficulty in creating perfect definitions. It is a difficult task to create well-defined concepts, especially for ordinary people. It is difficult to cover all requirements of people and the requirements evolve. All possibilities and constraints cannot be conceived at a time. There may always be exceptions and unanticipated data. Imposing constraints may render the conceptual model too brittle to accommodate unanticipated needs. In existing collaborative platforms, like the semantic wikis and Freebase, users need to define the concepts with proper data type constraints. This may cause difficulty both at the time of concept definition and instance data entry.

Structuring tags instead of data. Tags can serve collaborative organization of data objects. However, it should be noted that tags are after all informal labels assigned to the objects. The meaning of the tags can be well-defined, the tags may be organized into ontologies by identifying semantic relations but the actual data objects are still left unstructured and cannot be processed by machines. The actual relation between the tag and the data object is also not defined.

2.4.4 Scope of interest and specific problems

The presented study shows that the area of structured content creation in the social Semantic Web is wide. However, the interest of the thesis is mainly on obtaining structured data from people and sharing wide range of data by forming ontologies. Limitations of the existing approaches have also been uncovered by the study. Therefore, the scope of interest can be further focused along a promising direction. Referring back to Figure 4, approaches for structured instance data creation alone can only produce limited types of data. Semantification of existing contents and systems also produces limited types of data. Similarly, deriving semantics from totally unstructured text has limitations and can never be perfect. Defining semantics of tags and emergent semantics leaves the real data objects unstructured. Instead of relying on the inherent structure of existing systems or trying to derive structure from totally unstructured contents, it would be better to have structured contents directly from the user. Therefore, to advance the state-of-art, interest will be focused on collaborative creation of structured data and ontologies. Nevertheless, all approaches have their own significance and multiple approaches can be used in appropriate combination. The thesis follows the same strategy.

Based on the above study, the characteristics of some representative works in collaborative knowledge base creation are roughly summarized in Table 1. The existing works are compared, limitations are observed and the desired solutions are mentioned. The closely related works include the semantic wikis, Freebase, myOntology and the ontology maturing approach (Braun et al., 2007). The works are characterized along several dimensions as follows.

1. Ease of use. Regarding the ease of use, semantic wikis still seem to be bit complex for ordinary users because they need to use some extended wiki syntax. Although the systems are powerful, some initial training is required before getting started. Freebase seems to be more usable with an interactive graphical user interface. However, the interface is elaborate with many features that may be overwhelming to the casual user. myOntology is directly aimed at building ontologies collaboratively. So understanding of ontologies is needed which is not likely to be the case with ordinary users. The ontology maturing approach is much easier as concepts are defined freely as tags. Building the tags into a taxonomy adds some difficulty. However, the concepts are not richly structured as in the former approaches. A solution easy to use for ordinary users would be desirable.

2. Expressiveness. Semantic wikis mainly produce semantically structured instance resources. There may be arbitrary relations among the semantically annotated pages. Many semantic wikis can be used to create ontologies. The Semantic MediaWiki can also be used to create concept schemas. The expressiveness of Freebase is limited to conceptual schemas and instance data. myOntology can be used to create lightweight ontologies by defining concepts, relations and instances. Most of these works are moderately expressive producing lightweight ontologies. Usually, formal axioms towards building heavyweight ontologies, which enable powerful inferencing, are not expressed. The ontology maturing approach offers lower expressiveness as it is only limited to building a concept hierarchy. A moderate expressiveness for representing at least conceptual schemas and instances would be enough in our case.

3. Constraints. It is seen that most of the works including the semantic wikis, Freebase and myOntology impose strict constraints to control input and quality of the data contributed by the users. The semantic wikis and Freebase mainly impose constraints on the data type. This is done to ensure that the data is in the form which can be automatically processed later. In myOntology, further constraints, besides type specifications, can also be expressed. However, thinking of appropriate constraints is difficult in itself and moreover the restrictions may make it difficult to input unforeseen data. Data input is much easier in the ontology maturing approach because it is like free tagging activity without imposing any constraints. It would be desirable to keep the constraints to a minimum to encourage free contribution.

4. Multiplicity. Almost all existing collaborative systems do not support multiple conceptualizations of the same thing. In Freebase, users can define their own types with the same name. However, these types are kept separate in their own spaces. These are not consolidated or related in any way. The thesis proposes allowing multiple conceptualizations of the same thing.

5. Consensus. In most of the collaborative approaches like the semantic wikis and myOntology, general consensus is required over each resource or concept. Consensus is maintained using the wiki philosophy over which the platforms are built. Conflicts are resolved by collaborative discussions and occasionally with the help of a moderator. The wikis provide the necessary mechanisms. In Freebase, the schemas in individual user spaces do not require consensus. However, when the schemas are in the common shared space, general consensus is assumed. The administrator can select and promote better contributions to the public space. The ontology maturing approach also requires consensus over building the concept hierarchy. It provides interactive online communication mechanisms to discuss issues face-to-face and resolve disputes. However, it may not be difficult to get concerned parties online at the same time. Having consensus is desirable but we do not want to enforce consensus for all cases.

Table 1. Analysis of existing collaborative knowledge base creation systems.

	Ease of use	**Expressiveness**	**Constraints**	**Multiplicity**	**Consensus**
Semantic Wikis	*Complex* - extended wiki syntax - some training needed	*Moderate* - Mainly instances, concept schemas possible	*strict type constraints*	*No*	*Needed* - Wiki way
Freebase	*Moderate* - Interactive but elaborate interface	*Moderate* - Concept schemas, instances	*strict type constraints*	Allowed but concepts not related	*Mostly needed* - Wiki way - selected by admin
myOntology	*Complex* - understanding of ontology needed	*Moderate* - Concepts, relations, instances	*Strict logical constraints*	*No*	*Needed* - Wiki way
Ontology maturing approach	*Fairly easy* - need to build taxonomy	*Low* - Concept hierarchy	*free tagging*	*No*	*Needed* - By interaction
Desired solution	*Easy*	*Moderate*	*Minimum*	*Yes*	*Optional*

Having presented the state-of-art, scope of interest and analysis of existing works in the area, the focus of the thesis is narrowed down to the following specific problems. As the details of these problems have already been discussed, these are just briefly listed below.

- *Complexity and learning curve*. Existing systems for creating concepts and sharing structured data are still complex for ordinary people. Especially, more powerful and expressive systems tend to be more complex. The strict constraints imposed also make the systems difficult to use.
- *Difficulty of concept definition and ontology creation*. It is difficult to create perfect concept definitions considering all possibilities. It is difficult to specify modeling constraints so as to accommodate most of the possibilities and cases. Enabling expressive and complex definitions rather makes the system difficult to use. Hence, ontology creation is still difficult.
- *Existence of multiple conceptualizations*. Different people may have multiple conceptualizations of the same thing due to different perspectives or contexts.
- *Difficulty of collaboration and consensus*. It is not always possible to achieve consensus and the process of collaborative interaction is itself difficult.

These problems are considered in more detail in the next chapter and some new solutions are proposed.

2.5 Summary

This chapter briefly introduced the Semantic Web, ontologies and their classification and some major Semantic Web technologies including linked data. The social web was also discussed. It was discussed how information can be shared in social web communities. The benefits and problems of both the Semantic Web and the social web were pointed out. It was highlighted how these two trends can complement each other to form a more practical social Semantic Web. The difficulties in bringing these structured and unstructured worlds together were also discussed. A review of existing works on sharing structured data on the social Semantic Web was presented. Some limitations of the current state-of-art were pointed out. Finally, focusing the interest to collaborative creation of structured data and ontologies, existing systems were analyzed from multiple aspects and some specific problems were highlighted.

3. Sharing Concepts and Structured Data

3.1 Concepts and Cognitive Theories

The difficulty in creating perfect concept definitions, as pointed out in the previous chapter, is well recognized by modern cognitive theories about concepts. These theories about psychology of concepts are briefly discussed in this section.

A concept is a representation of a category/class of things (Murphy, 2004; Lakoff, 1987). Concepts play an essential role in human understanding of the world. People generalize their knowledge in terms of concepts rather than individuals of the category represented by the concept. The knowledge of the concepts is used in identifying objects as being in a certain category, drawing inferences about new objects and communicating about objects.

The *classical view* of concepts claims that concepts can be defined in terms of a set of definitional properties. All the members of the category share these common features. These are the necessary and sufficient conditions for membership of the category represented by the concept. All the things, and only the things, having all of these properties are considered to be members of the categories. However, a large number of observations and experimental studies have proved that concepts are far more complex and cannot be represented by a particular set of defining properties. It is not always possible to find the defining features for concepts and may be extremely difficult in many natural cases. In the real-world, concepts are usually fuzzy along with gradations.

Typicality phenomenon. According to the classical view, all the members of a category should be equally good examples of the category as they possess all the defining attributes. However, in the real world, we usually see that some members are more typical to the category than others. For, e.g., a robin is a more typical bird than an ostrich. *Typical* members are good examples and *atypical* members are poor examples of the concept.

Also there are many things that are not clearly in or out of a category. Furthermore, some features may be more important than others and there may be complex mixtures of properties. Many real-life observations cannot be explained by the simplistic classical view. Hence, the classical view has been abandoned and several new theories about the psychology of concepts have been proposed (Murphy, 2004; Lakoff, 1987).

The Prototype View

Eleanor Rosch, in her revolutionary works in the 1970s, revealed many shortcomings of the classical view and proposed alternatives which constitute the prototype view (as cited in (Murphy, 2004; Lakoff, 1987)). A concept can be seen as a *prototype* - a *summary representation* that is the description of the category as a whole. It gives a unified representation characterizing all the members of the category. It is not necessary to have all the features to be in the category. Objects more similar to the prototype are better examples of the concept and these become the central members of the category. Objects not very similar to the prototype are bad examples and become outlying or borderline members. This gradation of similarity fits very well with the typicality effects.

Schema. A concept schema is a structured representation showing the properties (dimensions or slots) of an instance of the concept as attributes and values of these attributes (fillers of the slots) for the instance. A schema has been considered as an enhancement to the prototype view (Cohen & Murphy (1984); Smith & Osherson (1984) as cited in Murphy, 2004). Although, the feature list of the prototype view is a convenient short-hand representation, the schema gives a better understanding of the concept. Using the schema, relations among the dimensions can be established and constraints on the values for slots may be specified.

The Exemplar View

Introduced by Medin and Schaffer (1978) (as cited in (Murphy, 2004)) this is basically an extensional theory which considers concepts to be formed from the set of all remembered examples of the category. The membership of a new object to a category is determined based on its similarity with existing exemplars. The degree of similarity reflects the typicality effect.

The Knowledge Approach

The knowledge approach considers that concepts are part of our general knowledge about the world. Concepts are influenced by what we already know. They are interlinked with other concepts existing in our knowledge. Conversely, learning new concepts can also change our general knowledge.

Basic Level of Concepts

Human knowledge is usually generalized in the form of hierarchies of concepts. It has been observed that there exists a natural, preferred level in this hierarchy of concepts. This has been known as the *basic level* of categorization. For e.g., people usually call a Siamese cat “a cat”, which is the basic level, rather than “a Siamese cat” or “an animal”. This basic level is most spontaneously used and understood by the people than the sub-ordinates (the child concepts) or super-ordinates (the parent concepts). This middle level of specificity has the special advantage because concepts are the most informative and distinctive at this level. Rosch (Rosch 1978; Rosch et al. 1976) has presented a series of highly influential studies about the basic level (as cited in (Murphy, 2004; Lakoff, 1987)).

Further, it has been observed that the basic level might depend on the person’s level of expertise in the domain (Rosch, et al. 1976, Berlin 1992; Berlin, et al. 1973; Dougherty, 1978). The experts know more distinctive features in more specific levels of categorization than novices. Hence, experts may have a different basic level for concepts. Therefore, it is important to acquire and analyze concepts from the ordinary people rather than experts to obtain the general basic level of concepts.

The modern theories on concepts support our view about the natural vagueness and multiplicity of concepts (Takeda, 2008). The classical view still offers useful computational capabilities. Traditional theories of logic and reasoning are based on the implicit assumption of the validity of the classical view of concepts. However, it does not fully capture the actual richness of human conceptualizations. If we want to harness knowledge from the mass of people, to form a social Semantic Web, we cannot ignore the theories about human psychology of concepts. If we acquire concepts from people and ontologies emerge from mass contributions of ordinary people, the natural way of human conceptualizations, especially the basic level of concepts, will definitely be reflected. In fact, incorporating these principles in building knowledge structures will make the results more intuitive and usable for humans. The significance of the cognitive principles of concepts is recently being realized by Semantic Web and ontology researchers. For example, Peroni et al. (2008) have proposed an approach for identifying the key concepts in an ontology, which best summarize what the ontology is about, by combining several factors from cognitive science, network topology, and lexical statistics.

3.2 Integrating Heterogeneous Conceptualizations

3.2.1 Multiple conceptualizations and contexts

The above discussion indicates that conceptualization is not a definitive process. Concepts are vague approximate representations of the real world. Moreover, conceptualization depends upon the individual's perspective, knowledge and level of expertise in the domain. Therefore, different people will have multiple conceptualizations of the same thing. The vagueness and multiplicity of conceptualizations have also been discussed in (Takeda, 2008; Takeda et al., 1995). In fact, heterogeneity among different information sources is very common. Different types of heterogeneity, both syntactic and semantic, can be observed between different information sources (Stuckenschmidt & Van Harmelen, 2005).

The necessity for representing and relating multiple conceptualizations has been pointed out many. Takeda et al. (1995) modeled heterogeneous system of ontologies by introducing *aspects*. They have introduced a *combination aspect* to integrate various aspects and a *category aspect* as a collection of aspects about the same thing but with different conceptualizations. They proposed muti-agent communication by translating messages across different aspects.

Distributed Description Logics (DDLs) (Borgida & Serafini, 2003) is a formalism for loosely combining different DL knowledge bases preserving the identity and independence of each local ontology. C-OWL (Bouquet et al., 2004) is an extension to OWL using DDL for contextual ontologies. ε-connections (Kutz et al., 2004) is also a method for combining logical formalisms. Grau et al. (2004) have proposed extensions to OWL based on ε-connections.

The DOGMA approach (Meersman, 1999; Jarrar & Meersman, 2002; Jarrar & Meersman, 2008) for formal ontology engineering also recognizes the need for multiple perspectives and contexts. It distinguishes between domain and application-specific axiomatizations or conceptualizations as the "ontology double articulation principle". There may be multiple application-specific perspectives sharing the same domain conceptualization. The domain conceptualization is maintained as an ontology base consisting of context-specific binary fact-types called *lexons*. Lexons serve as incremental conceptualization units. In the DOGMA approach, contexts can accommodate different, even inconsistent, conceptualizations in the same ontology base. As the application-specific axiomatizations, applications establish ontological commitments using constraints and rules specific to their perspectives.

De Leenheer et al. (2009) have proposed an approach for business semantics management (BSM) based on the foundations of DOGMA and DOGMA-MESS (De Leenheer & Debruyne, 2008). BSM enables collaboration among business stakeholders and the reconciliation of heterogeneous business metadata in different organizations. In a phase of community-based semantic reconciliation, business semantics are modeled by extracting, refining, articulating and consolidating fact-types from existing sources. The consolidation is based on semantic equivalence and removal of redundancies resulting into consolidated semantic patterns stored in the community-shared semantic pattern base. The semantic patterns correspond to the lexons and the semantic pattern base is actually the ontology base.

The eCOIN (extended COntext INterchange) framework (Firat et al., 2007; Firat et al., 2005) also emphasizes the capability to deal with multiple contexts. The eCOIN approach assumes the existence of a shared ontology which represents the minimal level of agreements between the local models. The ontological terms, called *semantic types*, reflect generic meanings irrespective of any context. However, the details of each ontological term may vary according to the context. So each ontological term is specialized to individual local contexts through *modifiers*. The set of such modifiers constitute a *context model* for each of the multiple contexts. For example, 'air fare' may be a generic semantic type. But the specific details of the air fare may be different for different contexts. The currency may be different. Whether the fare is for round trip or one way, whether it includes the tax or not may be different. To reconcile such differences, mappings are defined as a *conversion function network*. Each of the contextual modifier dimensions are atomically related through conversion functions in this network. Thus, a generic ontology can be specialized to multiple contexts and multiple conceptualizations can be related. However, the formation of the generic ontology is itself challenging at first.

3.2.2 Data integration and schema matching

Multiple conceptualizations are inevitable, especially in a widely distributed large scale system like the web. However, people and organizations need to exchange information in spite of having heterogeneous information systems and multiple conceptualizations. Data from various systems, structured under different information models, need to be integrated and accessed uniformly for various purposes. A possible way of dealing with this problem is to consolidate multiple conceptualizations into a unified form. Corresponding elements of the multiple conceptualizations may be mapped and treated uniformly. The well-investigated fields of data integration and schema matching/ontology alignment can help in achieving this.

Data integration approaches

Data integration is the process of combining data from multiple sources so that they can be queried together as a single information source. Data integration has been a long tradition in research (Lenzerini, 2002). There are two main types of approaches for data integration – Global-as-View and Local-as-View.

In the Global-as-View (GaV) approach, a global schema is defined as a view on the local source schemas. The main advantage of the GaV approach is that queries on the global schema can simply be unfolded to the source schemas by substituting the corresponding terms. The union of individual results produces the total result. A downside of the GaV approach is that the global schema has to be maintained constantly as new sources are added or existing ones are updated.

In the Local-as-View (LaV) approach, the source schemas are instead defined as views on the global schema. Querying processing is difficult in LaV because reformulating the queries on the global schema in terms of the local sources is a difficult process. The advantage of LaV is that new sources can be added easily. The global schema need not be updated because the local sources are defined in terms of the global schema. However, a global schema needs to be in place first.

Data integration by ontologies

Ontologies are consensual representations of conceptualizations. So these can provide a common basis for data integration. Stuckenschmidt and Van Harmelen (2005) have discussed 3 different types of approaches for data integration based on ontologies.

Single global ontology. In this approach all different information sources use a single global ontology to model their data. Data integration becomes straightforward if such an approach can be followed. However, this cannot accommodate multiple conceptualizations for different sources, which is a common case.

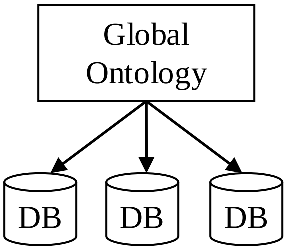


Figure 7. Single global ontology.

Multiple local ontologies. In this case, each information sources uses its own local ontology. Therefore, modeling multiple conceptualizations would not be a problem. However, information sharing and integration across different sources becomes very difficult. The Piazza system, mentioned in Section 2.3.1 follows this approach by maintaining semantic mappings locally between schemas of the peers. A major research challenge for the system is about distributed query processing. A query should be rewritten for sources reachable through the transitive closure of the mappings. The query answering scheme has to combine aspects of both Global-as-View and Local-as-View (Halevy et al., 2003).

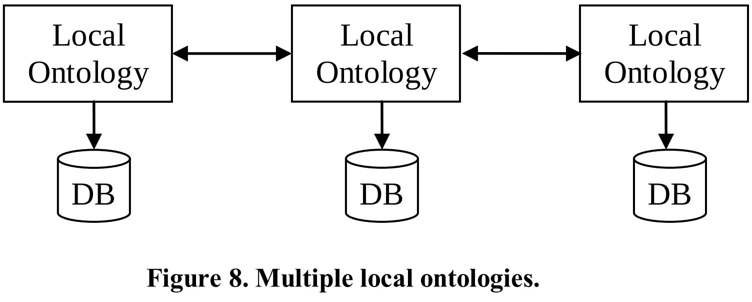


Figure 8. Multiple local ontologies.

Hybrid approach. The difficulties of the above approaches can be overcome by combining them into a hybrid approach. Different information sources have their local ontologies but they are built using one global shared vocabulary, as shown in Figure 9. Hence, multiple conceptualizations are allowed and information can be translated among the sources based on the shared vocabulary. Stuckenschmidt and Van Harmelen (2005) have proposed a detailed framework based on this hybrid approach.

Actually, this approach corresponds to the LaV approach just described as each local ontology is expressed in terms of the shared vocabulary. Therefore, the advantages and disadvantages of LaV apply. However, the biggest challenge for this approach is creating the shared vocabulary itself. Stuckenschmidt and Van Harmelen intend to use elaborate ontology engineering process and also propose a detailed methodology for ontology engineers and trained domain experts. This is a difficult process and fails to capture the requirements of the mass. Moreover, each local ontology has to be built from scratch using the shared vocabulary and existing ontologies cannot be reused.

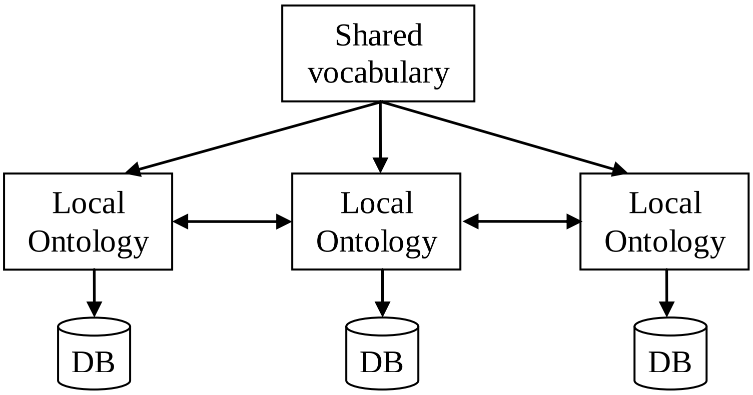


Figure 9. Hybrid approach with shared vocabulary.

Schema matching and Ontology alignment

There is already a large body of research about schema matching (Rahm & Bernstein, 2001) and ontology alignment (Euzenat et al., 2004). Various automatic and semi-automatic methods for aligning schemas or ontologies have been proposed. Clustering of schemas based on their similarity can help in identifying possible pairs of schemas that can be matched. There are sophisticated approaches for calculating the schema similarity (Castano et al., 1998; Rahm & Bernstein, 2001).

Ontology alignment is a difficult problem depending upon the complexity of the ontologies, complexity of the mappings and the level of accuracy needed. Traditionally, the focus has been on implementing various types of algorithms to automate this process. However, no matter how sophisticated and intelligent algorithms are employed, the process cannot be fully automated and accurate. Human intervention is essential at some point to verify, correct, refine and complete the process of alignment.

Recently, a complementary approach has been pointed out for solving this problem by using the power of people. Zhdanova and Shvaiko (2006) have introduced the notion of community-driven ontology matching. Potluck (Huynh et al., 2007c) is a user-friendly interface enabling casual users to align structured data schemas. Potluck is a data mash-up tool for casual users which can align, mix and clean structured data from Exhibit-powered pages. Fields can be merged by simple drag-and-drop, so that different data sources can be uniformly sorted, filtered and visualized. Merged fields are implemented as query unions. Currently, Potluck can only handle Exhibit-powered pages and not dynamic pages and other semantic formats. The schema alignment is mainly manual.

If the automatic methods are complemented by the community, the problem of schema matching or ontology alignment can be solved more effectively and accurately. However, the process has to be kept simple enough to enable ordinary people to understand and perform the alignment task.

3.3 Collaborative Knowledge Base Creation

A knowledge base includes the ontology and instance data. It should be noted that collaborative knowledge base creation mentioned here is in a limited sense of our scope of interest. Here, the term is mainly intended to refer to the collaborative creation of structured data and ontologies in the social Semantic Web as introduced in Section 2.4. In general, knowledge bases may range from simple databases to sophisticated formal ontological repositories. The notion of knowledge creation may reach out too far beyond our scope - see for e.g., (Nonaka & Takeuchi, 1995).

In this limited context, it is seen that most of the existing works on community-driven collaborative ontology creation follow the approach of constructing a single global ontology. A common knowledge base is maintained collaboratively by having consensus as shown in the Figure 10. As mentioned earlier, besides the difficulty for consensus, such direct collaboration by interaction may itself be a difficult process.

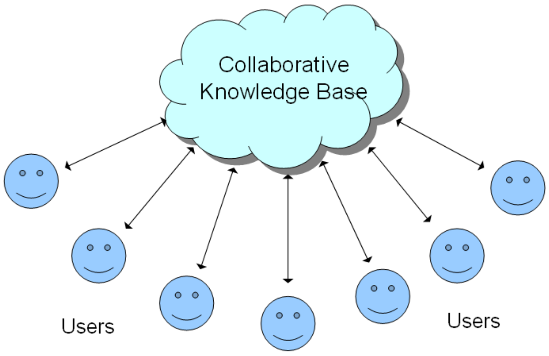


Figure 10. Existing collaborative knowledge creation approaches.

The thesis proposes a new approach for collaborative knowledge base creation. Each group of users sharing the same perspective can have their own local knowledge base. Therefore, multiple conceptualizations in the different local knowledge bases would be allowed. At the same time, these local knowledge bases are consolidated to form a global collaborative knowledge base which is shared by all. This is illustrated in Figure 11. The consolidation process itself can be community supported in addition to automatic computations of mappings among the local sources.

This process is quite similar to the hybrid approach described above. However, unlike the hybrid approach, this approach corresponds to the GaV approach. A global ontology is built in terms of the local ontologies by combining them. Hence, there is no need to construct a shared vocabulary beforehand. It emerges by the combination of local knowledge bases which are built by the community.

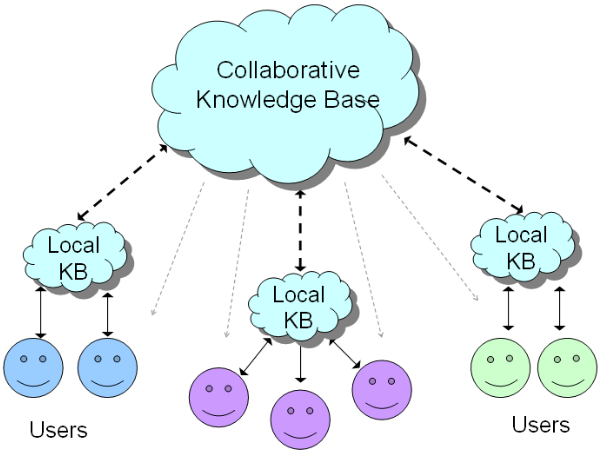


Figure 11. Proposed collaborative knowledge creation approach.

Global consensus is not required as people can maintain their own conceptualizations. These can be related while combining the conceptualizations, thus, enabling information sharing and integration. Also direct collaborative interaction is not essential making it easy for anyone to participate and contribute. Therefore, this serves as a loose collaborative approach for community-driven knowledge creation. The details of the approach are described in the sections to follow.

3.4 Overview of the Proposed Approach

The basic motivation of this thesis is to enable communities of ordinary people with multiple heterogeneous perspectives to share various types of structured data in the Semantic Web and, in the process, to derive ontologies for structuring different things. The proposed approach for achieving this consists of providing an online social platform intended to be easy for ordinary people to share data about various things. Users are given the freedom to define their own concepts thereby allowing multiple conceptualizations. The approach includes consolidation of these multiple conceptual schemas. This can be done by mapping the schemas semi-automatically with the help of schema alignment techniques and the community. Further, the concepts are organized by grouping them based on similarity. Concept grouping can also be done semi-automatically by applying algorithms for calculating the schema similarity. As a result of this consolidation and grouping of concepts, informal ontologies can gradually emerge that combine multiple perspectives unifying common elements.

The block diagram of the proposed approach is illustrated in Figure 12. The main parts of the overall approach are introduced below briefly.

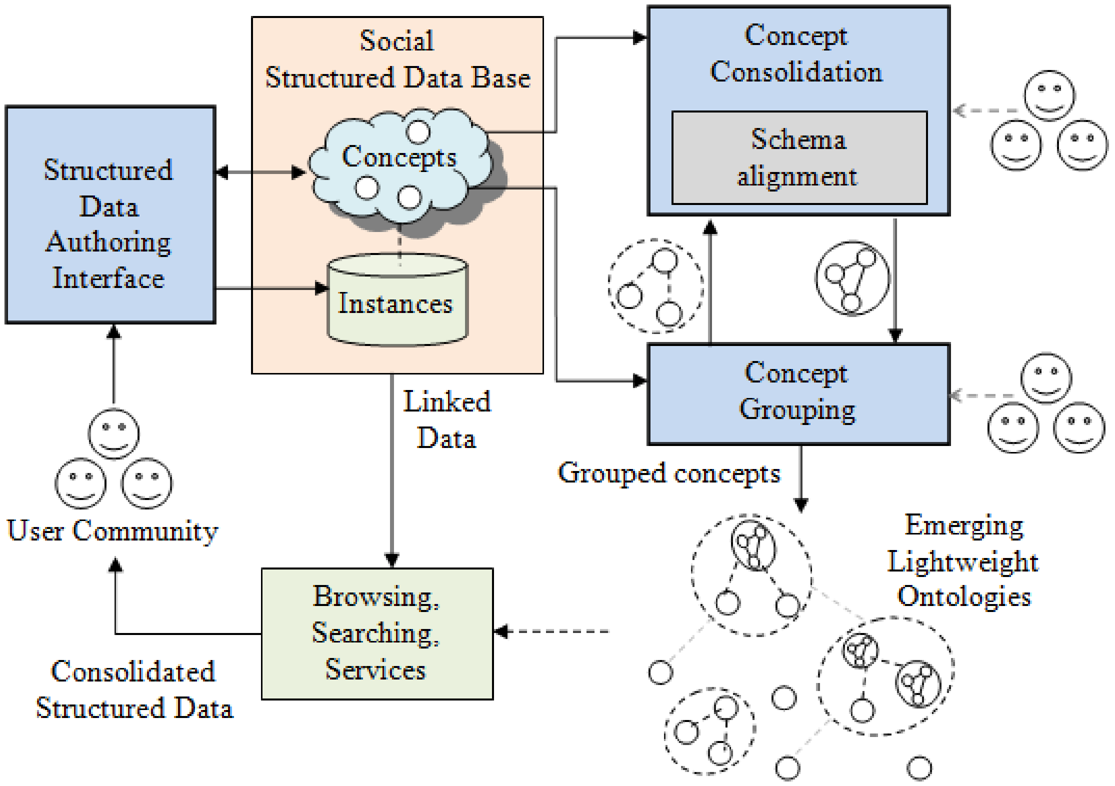


Figure 12. Block diagram of the proposed approach.

Structured data authoring interface. A social platform is proposed for structured data authoring to enable the ordinary user community to publish structured data directly on the Semantic Web. The users contribute concept schemas for the data they want to share or save. They may also directly use or adapt existing concepts defined by the community. The users post data instances though the authoring interface, with the help of the schemas. All the concepts and linked data instances are maintained in the social structured data base shared by the community.

Concept consolidation. Multiple versions of the same concept or similar concepts can be consolidated. This process includes alignment of the schemas, i.e., mapping of corresponding elements in the schemas. Concept consolidation is a semi-automatic process supported by the user community. This produces consolidated concepts combining the features of the constituent concepts.

Concept grouping. Similar concepts are grouped together. Groups with multiple versions of the same or similar concepts that may be consolidated are forwarded to the concept consolidation process. Moreover, grouping similar concepts helps in organizing the concepts, both individual and consolidated, so that relations among them become more apparent. Concept grouping is also a semi-automatic process supported by the community. The groups of concepts may further be linked and organized. Hence, emerging lightweight ontologies can form gradually in a bottom-up fashion. Popularity in the community also helps in the emergence of widely acceptable definitions from the cloud of concepts.

Services for using structured data. The consolidation of concepts forms a consolidated vocabulary to uniformly represent structured data from heterogeneous sources. The integrated collection of structured data can be viewed as consolidated structured data following this consolidated vocabulary. User interfaces can be provided for browsing and searching the different types of structured linked data. The grouping and organization of concepts by similarity makes browsing and locating information easy. The structured representation of data enables searching with detailed criteria. Other useful services may also be introduced to utilize the structured linked data.

3.4.1 Assumptions

There are some implicit assumptions for the approach to be applicable and to work properly. Some of them are as follows.

- First, the approach assumes that most concepts can be expressed with the help of flat non-hierarchical schemas. Hence, there would be no blank nodes in the graph representation of the concept schemas. In fact, use of blank nodes is discouraged for linked data (Bizer et al., 2007a). Statistical observations by Halpin (2009) have shown that blank nodes make up only a very tiny fraction of the total data set. Hence, we can safely assume the sufficiency of simple and flat schemas. Moreover, hierarchical schemas can also be flattened by combining the parent nodes.
- The approach assumes that the process of concept evolution and consolidation converges to not more than few versions. This is reasonable because usually different people do not have too many perspectives over the same thing. People can have consensus on most of the parts and settle down on few distinct perspectives.
- It is assumed that multiple conceptualizations over the same thing overlap and complement significantly and do not conflict much. The complementary parts can be combined and corresponding parts can be mapped.
- It is also assumed that most of the concept alignments are simple alignments that can be expressed conveniently by relations such as one-to-one or many-to-one equivalence.

The validity of these assumptions is also verified with some experimental evidences as presented later in Section 5.3 and Section 5.4. In the following text, the main aspects of the approach are described in more detail.

3.5 Structured Data Authoring by People

People want to share a wide variety of data. However, there are online systems only for sharing limited popular types of data. Also it is cumbersome to look for new systems for every new type of data, learn the system, understand the underlying data models and adapt ones data to fit to the existing data model and constraints. Therefore, a new data authoring platform is proposed where people can define their own structured concepts and share various types of structured data instances though a single platform. The following aspects are considered for the proposed platform.

Defining structured concepts. Users should be allowed to define their own structured concepts as they need. A concept can be structured by defining a schema composed of the attributes or properties of the concept. Defining one's own concept can be easier than understanding concepts defined by others and adapting own perspective and requirements to fit that. A schema is helpful in guiding people to input structured data. Although a schema-less environment offers full flexibility, users may not know what to input. So it would be better to maintain flexible and extensible schemas.

Publishing on the web of data. Today, social software has enabled ordinary people to publish documents on the web easily. In the web of documents, people mostly publish unstructured documents and interlink those using hyperlinks. The Semantic Web shifts the paradigm to *data publishing* and *data-linking*. This paradigm shift has to come in the publishing interface too that people use to share data. Thus, we propose enabling ordinary people to publish *data* on the web instead of unstructured documents and *data-links* instead of hyperlinks.

Flexible definitions and relaxed data entry. Creating perfect concept definitions with strict constraints is not easy and practical. It is difficult to think of all attributes and all possible value ranges at the time of concept definition. It may also be difficult to say whether an attribute value would be a literal or a resource and whether the attribute would have a single value or multiple values. While defining a concept A, if an attribute takes a resource of type B, we would need to ensure that concept B has already been defined. If concept B has an attribute which takes values of type C, then concept C must be defined first, and so on. Also we may not always have perfect data, or it may be difficult for the user to enter perfect data as mandated by a schema, at the time of data entry. All attribute values may not be known. Proper resource URIs for attribute values may not exist or the user may not be able to find it at the time. Moreover, exceptions may always exist no matter how well the schema has been designed and unpredicted new data instances may appear. These difficulties in data modeling and data entry can be avoided by allowing flexible and relaxed definitions. With such relaxed interface, of course, we may get some imperfect, incomplete or heterogeneous data. However, users generally enter appropriate or sensible data for their purpose. This has been evidenced by systems like tagging and wikis which accumulate large volumes of good data in spite of having completely relaxed interface.

3.5.1 User motivation for data contribution

As discussed in Section 2.3.2, one the major bottlenecks for collaborative creation of structured data and ontologies is how to motivate the users (Hepp, 2007). Following are some aspects that would motivate the users of the proposed system to contribute structured contents.

Data bookmarking. Using the system users may bookmark a wide variety of things they are interested in and care about. They need not be limited to bookmarking only web URLs, one at a time. While social bookmarking helps us to remember data sources, data bookmarking would help us remember the data as well. We would not need to go through all the documents again to find the important or useful facts, thus, saving us from a lot of effort in the future. The system can act as a personal knowledge management system to organize own data collection.

Social information sharing. As an online social platform the system inherits some motivating features of social software in general. Users may freely share interesting and useful things in the community modeled in their own formats. The system can also be an effective way of collecting data from the community in desired formats. By posting data instances and having others post to the system the user can have a useful collection of information in a structured way.

Information utilization. The users' data get organized under different concepts and can be retrieved by desired criteria. Useful operations like sorting, filtering, exporting, etc. and other automatic operations become possible. Moreover, data from different sources can be viewed and processed homogenously at one place.

Ease of use and freedom. The proposed structured data authoring interface is intended to be easy to get started without requiring any special knowledge or training. It allows publishing different types of data though a single platform. Users are free to create their own concepts to suit own needs. Most of the users may simply post their data using concepts created by others or modify and reuse existing concepts with little effort. A relaxed interface would allow the users to type in any data freely.

Targeting specific users. As described earlier, as the contribution inequality, most of the contributions come from only a small percentage of users. We may target users who have specific requirements and deploy the proposed system in targeted communities.

3.5.2 About the community

The community mentioned in the thesis does not have a fixed definition. The size and coverage of the community depends on the particular application where the approach is implemented. Some possible types of communities may be as follows.

- *Open online community*. This is the global community in which anyone can join and share any information. Such user community should be considered for online applications meant for all and not dedicated to any specific field of interest or group of people.
- *Community of common interest*. If the application is for a particular domain of interest, the coverage of the active community may be limited to the group of people sharing the common interest. The members may still be spread worldwide across organizational boundaries.
- *Closed community*. The community may also be closed within a particular group of people, organization or group of organizations. Applications specifically built for a group or organization serve such a community. The use of the application may range from informal information sharing to formal corporate use.

In some cases, these categories may overlap or co-exist within the same application.

3.6 Concept Consolidation

Multiple heterogeneous or overlapping conceptualizations always exist due to different requirements, perspectives or contexts (Ankolekar et al., 2007). Thus, multiple definitions for the same concept should be allowed. As illustrated in Figure 13, the same concept may be defined by different users in different ways. Even the same user may have multiple versions for the concept in different contexts. These can be grouped together and consolidated into a single virtual concept combining all the features of the individual definitions. Then, the user may retrieve all instances of a concept regardless of the concept version.

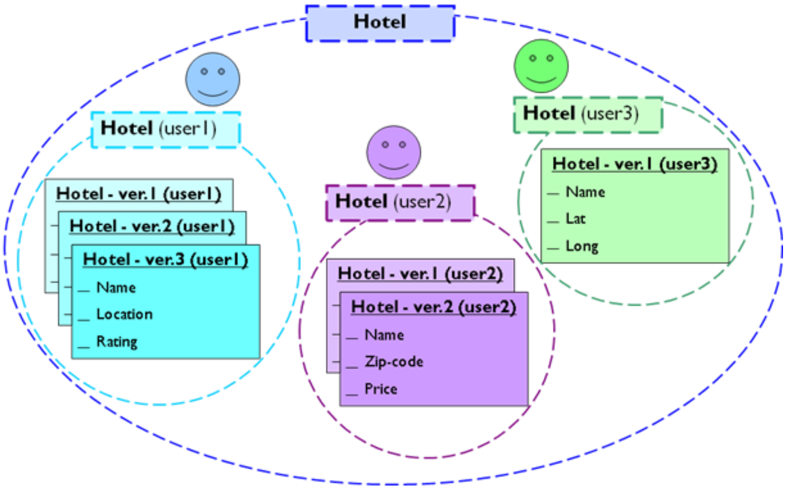


Figure 13. Concept consolidation.

3.6.1 Concept consolidation example

A detailed example is presented here. The example has been adapted from the tourism domain example on heterogeneity, described by (Stuckenschmidt & Van Harmelen, 2005), and some real observations from our experiment on conceptualization described later in Section 5.3. The example is an idealized case to illustrate the aspects covered by our approach.

One hotel owner may describe a hotel with a list of attributes as shown below (hotel 1). Suppose the rating of the hotel is expressed as the number of stars. The hotel only has single rooms. So all rooms have the same price represented by a single ‘price’ attribute.

Hotel 1

- name
- rating
- price
- amenities
- capacity
- contact
- access

Suppose another hotel owner describes the concept as follows (hotel 2). Instead of star rating he may prefer to use category (for e.g., luxury, standard, budget hotel, etc.). Moreover, if the hotel has single and double rooms, two separate attributes would be required to show the price. Further, suppose that the city has a good metro network. So the information about the nearest station would be useful information for access.

Hotel 2

- name
- category
- single room price
- double room price
- facilities
- no. of rooms
- phone-number
- address
- nearest station

An international tourist site may describe the same concept slightly differently (hotel 3). In this case, it would be more important to know the country and city first than the detailed address. Moreover, tourists would be interested in the near-by attractions around the hotel.

Hotel 3

- name
- rating
- price
- city
- country
- near-by attractions

Finally, suppose the government city office also maintains details about hotels in the city (hotel 4). It would need detailed postal information like the zip-code. Suppose the office also has a mapping application to map the locations of all hotels. The latitude and longitude coordinates may be used for such purpose. Information like the number of stories of the hotel building may also be useful if the office is concerned about the cityscape and planning.

Hotel 4

- name
- zip-code
- phone-number
- Latitude
- Longitude

- no. of stories

Therefore, the same concept may have different versions defined by different parties from multiple perspectives or contexts. Even the same person may be having multiple roles. For example, the same person may be a hotel owner and working at the city office. So he may model the same thing in different ways in the different contexts of his roles. There may be different types of heterogeneities, both syntactic and semantic, in the multiple concept definitions. For the above example of the hotel concepts, the heterogeneities in the attribute definitions are illustrated in the following Table 2. The table also shows how these multiple concept definitions can be consolidated by mapping corresponding attributes and combining complimentary attributes into a single consolidated concept.

Many of the corresponding attributes can be mapped one-to-one. The attribute labels may be the same, similar or synonymous or even quite different. But if they have the same intended meaning, they can be mapped one-to-one. Some attribute definitions may be different due to different contexts. For example *hotel 2* has single and double rooms. So it has two separate price attributes. However, for *hotel 1* and *hotel 3,* there are only single rooms. So the price means the single room price. Therefore, the price can be mapped to the single room price and the consolidated concept will have both the price attributes to generalize for both the cases. Similarly, in context of *hotel 2*, the access information corresponds to the nearest metro station while that is not the case in the context of *hotel 1*. From the perspective of *hotel 1* and *hotel 3*, the star ratings characterize the hotel. However, *hotel 2* characterizes the hotel by different levels of categories. Nevertheless, both of these attributes are intended to characterize the quality of the hotel and hence they can be mapped.

There may also be cases where multiple attributes combined map to one attribute. For example, *hotel 3* defines the address in terms of the city and country while *hotel 2* has a single address attribute. The multiple concept definitions from different perspectives also contribute many complimentary attributes that are only significant from the particular perspective and do not have a counterpart in other concept definitions. For example, the zip-code, geographical coordinates and number of stories of the hotel is only defined by the city office. Similarly, the near-by attractions attribute is contributed by the tourist site perspective.

When the multiple concept definitions are consolidated, a rich consolidated concept is formed which unifies all the definitions. The table clearly shows that the consolidated concept has much larger number of attributes than any of the individual definitions. Hence, consolidation combines the knowledge of different parties to form richer and generalized conceptualizations. At the same time, the process also establishes the relations among the multiple definitions, thus, enabling interoperation.

Table 2. Concept consolidation example.

Attribute mappings		**Hotel 1**	**Hotel 2**	**Hotel 3**	**Hotel 4**	**Consolidated**
One-to-one	Same label	name	name	name	name	name
	Similar, synonymous or different labels	amenities	facilities			facilities
		capacity	no. of rooms			capacity
		contact	phone-number		phone-number	contact
	different context or perspective	price	-single room price -double room price	price		-single room price -double room price
		access	nearest station			access
		rating	category	rating		rating
Many-to-one			address	-city -country		address
Complimentary					zip code	zip code
					-latitude -longitude	-latitude -longitude
				near-by attractions		near-by attractions
					no. of stories	no. of stories
No. of attributes		7	9	6	6	14

3.6.2 Formalization

In this section, our approach of consolidating multiple concept definitions is formalized. The proposed approach for consolidation is based on the Global-as-View (GaV) approach for a data integration system where a global schema is defined in terms of the source schemas (Lenzerini, 2002). The approach is simplified in our case because a concept schema does not have multiple relations and integrity constraints as in relational database schemas.

Definition 1. Concept and Instances. A concept C is an entity characterized by a set of attributes given by the function $att(C) = \{a_1, a_2, \ldots.. a_r\}$

The fact that x is an instance of C is denoted by the relation *instanceof*(x, C). C may have a set of instances I. The value for an attribute a of an instance k of C is given by the function $v(k, a)$.

Definition 2. Concept Consolidation. A concept consolidation $\mathcal{C}$ is defined as a triple $<\overline{C}, S, \mathcal{A}>$ where

- $\overline{C}$ is called the *consolidated concept*
- S is the set of *constituent concepts* $\{C_1, C_2, \ldots..C_n\}$, n is the number of constituent concepts
- $\mathcal{A}$ is the *alignment* between $\overline{C}$ and S.

Let the set of attributes of $C_i \in S$ be $att(C_i) = \{a_i^1, a_i^2, \ldots\ a_i^{n_i}\}$where n_i is the number of attributes of C_i. Let the set of attributes of $\overline{C}$ be $att(\overline{C}) = \{\overline{a}_1, \overline{a}_2, \ldots.. \overline{a}_m\}$, called *consolidated attributes*, where m is the number of attributes of $\overline{C}$.

Definition 3. Alignment between Attributes. For each concept $C_i \in S$, if attribute $b_i^k \in att(C_i)$ is aligned to $\overline{d}_l \in att(\overline{C})$, it is denoted as

$$aligned(\overline{d}_l, b_i^k)$$

for $l = 1, 2, \ldots\ r$ ($r \leq m$). All $\overline{d}_l$ are different. The mapping between $\overline{C}$ and C_i is defined as a set of ordered pairs

$$M_i = \{(\overline{d}_l, b_i^k) \mid \forall\ \overline{d}_l \in att(\overline{C})\ aligned(\overline{d}_l, b_i^k) \wedge b_i^k \in att(C_i)\}$$

aligned represents a correspondence between the aligned attributes. Some relation may hold between the aligned attributes asserted by the correspondence.

Then, alignment $\mathcal{A}(\overline{C})$ between $\overline{C}$ and concepts in S is defined as the set of mappings $\{M_1(\overline{C}), M_2(\overline{C}), \ldots..\ M_n(\overline{C})\}$.

The Figure 14 below illustrates the formal definition of concept consolidation according to the definitions given above. The notion of image and view are described in the following text (definitions 6 and 7).

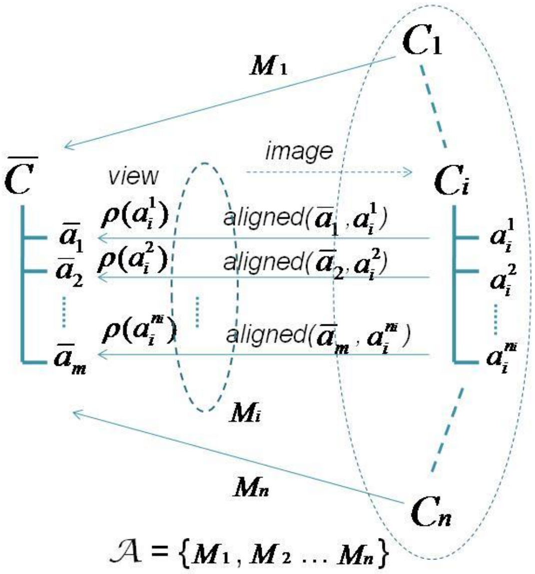


Figure 14. Formalization of concept consolidation.

Definition 4. Mapped Concepts in a Concept Consolidation. A concept $C_i \in S$ in the concept consolidation $<\overline{C}, S, \mathcal{A}>$ is said to be *mapped* if and only if

$$\exists x \in att(C_i)\ \exists y \in att(\overline{C})\ aligned(y, x)$$

i.e., at least one of its attributes is aligned to a consolidated attribute.

Definition 5. Grounded Consolidated Concept. The consolidated concept $\overline{C}$ in $<\overline{C}, S, \mathcal{A}>$ is said to be grounded if and only if

$$\forall\ z \in att(\overline{C})\ \ \exists x\ \in \bigcup_{i=1}^{n} att(C_i)\ \ aligned(z, x)$$

i.e., all the consolidated attributes are aligned to some attribute of the constituent concepts.

Definition 6. View of an attribute in a consolidated concept. The view of an attribute $b \in att(C_i)$ of concept C_i in the consolidated concept $\overline{C}$ for $\mathcal{C}$ is given by the following function.

$$\rho\ (b, C_i, \mathcal{C}) = \left|\begin{array}{ll} a & \text{if } \exists\, a \in att(\overline{C})\ \ (a, b) \in M_i(\overline{C}) \in \mathcal{A}(\overline{C}) \\ \phi & \text{otherwise} \end{array}\right.$$

Definition 7. Image of a consolidated attribute. The image of an attribute $a \in att(\overline{C})$ of the consolidated concept $\overline{C}$ for a constituent concept C_i in $\mathcal{C}$ is given by the following function which is the inverse function of ρ.

$$\sigma\ (a, C_i, \mathcal{C}) = \left|\begin{array}{ll} b & \text{if } \exists\, b \in att(C_i)\ \ (a, b) \in M_i(\overline{C}) \in \mathcal{A}(\overline{C}) \\ \phi & \text{otherwise} \end{array}\right.$$

Definition 8. Consolidated views of instances. The view of an instance *k* of concept C_i in the concept consolidation $\mathcal{T}$ is given by the following function

$$\bar{k} = w(k, C_i, \mathcal{T})$$

where $instanceof(\bar{k}, \bar{C})$ and the value of each attribute $\bar{a}_j \in att(\bar{C})$ (j = 1, 2, … m) for $\bar{k}$ is given by

$$v(\bar{k}, \bar{a}_j) = \left| \begin{array}{ll} v(k, \sigma(\bar{a}_j, C_i, \mathcal{T})) & \text{if } \sigma(\bar{a}_j, C_i, \mathcal{T}) \neq \phi \\ \phi & \text{otherwise} \end{array} \right.$$

The value $v(k, a)$ of each attribute a of C_i is known. The set of instances of $\bar{C}$ is exactly

$$\bar{I} = \{ \bar{k} : \bar{k} = w(k, C_i, \mathcal{T}) \wedge instanceof(k, C_i) \wedge C_i \in S \}$$

$\bar{I}$ is disjoint from the set of instances I_s of the constituent concepts in S.

Translation of instances

Using the alignment in the concept consolidation, translation of structured data instances from one conceptualization into another is also possible. This can be done by first converting the data instance of one concept into the consolidated concept form and then converting this consolidated concept instance into the target concept form, using the alignment mapping. This is formalized in the following simple theorems.

Theorem 1. Translation of instances. The translation of an instance *k* of concept C_i to another concept C_j in the concept consolidation $\mathcal{T}$, denoted by the function

$$k' = \gamma(k, C_i, C_j, \mathcal{T})$$

can be obtained as follows. If $\bar{k} = w(k, C_i, \mathcal{T})$ is the consolidated view of instance *k*, the value of each attribute $a_j^l \in att(C_j)$ (l = 1, 2, … n_j) for k' is given by

$$v(k', a_j^l) = v(\bar{k}, \rho(a_j^l, C_j, \mathcal{T})) \qquad \text{(from def. 6)}$$

$$= v(k, \sigma(\rho(a_j^l, C_j, \mathcal{T}), C_i, \mathcal{T})) \qquad \text{(from def. 8)}$$

Attributes of k' are exactly $att(C_j)$. However, $k' \notin I_s$.

Theorem 2. Lossless Translation. Instances of concept C_i can be translated to instances of C_j without any loss of information iff the following conditions hold.

$$\forall a \in att(C_i)$$

$$\rho(a, C_i, \mathcal{T}) \neq \phi \text{ and}$$

$$\sigma(\rho(a, C_i, \mathcal{T}), C_j, \mathcal{T}) \neq \phi$$

$| att(C_i)| \leq | att(C_j)|$ is a necessary condition for the lossless translation of an instance from C_i to C_j. If $k_j = \gamma(k_i, C_i, C_j, \mathcal{T})$ is lossless, $k_i = \gamma(k_j, C_j, C_i, \mathcal{T})$.

Query over a Concept

The main advantage of a GaV is that queries on the global schema can simply be unfolded to the source schemas by substituting the terms. In our case, queries on the consolidated concept $\overline{C}$ can be unfolded to queries on the constituent concepts using the attribute alignments in $\mathcal{A}$. The union of individual results produces the total result. Similarly, we can also translate queries over one concept schema into queries over another. The following theorems for unfolding and translating queries formalize this. The proofs follow from the literature for the GaV approach (Lenzerini, 2002).

Theorem 3. Unfolding Queries over $\overline{C}$ in $\mathcal{C}$. Any query $Q(\overline{C})$ over $\overline{C}$ can be unfolded into the union of queries $Q_1(C_1) \cup Q_2(C_2) \cup \ldots \cup Q_n(C_n)$, where $C_i \in S$ (i = 1, 2, … n). Let the queries be defined over the concept attributes as follows

$$Q(\overline{C}) = Q'(\overline{a}_1,\ \overline{a}_2\ , ...\ \overline{a}_r) \text{ where } \overline{a}_j \in att(\overline{C})\ (j = 1, 2, \ldots r)$$

$$Q_i(C_i) = Q_i'(a_i^1, a_i^2, \ldots\ a_i^r) \text{ where } a_i^j \in att(C_i)$$

Each Q_i can be obtained by unfolding the attributes in Q using $\mathcal{C}$

$$Q_i'(a_i^1, a_i^2, \ldots\ a_i^r) = Q'(\sigma_i(\overline{a}_1), \sigma_i(\overline{a}_2), \ldots\ \sigma_i(\overline{a}_r))$$

where $\sigma_i(a)$ is the short form of $\sigma\ (a, C_i, \mathcal{C})$.

Theorem 4. Query Translation. The query $Q_i'(a_i^1, a_i^2, \ldots\ a_i^r)$, $a_i^k \in att(C_i)$ (k = 1, 2, … r) over C_i can be translated into a query $Q_j'(a_j^1, a_j^2, \ldots\ a_j^r)$, $a_j^k \in att(C_j)$ over C_j in the concept consolidation $\mathcal{C}$ as following

$$Q_j'(a_j^1, a_j^2, \ldots\ a_j^r) = Q_i'(\sigma_j(\rho_i(a_i^1)), \sigma_j(\rho_i(a_i^2)), \ldots\ \sigma_j(\rho_i(a_i^r)))$$

where $\rho_i(a)$ and $\sigma_j(b)$ are short forms of $\rho\ (a, C_i, \mathcal{C})$ and $\sigma\ (b, C_j, \mathcal{C})$ respectively.

Maintaining Multiple Conceptualizations

Although multiple concepts are consolidated into a single unified view, the individual concepts are also retained along with their own definitions and descriptions. This maintains the multiple perspectives different individuals hold. Commonalities and differences between the intensions of the concepts can be identified by people with the help of individual descriptions of the concepts. The consolidation process only abstracts the compatible and complementary attributes from the individual concepts into a virtual unified concept. Thus, multiple conceptualizations are maintained and, at the same time, related and unified through the mechanism of concept consolidation.

3.7 Concept Organization by Grouping

Similar concepts are grouped semi-automatically. This can serve two purposes (illustrated in Figure 12). Firstly, it becomes easy to find out same or similar concepts that can be consolidated. Secondly, grouping similar concepts helps in organizing the concepts so that browsing and locating information becomes easy and relations among concepts become more apparent. Concepts are grouped under some similarity threshold. A higher threshold will result in tight concept groups with higher similarity. However, the coverage of concepts will decrease. On the other hand, a lower threshold will have better coverage at the cost of allowing lower similarities. An appropriate threshold value may be reached by testing iteratively until satisfactory accuracy and coverage is attained.

3.7.1 Concept schema similarity

The similarity (*ConceptSim*) between two concepts, C_1 and C_2, is calculated as the weighted sum of the concept name similarity (*NameSim*) and the schema similarity (*SchemaSim*).

$$\mathit{ConceptSim}(C_1,C_2) = w_1*\mathit{NameSim}(N_1, N_2) + w_2*\mathit{SchemaSim}(S_1, S_2) \qquad (1)$$

where N_1 and N_2 are the names of the concepts, S_1 and S_2 are the associated schemas respectively and w_1 and w_2 are the percentage weights ($w_1 + w_2 = 1.0$).

Appropriate values for the weights are also determined by iterative testing with a fixed threshold. As described later in Section 5.4.4, it has been observed that w_1, the name similarity, is much more significant than w_2, the schema similarity, and has to be assigned a higher weight accordingly.

Schema Similarity

The schema similarity *SchemaSim*(S_1, S_2) is calculated in the following steps.

1) For all possible pairs of attributes, calculate the name similarities between attribute labels (as explained next).
2) Create an n_1* n_2 matrix of these name similarities, where n_1 and n_2 are the number of attributes of S_1 and S_2 respectively.
3) Determine the best matching pairs of attributes between S_1 and S_2 employing the Hungarian algorithm using the similarity matrix from step 2.
4) Calculate *SchemaSim*(S_1, S_2) as the matching average of the attribute similarities for the best matching pairs found in step 3.

$$\text{matching average} = 2*\sum \text{name similarity of matching attribute pairs}/(|A_1|+| A_2|) \qquad (2)$$

where $A_1$and A_2 are the attribute sets of S_1 and S_2 respectively.

Hungarian algorithm. The Hungarian algorithm (Kuhn, 1955) is a combinatorial optimization method for solving the assignment problem. The assignment problem can be stated as follows.

> Given a weighted complete bipartite graph G = (X∪Y, X×Y) where edge *xy* has weight w(*xy*), find a matching M from X to Y with maximum weight.

Simpson and Dao (2005) used the Hungarian algorithm to find the semantic similarity between two sentences. Their technique has been adapted to find the similarity between two schemas. The time complexity of the Hungarian algorithm is $O(n^3)$. There are more sophisticated approaches for calculating the schema similarity (Castano et al., 1998; Rahm & Bernstein, 2001) depending on the complexity of the schema and accuracy needed. In our case, there are no complex hierarchical schemas and strict data types and perfection is not expected in informal user-defined schemas. So a simple and fast method with acceptable results has been used.

Name Similarity

The name similarity *NameSim* between the concept labels, or attribute labels, is calculated using the Lin's algorithm for WordNet-based similarity (Lin, 1998) (WordNet 2.1 has been used). The Lin's algorithm computes the semantic relatedness of word senses using the *information content* of the concepts in WordNet and the *similarity theorem* described in (Lin, 1998). The subsumption hierarchy of the concepts in WordNet is taken into account by this method. The information content is a corpus-based likelihood measure. The more generic a concept is, the lower its information content.

However, if a word is not found in WordNet, the Levenshtein distance is used to calculate the edit distance similarity. The Levenshtein distance[59] measures the difference between two strings by the minimum number of operations needed to transform one into the other.

For all possible pairs of concepts C_1,C_2 *ConceptSim* is calculated using equation 1. Pairs of concepts with *ConceptSim* above the threshold are considered to be related. Finally, all related concepts are collected into groups.

3.7.2 Emergence of lightweight ontologies

Enabling people to contribute concepts freely would result in a huge cloud of concepts. However, there are several ways by which prominent, stable and converging knowledge structures can emerge from the user contributions. The following ways enable the emergence of lightweight ontologies in the proposed approach.

1. Firstly, provision for collaborative maintenance and reuse helps in evolution and refinement of existing concepts. This helps to keep up with the conceptual dynamics (Hepp, 2007) in the community.
2. Secondly, consolidation of the user-defined schemas, which may be partial definitions from different perspectives, results into more complete definitions satisfying wider requirements. These consolidated concepts act as common consolidated vocabularies for the community to share structured data.
3. Thirdly, popular concepts can emerge out in the same way as popular tags emerge in folksonomies. The large number of concepts contributed by the community, including multiple versions, forms a cloud of concepts similar to a tag cloud. Popular concepts can emerge out from this concept cloud. The popularity may be decided by various indicators like number of instances,

[59] http://en.wikipedia.org/wiki/Levenshtein_distance

usage, ratings, etc. Popularity of a concept reflects consensus in the community about the concept. As popular concepts emerge out they gain even more attention and become more popular and more widely used. Hence, convergence and stability of the emerging ontology can be achieved.

4. Finally, concepts can be organized systematically by grouping or clustering similar concepts. This can be done semi-automatically with computations and community effort while organizing contents for themselves.

These mechanisms facilitate the emergence of ontological structures embodying the knowledge and consensus of the community. The resulting ontologies fall on the lightweight side of the spectrum of expressiveness defined by Corcho et al. (as cited in Schaffert et al., 2005, p. 7) as emerging informal vocabularies of concepts and relations for structured information sharing (see Figure 1). The concepts act as term list or vocabulary to categorize things in the domain. The schemas provide the class/property frame definitions. The concepts grouped by semantic proximity may serve as a thesaurus.

Ontology emergence in the proposed approach is similar to that in the business semantics management approach (De Leenheer et al., 2009) based on the DOGMA approach. In their approach, a common shared ontology base is formed by the consolidation of multiple perspectives of the stakeholders in the community. The proposed approach also fits quite well into the model of ontology maturing described by Braun et al. (2007). In the first phase, emergence of ideas, the community freely contributes structured concept schemas. In the second phase, consolidation in communities, people use each others' concepts as common vocabulary to share structured data, concepts evolve with refinements along with appropriate versions and these are explicitly consolidated by aligning corresponding features. The third phase, formalization, is partly covered by grouping and organizing similar or related concepts to form lightweight ontologies. However, the organization is quite informal and not headed towards the final phase of axiomatization for making heavyweight ontologies.

3.8 Application Scenarios

The proposed approach has been realized by implementing a social web application called StYLiD (an acronym for Structure Your own Linked Data). It is available online[60]. Various application scenarios may be conceived with the proposed platform for collaborative modeling and sharing of different types of structured data. Two important scenarios are discussed below.

3.8.1 Information sharing social platform

StYLiD can be used as or may be adapted to create a social website for structured information sharing as illustrated in Figure 15. It provides a CMS (Content Management System) where users can freely contribute their own concept schemas and share structured instance data. Data integration is done by concept consolidation using semi-automatic schema alignment techniques supported by the community. Concepts are also grouped and organized by the community. The structured data can further be annotated with external resources like Wikipedia.

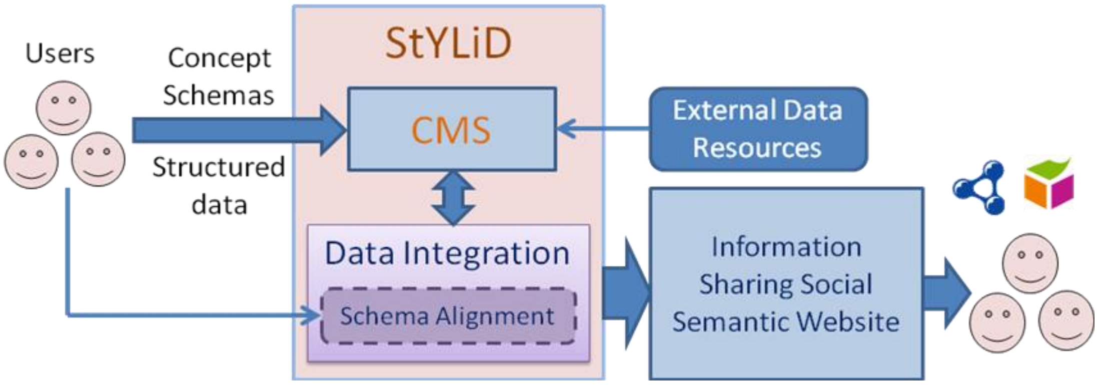


Figure 15. Information sharing social platform scenario.

The system can be used to bookmark and share things of personal interest and invite data from the community in desired schematic formats. With structured linked data and Semantic Web formats, users can enjoy various semantic capabilities while sharing data they are interested in.

3.8.2 Integrated semantic portal

Another application scenario for StYLiD may be as an integrated semantic portal as illustrated in Figure 16. In this scenario, the concept schemas and structured data may come from different information sources, websites or online systems, besides the users. Wrappers may be needed to export data from the systems into StYLiD acting as a data backend. The different information sources can maintain their own conceptual schemas and continue to serve their consumers. At the same time, these are also integrated in StYLiD by concept consolidation with schema alignment which can be handled by the system administrators. The data may further be enriched by linking to

[60] http://www.stylid.org/

external data resources. The system can act as a semantic portal providing integrated linked data and semantic capabilities to the user community or drive such vertical portals. In this scenario, StYLiD can comfortably be used with legacy systems. It would be easier to convince data providers when they can still maintain their own local systems while enjoying exposure to the linked data web through vertical portals.

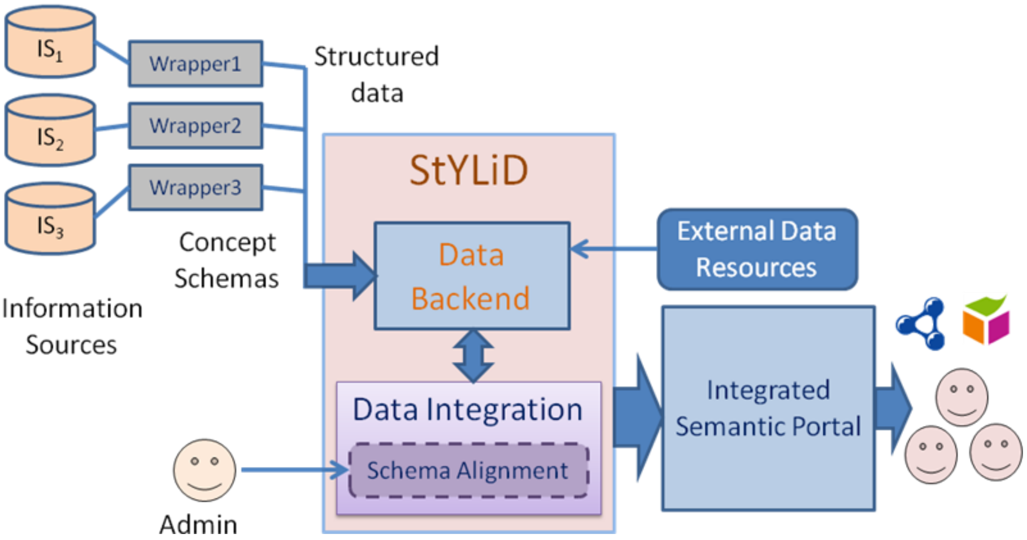


Figure 16. Integrated semantic portal scenario.

Besides these, StYLiD may also be used in other scenarios, both in public and closed settings. Some of them are as follows.

- It may be used for inter-departmental or inter-organizational information exchange and integration over separately maintained information systems.
- It can be used as a simple content management system or a data backend to drive dynamic online applications and websites.
- It may also be adapted as a structured blogging platform for personal or corporate use.
- It can be used for collaborative designing of conceptual schemas and serve as an inexpensive tool for rapid prototyping when the requirements are not well-defined initially.

3.8.3 Adaptation of the system to different scenarios

In principle, the approach is applicable for different scenarios as discussed above, from social information sharing to data integration to corporate and business use. Practically, the platform should be adapted to fit into such diverse scenarios. There are several control factors that vary across different purposes. Some of the factors are as follows.

- *Concepts and data acquisition method*. The nature of the concepts and the degree of detail and perfection in the definition would depend on the application requirement. In some cases, the concept schemas may be relatively stable, in other cases, they may be evolving rapidly and diverse. The concepts and data may be acquired from pre-existing sources or totally contributed by the community or a combination of both.

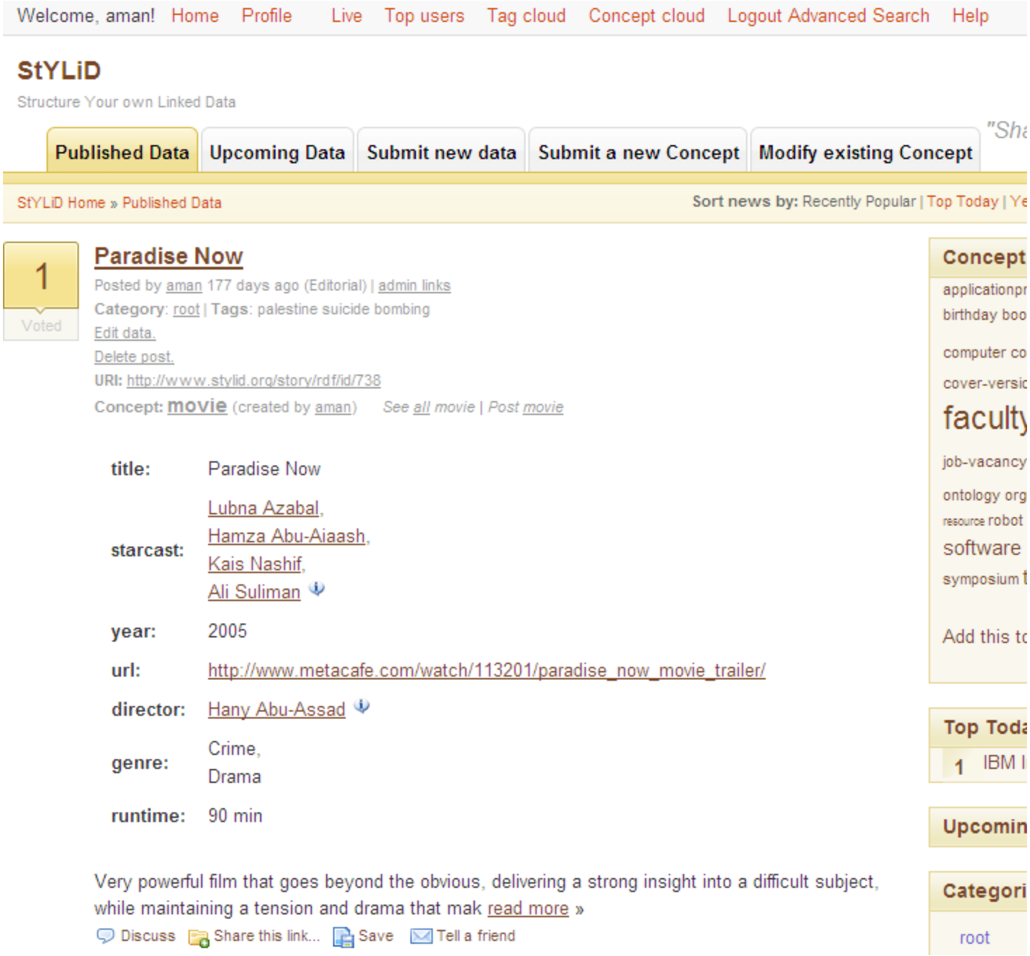
- *Motivation*. The degree of user motivation and the way to motivate also differs by the application.
- *Functionalities and constraints*. The functionalities and constraints required naturally depend upon the domain of application and requirements of the community.
- *Data quality*. The required degree of quality and consistency of data also depends upon the application.

The platform offers some flexibility to be adapted along such different factors. It supports multiple ways of creating concepts. Wrapper technologies may be used to import concepts and data from existing sources. The underlying framework provides a plug-in architecture. Hence, new functionalities can be added easily as required. The open source code can further be extended and the interface can be adapted to suit the purpose. The implementation of specialized functionalities may be delegated to the specific applications. Functionalities can be added to operate on domain specific types of data. Some constraints may be introduced to control the nature and quality of data or application specific heuristics may be used to clean the data later. Also application-specific queries and views can be created over the data to serve different information needs of the community.

In case of open online communities or open communities of common interest, motivation has a crucial role to gain the participation of ordinary people. In such scenarios, the application should be designed to be very easy and offer some enjoyment or instant benefit to the users. In case of closed communities or corporate use, personal motivation may not be essential, as the application would be meant to serve their requirements in the first place. The application should be tailored to serve their purpose well and kept easy enough. In such targeted applications, most concepts can be created beforehand, or extracted from existing data or systems. Some real applications of the system adapted for different purposes are demonstrated later in the Section 5.6.

3.9 Implementation

StYLiD has been implemented basically as a social web application to share structured linked data. It provides a structured data authoring interface for ordinary users without any knowledge of Semantic Web technologies. It allows users to define their own concept schemas freely and share different types of structured data they are interested in. It serves as a content management system to produce structured linked data. A screenshot of the system is shown in Figure 17.



Paradise Now

Posted by aman 177 days ago (Editorial) | admin links

Category: root | Tags: palestine suicide bombing

Edit data.

Delete post.

URI: http://www.stylid.org/story/rdf/id/738

Concept: movie (created by aman) See all movie | Post movie

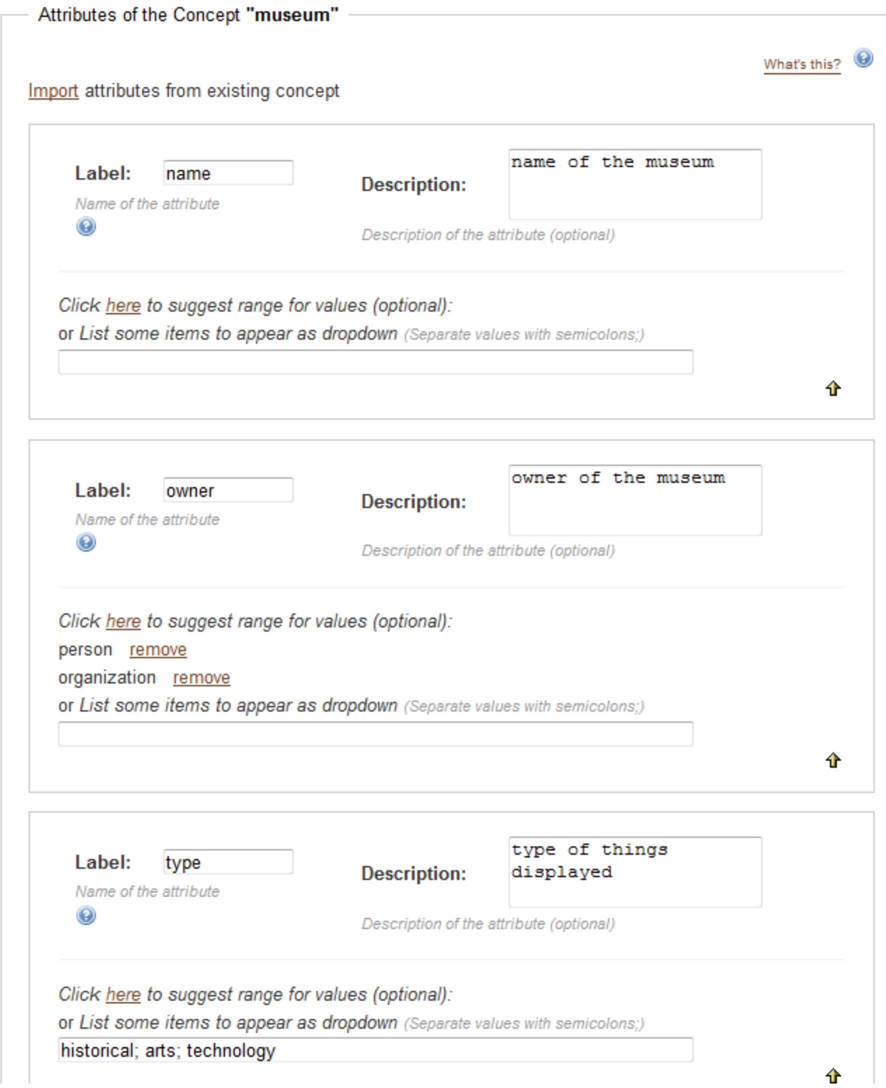
title:	Paradise Now
starcast:	Lubna Azabal, Hamza Abu-Aiaash, Kais Nashif, Ali Suliman
year:	2005
url:	http://www.metacafe.com/watch/113201/paradise_now_movie_trailer/
director:	Hany Abu-Assad
genre:	Crime, Drama
runtime:	90 min

Very powerful film that goes beyond the obvious, delivering a strong insight into a difficult subject, while maintaining a tension and drama that mak read more »

Figure 17. StYLiD screenshot.

3.9.1 Defining structured concept schemas

The users may may freely their own concept schemas by specifying the concept name, some description (optional) and a set of attributes. Each attribute has a name and some description (optional) as shown in Figure 18. This description is not necessarily the independent definition of the attribute. It is usually for clarifying the role of the attribute in the context. Trying to give abstract dictionary definitions to common labels would rather confuse the ordinary users. However, when the labels are not obvious more explanatory descriptions would be desired. Further, the user may select a set of concepts as the suggested value range (optional). Some possible values may also be enumerated which would appear as a drop-down list to help in data input (see Figure 25).


Attributes of the Concept **"museum"**

What's this?

Import attributes from existing concept

Label: name
Name of the attribute

Description: name of the museum
Description of the attribute (optional)

Click here to suggest range for values (optional):
or *List some items to appear as dropdown* (Separate values with semicolons;)

Label: owner
Name of the attribute

Description: owner of the museum
Description of the attribute (optional)

Click here to suggest range for values (optional):
person remove
organization remove
or *List some items to appear as dropdown* (Separate values with semicolons;)

Label: type
Name of the attribute

Description: type of things displayed
Description of the attribute (optional)

Click here to suggest range for values (optional):
or *List some items to appear as dropdown* (Separate values with semicolons;)
historical; arts; technology

Figure 18. Interface to create a new concept.

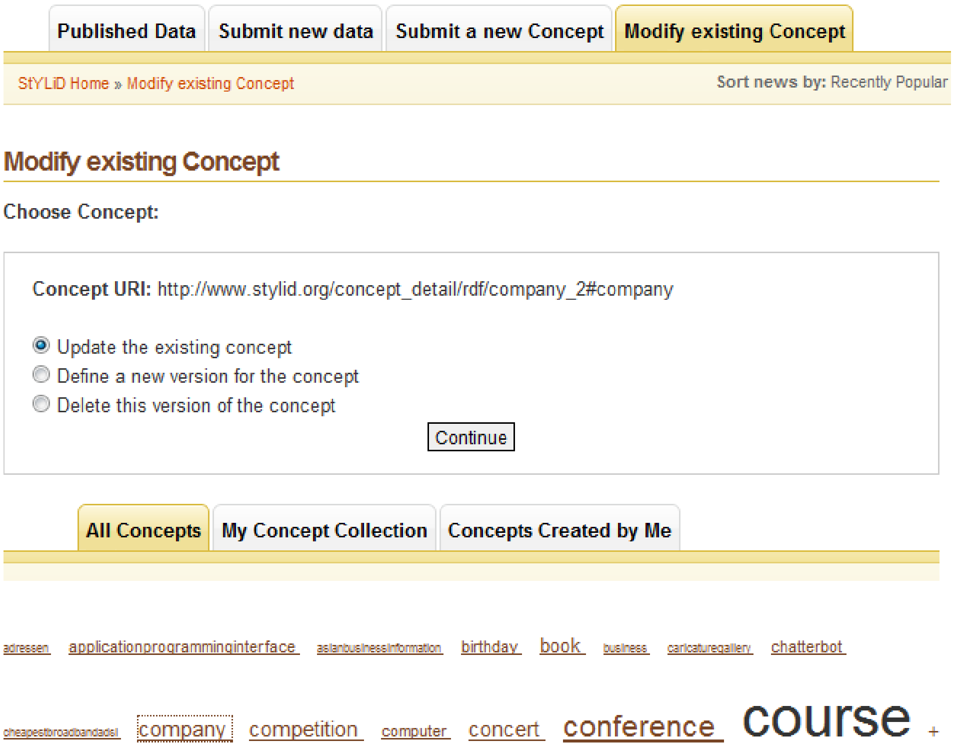
Reusing and updating existing concepts

Users do not need to define concepts from scratch. They may modify an existing concept into their own version using the interface shown in Figure 19. If the user is not aware of the existing concept and starts defining his own concept, the system automatically prompts that the concept already exists and provides the option to modify it or re-use it, as shown in Figure 20. However, users are not allowed to tamper with others' concept definitions. The system creates a copy of the concept and makes modifications on it. It keeps record of the source from which the concept was derived using the *dc:source*[61] property. Schema attributes from an existing concept can also be imported to define a new concept with similar structure, as shown in Figure 21.

Users may update definitions incrementally as and when needed. Users can update their own concept definitions keeping the existing instances consistent. Attributes can be added. However, if we need to rename or delete attributes of the concept, a new version of the concept should be defined to keep the existing data intact. Users may

[61] DC stands for the Dublin Core metadata standard

even define multiple versions of the same concept with the same name. Thus, concepts can evolve incrementally along with different versions. The system allows different users to define their own concepts having the same name.



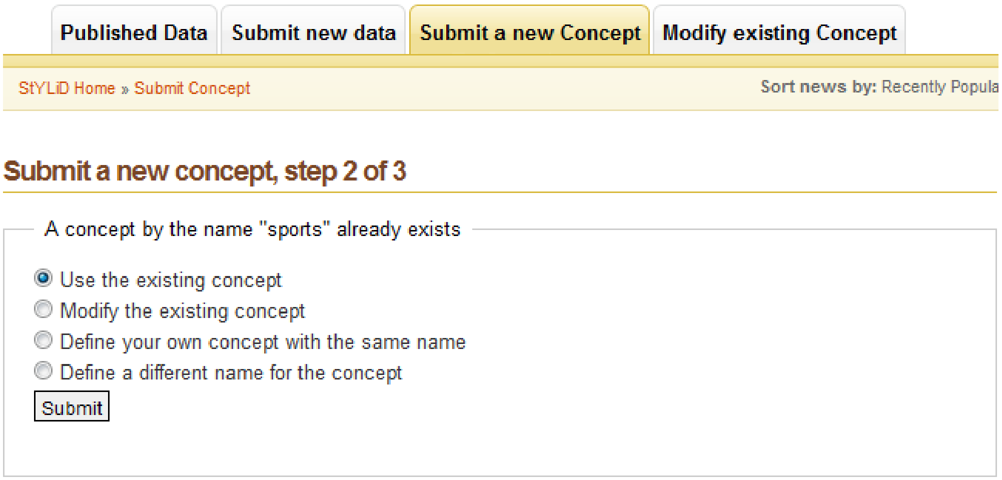
Modify existing Concept

Choose Concept:

Concept URI: http://www.stylid.org/concept_detail/rdf/company_2#company

● Update the existing concept
○ Define a new version for the concept
○ Delete this version of the concept

Figure 19. Interface to modify and reuse an existing concept.

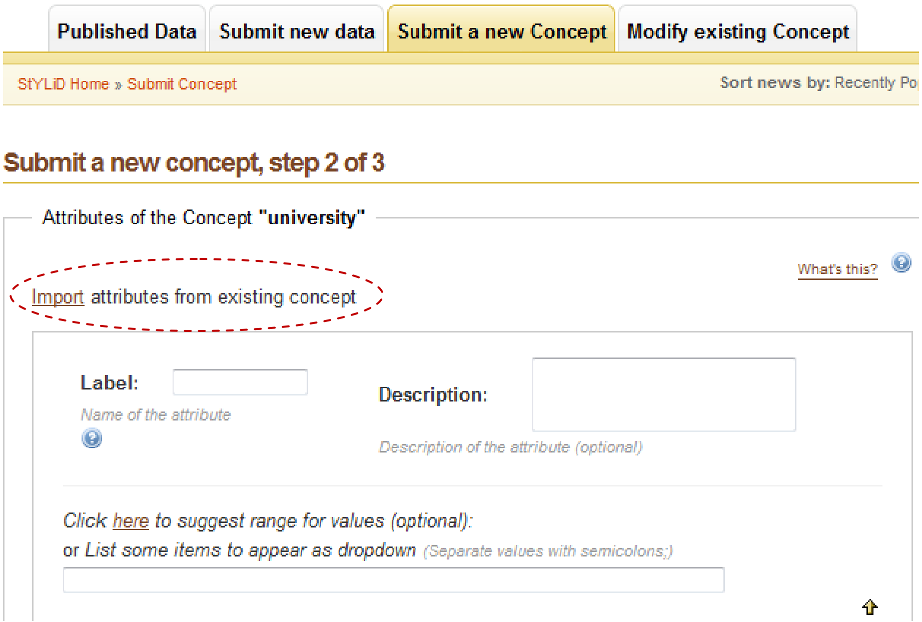


Submit a new concept, step 2 of 3

A concept by the name "sports" already exists

● Use the existing concept
○ Modify the existing concept
○ Define your own concept with the same name
○ Define a different name for the concept

Figure 20. Interface shown when defining a concept that already exists.



Submit a new concept, step 2 of 3

Attributes of the Concept **"university"**

Import attributes from existing concept

Label:
Name of the attribute

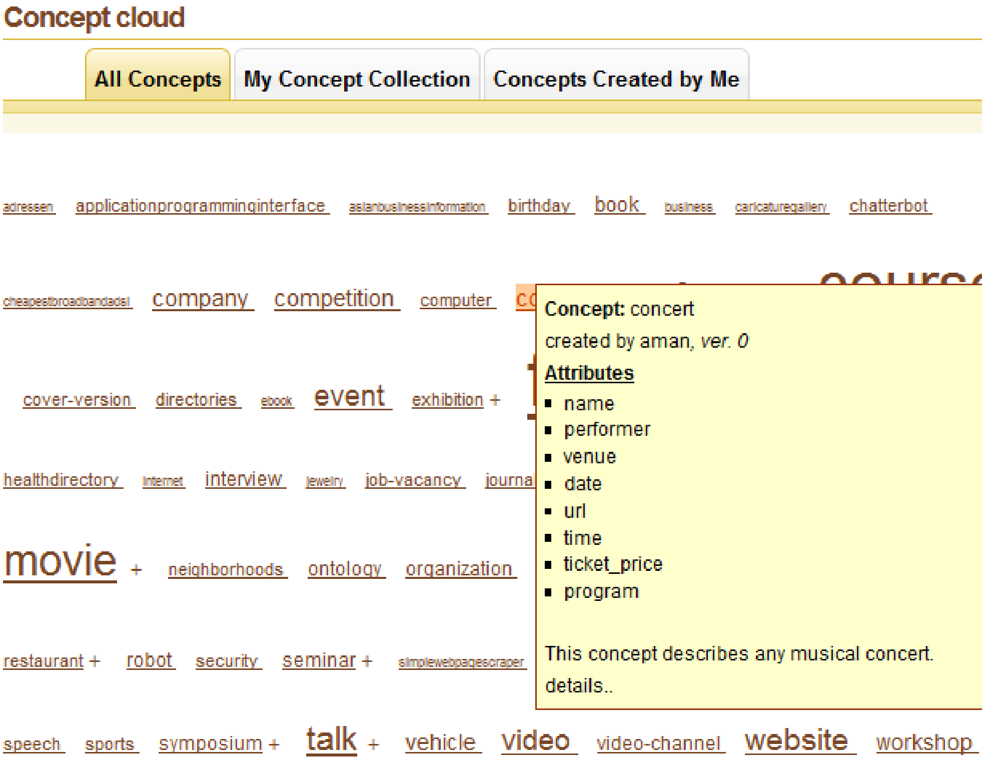
Description:
Description of the attribute (optional)

Click here to suggest range for values (optional):
or *List some items to appear as dropdown* (Separate values with semicolons;)

Figure 21. Importing attributes from existing concept.

Concept Cloud

All concepts are visualized in a *Concept Cloud* as shown in Figure 22. The user would be able to browse different types of data using the concepts in the concept cloud. Popular concepts appear bigger in the cloud. When the user hovers over any concept, the attributes and description of the concept are shown so that the concept and its structure can be understood instantly. Clicking on a concept retrieves all its instances.

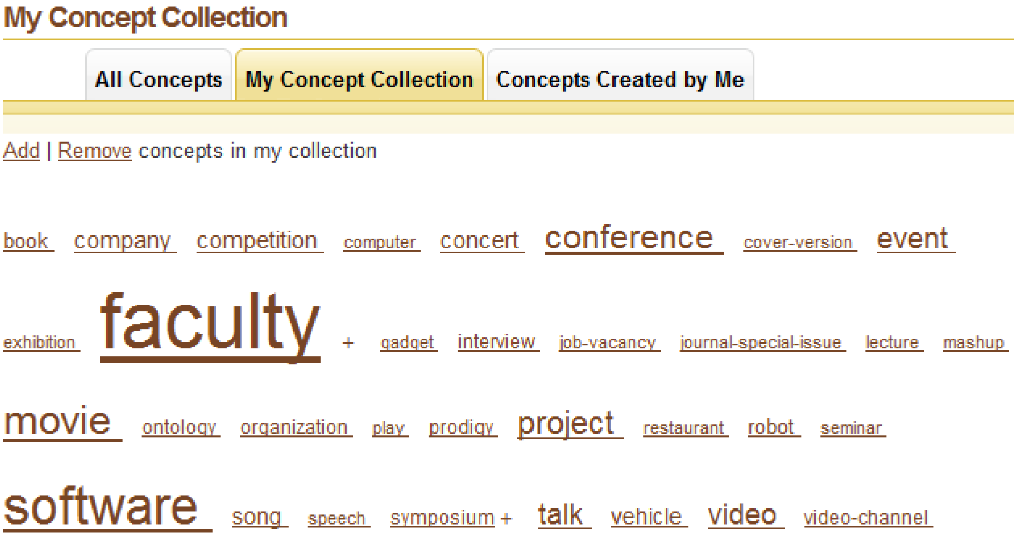


Concept: concert
created by aman, *ver. 0*
Attributes
- name
- performer
- venue
- date
- url
- time
- ticket_price
- program

This concept describes any musical concert.
details..

Figure 22. Concept Cloud in StYLiD.

A personal structured data space. The system also offers every user a personal structured data space called the “Concept Collection”, as seen in Figure 23. Concepts created or adopted by the user are automatically added to this collection. Besides these, users can also add any other useful concepts to their collection. The users need not be overwhelmed by the huge cloud of concepts defined by the large number of users. Moreover, the concept collection is also helpful to mark the concepts that the user has been using out of numerous concepts and different versions. The concepts actually created by the user are also shown in a separate tab.



My Concept Collection

All Concepts | My Concept Collection | Concepts Created by Me

Add | Remove concepts in my collection

book company competition computer concert conference cover-version event

exhibition faculty + gadget interview job-vacancy journal-special-issue lecture mashup

movie ontology organization play prodigy project restaurant robot seminar

software song speech symposium + talk vehicle video video-channel

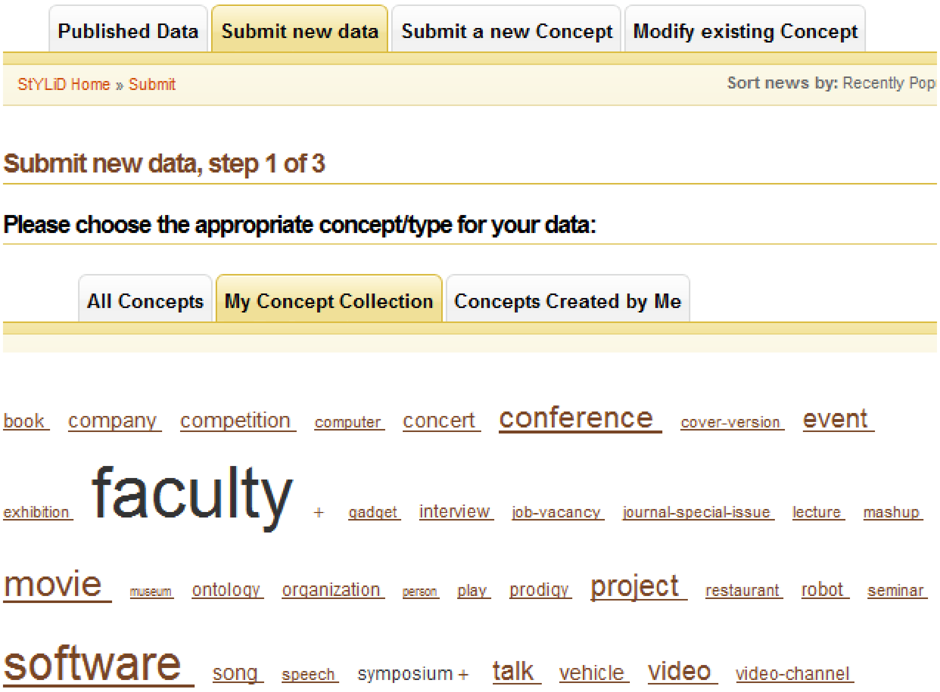
Figure 23. Personal concept collection.

3.9.2 Sharing structured data instances

Any user may enter instance data by selecting the desired concept, as shown in Figure 24, and filling the system generated online form, as shown in Figure 25. Data instances can be linked to each other by directly entering resource URIs as *data-links* for attribute values. The system helps the user to pick up suitable values by suggesting range of values for the attributes. The values can be picked up from a pop-up as shown in Figure 26. For the user, the data appears as usual hyperlinked entries (see Figure 17). However, the *data-links* behind can be crawled by machines to feed powerful linked data applications. The posted data instances are presented in a record view (by default) or a table view.

Flexible definitions and relaxed data entry

The concept definition may be incrementally updated later and new attributes may be added. New versions of the concept may be defined by different users or even the same user. The range of values defined for attributes, as seen in Figure 18 and Figure 25, is only suggestive and does not impose strict constraints. Rather the system assists the user to pick instances from the suggested range. However, any suitable value may be entered even though it is not in the suggested range. The concept may be updated later to change the suggestive range by including more range concepts or narrowing down to refine the range. An attribute of a concept can take a single value or multiple-values. The system accepts both literal values and resource URIs. Instances may be updated later to change a literal into a resource value by adding the URI. If the value is a resource URI, a human readable label may be entered along with the URI.

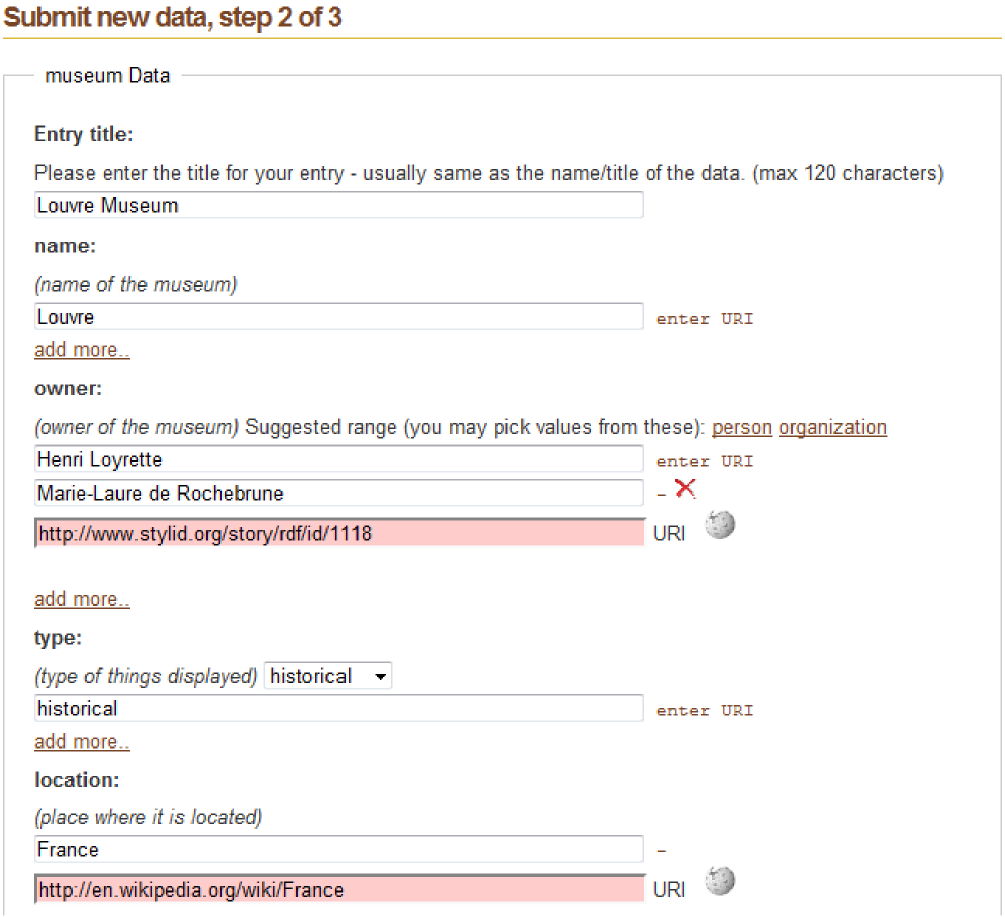


Submit new data, step 1 of 3

Please choose the appropriate concept/type for your data:

book company competition computer concert conference cover-version event exhibition faculty + gadget interview job-vacancy journal-special-issue lecture mashup movie museum ontology organization person play prodigy project restaurant robot seminar software song speech symposium + talk vehicle video video-channel

Figure 24. Selecting concept to input instance data.



Submit new data, step 2 of 3

museum Data

Entry title:

Please enter the title for your entry - usually same as the name/title of the data. (max 120 characters)

Louvre Museum

name:

(name of the museum)

Louvre enter URI

add more..

owner:

(owner of the museum) Suggested range (you may pick values from these): person organization

Henri Loyrette enter URI

Marie-Laure de Rochebrune - ✕

http://www.stylid.org/story/rdf/id/1118 URI

add more..

type:

(type of things displayed) historical

historical enter URI

add more..

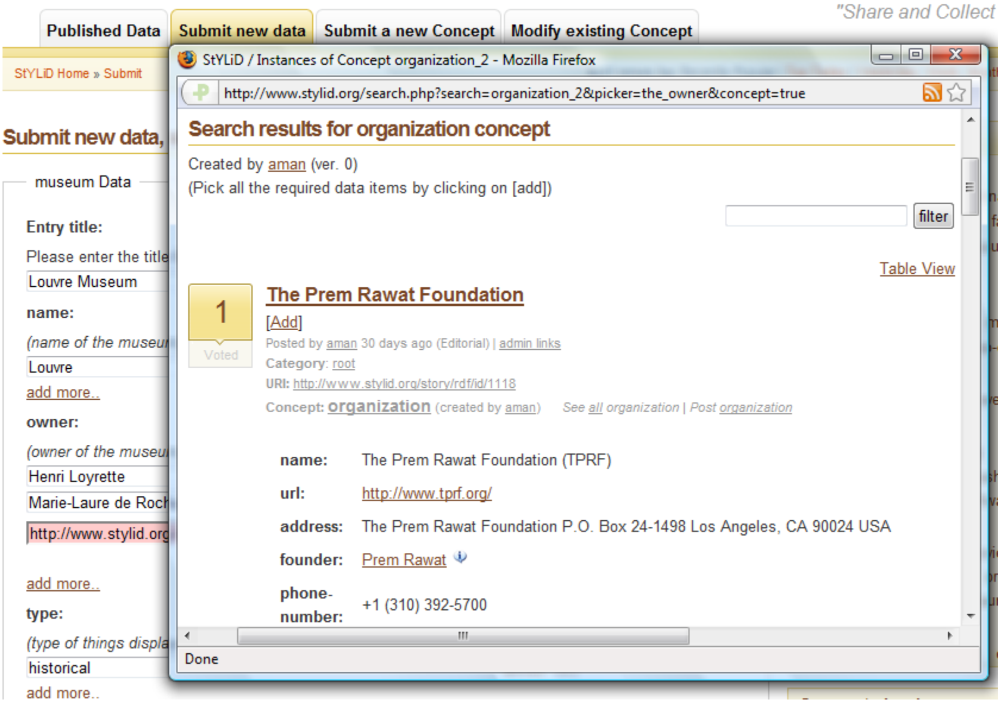
location:

(place where it is located)

France -

http://en.wikipedia.org/wiki/France URI

Figure 25. Interface to enter instance data.



http://www.stylid.org/search.php?search=organization_2&picker=the_owner&concept=true

Search results for organization concept

Created by aman (ver. 0)
(Pick all the required data items by clicking on [add])

filter

1 Voted

The Prem Rawat Foundation
[Add]
Posted by aman 30 days ago (Editorial) | admin links
Category: root
URI: http://www.stylid.org/story/rdf/id/1118
Concept: organization (created by aman) See all organization | Post organization

name: The Prem Rawat Foundation (TPRF)
url: http://www.tprf.org/
address: The Prem Rawat Foundation P.O. Box 24-1498 Los Angeles, CA 90024 USA
founder: Prem Rawat
phone-number: +1 (310) 392-5700

Submit new data,
museum Data
Entry title:
Please enter the title
Louvre Museum
name:
(name of the museu
Louvre
add more..
owner:
(owner of the museu
Henri Loyrette
Marie-Laure de Roch
http://www.stylid.org
add more..
type:
(type of things displa
historical
add more..

Figure 26. Pop-up list of suggested values.

3.9.3 Linked data generation

The system generates unique dereferenceable URIs for each concept, attribute and instance. The guidelines provided by Bizer et al. (2007a) for publishing linked data have been used.

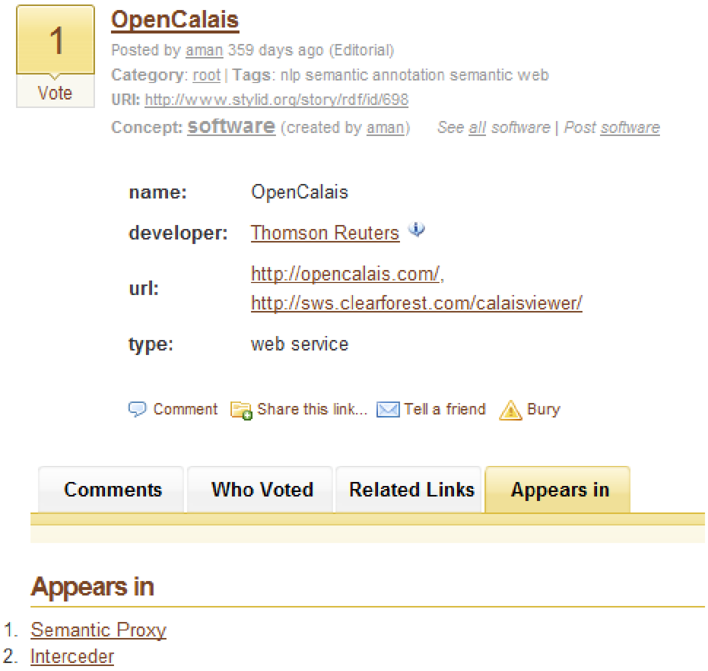
Each concept is uniquely identified by the concept name, its creator and the version number. An example URI for a concept "car", version 2, defined by the user with ID 1 would be *http://www.stylid.org/concept_detail/rdf/car_ver2_1#car*

An attribute is uniquely identified by the concept and the attribute name. For example, the URI for the price attribute of the car concept would be *http://www.stylid.org/concept_detail/rdf/car_ver2_1#price*

The hash URI retrieves the RDF document describing the concept and dereferences to the RDF description.

An instance is uniquely identified by the system generated ID. For example, the URI for an instance with ID 623 would be *http://www.stylid.org/story/rdf/id/623*. The URI dereferences to the RDF description of the instance by an HTTP 303 redirect.

For both types of URI, content negotiation is used to return the RDF description in case of "application/rdf+xml" request and HTML otherwise. The description also contains backlinks from other instances that link to the instance. For the users, the backlinks are shown under the instance as an "Appears in" list, similar to trackbacks. In Figure 27, the instances "Semantic Proxy" and "Interceder" are linked to the instance "OpenCalais". Hence, the backlinks are automatically shown for the latter.



OpenCalais

Posted by aman 359 days ago (Editorial)

Category: root | Tags: nlp semantic annotation semantic web

URI: http://www.stylid.org/story/rdf/id/698

Concept: software (created by aman) See all software | Post software

name: OpenCalais

developer: Thomson Reuters

url: http://opencalais.com/, http://sws.clearforest.com/calaisviewer/

type: web service

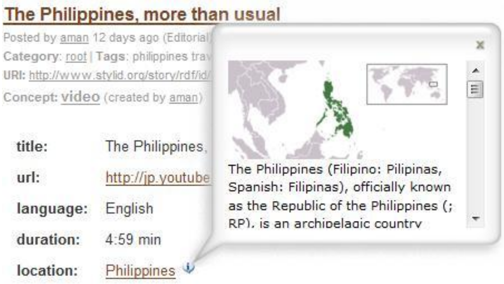
Appears in

1. Semantic Proxy
2. Interceder

Figure 27. Backlinks to a data instance in StYLiD.

Linking to Wikipedia and External Resources

The user may directly enter any external URI for an attribute value. The system currently provides some support to link to Wikipedia contents. The familiar Wikipedia icon appears next to the URI field (see Figure 25). When the user clicks on the icon it searches for the Wikipedia page about the text attribute value typed by the user. The user may copy the Wikipedia page URL as the URI. Transparent to the user, the system converts it into the corresponding DBpedia (Auer et al., 2007) URI. Unlike DBpedia, Wikipedia is well understood by general people. The users would be motivated to link to Wikipedia pages to make their data more informative, interesting and useful. Some short description and depiction from Wikipedia (through DBpedia) is pulled dynamically and shown as an annotation balloon as shown in Figure 28.



The Philippines, more than usual

Posted by aman 12 days ago (Editorial

Category: root | Tags: philippines tra

URI: http://www.stylid.org/story/rdf/id/

Concept: video (created by aman)

title: The Philippines,

url: http://jp.youtube

language: English

duration: 4:59 min

location: Philippines

The Philippines (Filipino: Pilipinas, Spanish: Filipinas), officially known as the Republic of the Philippines (; RP), is an archipelagic country

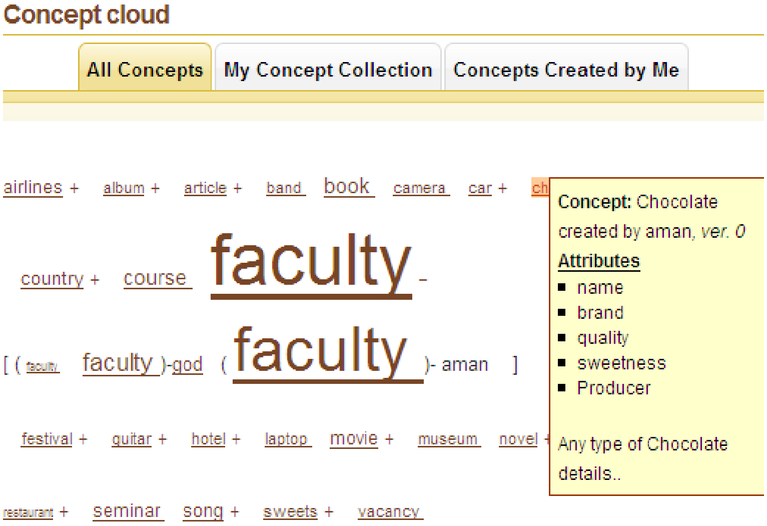
Figure 28. Annotation with Wikipedia contents using DBpedia linked data.

3.9.4 Concept consolidation

Concepts defined by different users with the same name are automatically grouped together in the concept cloud as shown in Figure 29. This group of concepts can be aligned to form a consolidated concept. However, the user is free to consolidate any concepts if he/she considers them the same or similar. The system helps in identifying similar concepts by grouping similar concepts. This is described and illustrated later in Section 3.9.5.

As shown in the figure, a consolidated concept can be expanded into a *sub-cloud* showing all the versions defined by different users, labeled with the creator name and version number. These are the *constituent concepts* of the *consolidated concept* as defined in Section 3.6 (def. 2). In the sub-cloud, multiple versions defined by the same user are sub-grouped together. In Figure 29, the “faculty” concept has been expanded to show two versions by the user “god” and one version by “aman”. The sizes of all versions in the sub-cloud add up to form the size of the consolidated concept.

Clicking on the consolidated concept retrieves all instances of all its versions. Instances of the versions defined by a single user can also be listed by clicking on the user name. Hence, the concept cloud helps in browsing concept instances at different levels of granularity.



Concept: Chocolate
created by aman, *ver. 0*
Attributes
- name
- brand
- quality
- sweetness
- Producer

Any type of Chocolate
details..

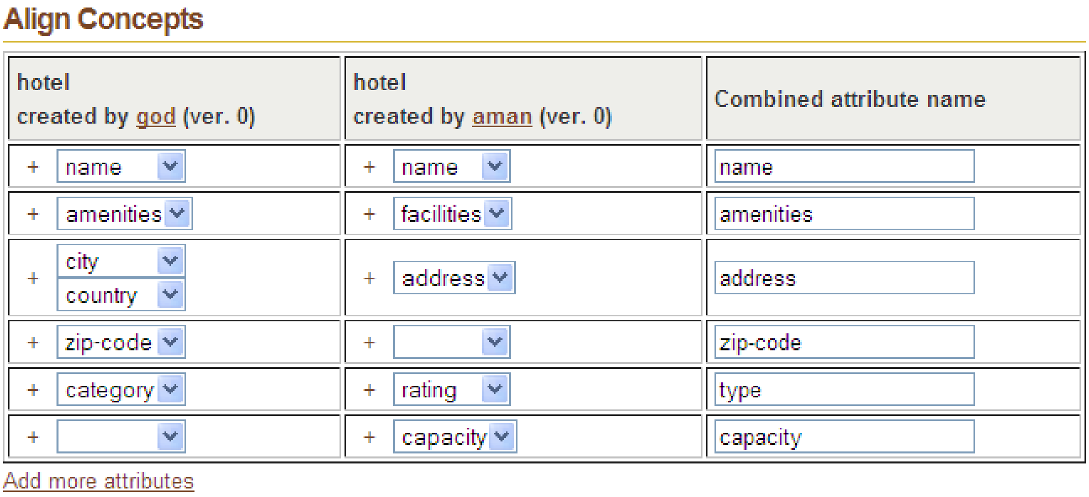
Figure 29. Consolidated concept cloud.

Semi-automatic schema alignment

The constituent concepts in a consolidated group can be aligned to produce a uniform and integrated view. The system automatically suggests alignments between the attributes, as shown in Figure 30. The semi-automatic alignment interface is invoked either by explicitly aligning the set of concepts or when a user attempts to view instances of a consolidated concept in a single table view.

Matching attributes are automatically selected in the form-based interface. The Alignment API[62] (Euzenat, 2004) implementation with its WordNet[63] extension has been used for the purpose. It utilizes a WordNet based similarity measure between attribute labels to find alignments. This may be replaced by more sophisticated alignment methods in the future. However, more sophisticated user interfaces may be required and it may be difficult to maintain usability keeping ordinary users in mind. Each set of aligned attributes forms a consolidated attribute.

[62] http://alignapi.gforge.inria.fr/
[63] http://wordnet.princeton.edu/



Align Concepts

hotel created by god (ver. 0)	hotel created by aman (ver. 0)	Combined attribute name
+ name	+ name	name
+ amenities	+ facilities	amenities
+ city country	+ address	address
+ zip-code	+	zip-code
+ category	+ rating	type
+	+ capacity	capacity

Add more attributes

Figure 30. Aligning the attributes of multiple concepts.

No matter how sophisticated techniques we use, it is not possible to make the alignment fully automatic and accurate. Sometimes the mappings may require deeper human understanding than mere linguistic similarity. So it is necessary to have the user in loop to complete the process by adding or modifying mappings not suggested by the system correctly. Any user, who wants to retrieve or search over all data from different sources in a unified form, can make the alignment, assisted by the automatic suggestions. Currently, the system only suggests one-to-one mappings. However, the user can add many-to-one mappings too as shown in Figure 30. While unfolding a query from one to many attributes, the union of the values of the multiple attributes is considered. Further, the alignments created in the systems are at a generic level. More complex mappings requiring transformations (for e.g., currency conversions, etc) can be handled through mechanisms such as the conversion function network proposed by Firat et al. (2007).

Collaborative schema alignment. Completing the alignment can be done collaboratively. An individual may perform the alignment up to his needs and understanding. The mappings are saved by the system. The alignment can be updated incrementally as more concepts are added to the group or the existing ones updated. Other users may successively add the missing parts and refine the alignment. In theory, conflicts may be resolved in a wiki manner. Thus, both machine intelligence and human intelligence are used in getting the concepts aligned. This forms the alignment $\mathcal{A}$ defined in Section 3.6 (def. 2, 3). Once a proper alignment is in place, rest of the users can directly access the unified data. Others need not do the alignment again. Hence, the action of one or few can benefit all in the community.

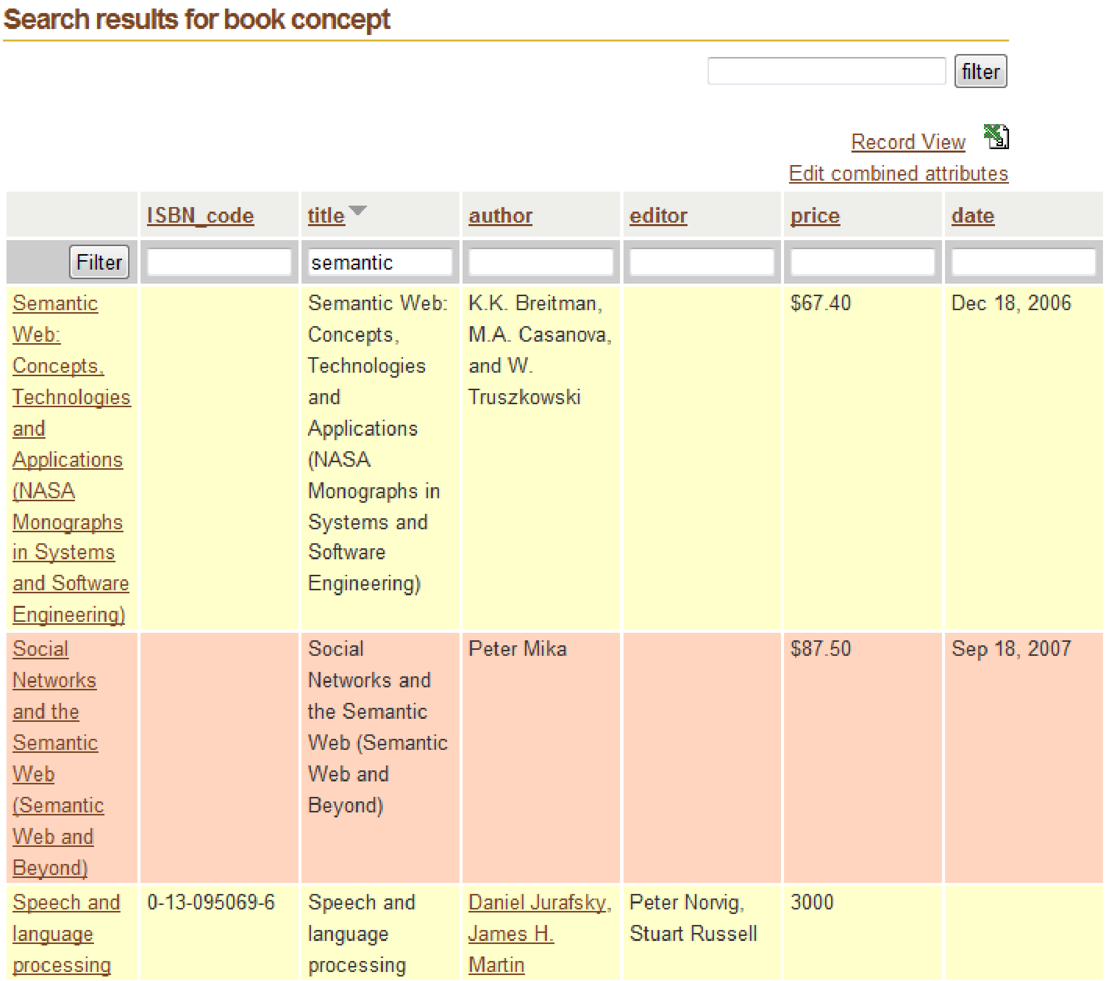
The alignment API represents the schema level linking in an expressive alignment specification language capable of representing complex alignments (Euzenat, 2004). Although the system currently does not determine complex alignments, this allows for more sophisticated mappings in the future. The alignments are also represented using the alignment ontology[64] (Hughes & Ashpole, 2004) and saved by the system. The API also has provisions to export the alignment in other formats like C-OWL, SWRL, OWL axioms, XSLT, SEKT-ML and SKOS (Euzenat, 2004). The alignments are

[64] http://www.atl.lmco.com/projects/ontology/

published as schema level linked data which allows machines to understand the relations among the data sources. This, in turn, can help in linking data instances by providing a basis to compare them schematically.

*A unified view***.** A unified schema is formed by consolidating the multiple concept schemas. Each set of aligned attributes is mapped to a single consolidated attribute. This *consolidated attribute* (def. 2) is the *view* of a corresponding attribute (def. 6) from each constituent concept as defined in Section 3.6. The system automatically fills a name for each consolidated attribute, as shown in Figure 30, though the user may rename it as desired. The user may even remove attributes from the unified view, if not required. Thus, the user can create a unified view, customized according to his need, and view heterogeneous data in a uniform table (as shown in Figure 31). This table corresponds to the *consolidated view of instances* described in Section 3.6 (def. 8). The table of structured data can be sorted and filtered by different fields. For example, the figure shows the unified list of books sorted by title and filtered with the word 'semantic' in the title. The table of data can even be exported to spreadsheet applications like Microsoft Excel for desired processing.

To have all instances of all the concepts listed, all the concepts should be *mapped* (def. 4). The consolidated concept should be *grounded* (def. 5) to have no empty attributes in the unified table. The user is notified if all concepts are not mapped or the consolidated concept is not grounded.



Search results for book concept

	ISBN_code	title ▼	author	editor	price	date
Filter		semantic				
Semantic Web: Concepts, Technologies and Applications (NASA Monographs in Systems and Software Engineering)		Semantic Web: Concepts, Technologies and Applications (NASA Monographs in Systems and Software Engineering)	K.K. Breitman, M.A. Casanova, and W. Truszkowski		$67.40	Dec 18, 2006
Social Networks and the Semantic Web (Semantic Web and Beyond)		Social Networks and the Semantic Web (Semantic Web and Beyond)	Peter Mika		$87.50	Sep 18, 2007
Speech and language processing	0-13-095069-6	Speech and language processing	Daniel Jurafsky, James H. Martin	Peter Norvig, Stuart Russell	3000	

Figure 31. Unified table view of instances.

Multiple concept generalizations

Similar or related concepts may also be consolidated to form a more generic concept. For example, ‘hotel’ and ‘apartment’ concepts can be consolidated to form an ‘accommodation’ concept. Then all hotels and apartments can be searched together conveniently as ‘accommodation’. Different users may group and consolidate the same concepts in different ways depending upon their requirements or perspectives. For example, another user may group ‘hotel’ with ‘building’ to create a ‘landmark’ concept if he is interested in sight-seeing landmarks. Hence, the same set of concepts may be organized in bottom-up fashion in multiple ways by different people.

Consolidated linked data instances

If we consider the concept instances, consolidation results in two levels of linked data – *consolidated/global* linked data and *local/contextual* linked data.

The data originating from an individual source fully confirms to the conceptualizations within the context of the source though it may not be consistent in a different context for a different source. So this data can be treated as *local/contextual* with respect to the source. It includes all and only the original data instances from the source. The local/contextual linked data mainly serves the local requirements that need to be fulfilled for the application and context associated with the source, irrespective of other sources. These are the requirements of the direct users of the information source who share the same local context and perspectives. With the local linked data, it is easy to maintain compatibility with existing legacy systems and useful semantic applications may be provided at the local level.

On the other hand, the data source is also exposed for integration with other sources. The *consolidated/global* linked data is the result of integration of several local linked data at the schema level. It provides an integrated view of the complete collection of data instances derived from all the sources. All global data can be treated uniformly irrespective of the source of origin. It confirms to the unified model compatible to all the sources involved. Schema elements that are not consistent with a source would not appear in the consolidated view. So some context-dependent information and requirements may not be retained in the consolidated view. Therefore, both local and global linked data are maintained and shared.

While combining partial schemas from multiple sources a consolidated vocabulary to structure data instances gradually emerges. Hence, while the Semantic Web is usually considered for data integration at the record level, data integration, in the first place, can produce rich linked data and emerging vocabulary for the Semantic Web. The consolidation also serves information exchange among the different local linked data sources. The schema level mappings relate these two levels of linked data and allow information translation to suit different needs.

The consolidated/global linked data may be materialized or simply used as immaterialized views depending upon the situation and implementation choice. If the local sources have stabilized, i.e., further updates or additions to the local database would rarely be done, and integrated data is more significant, it would be better to materialize the consolidated/global linked data. When the local sources update rapidly and the local view is more important, the consolidated linked data may be computed only when needed without materializing the instances. Otherwise it would be difficult to keep the consolidated linked data up-to-date.

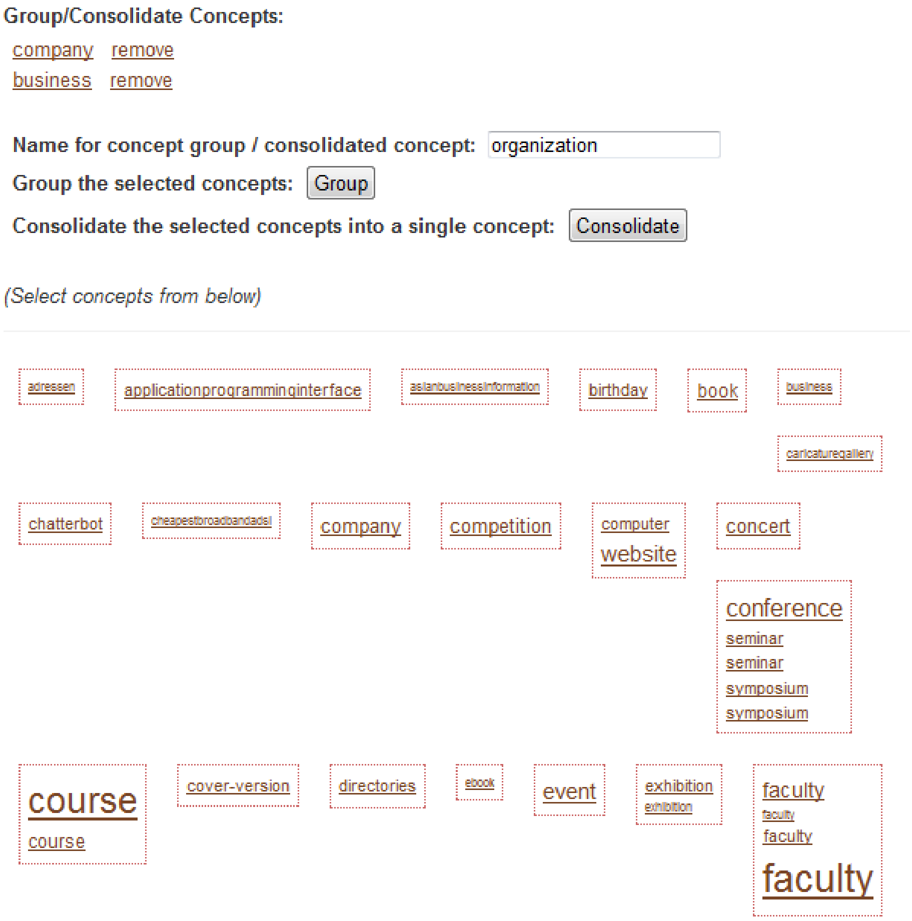
3.9.5 Concept grouping and organization

Concepts can be grouped and organized semi-automatically by the community. Concepts are automatically grouped by the system, as shown in Figure 32, using the algorithm for concept schema similarity calculation described earlier in Section 3.7.1. However, the user is free to create any concept grouping as desired. The concept groups can be maintained with appropriate names as shown in Figure 33.

Browsing concept groups. Browsing different types of concepts and data becomes more convenient and intuitive when related concepts are grouped together. The system provides an interface to explore the named concept groups as shown in Figure 34. Besides this, clusters of similar concepts can be effectively visualized using tools like Cytoscape[65] which is an open source platform for visualizing graphs with large number of nodes and relations. A screenshot of Cytoscape is shown in Figure 35.

Concepts are the nodes and relation edges may be drawn between similar concepts weighted by the similarity value (*ConceptSim*). Visualization techniques are available to show more similar nodes close to each other than less similar nodes. This provides a clearer and meaningful visualization of the groups of similar concepts.

[65] http://www.cytoscape.org/



Group/Consolidate Concepts:

company remove
business remove

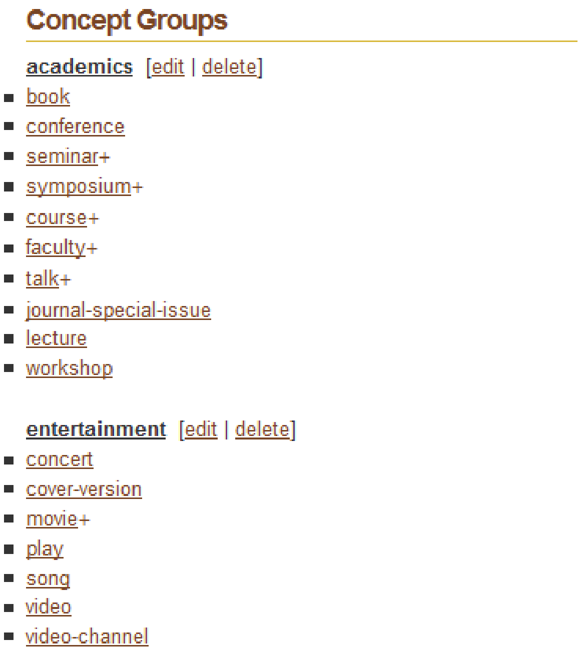
Name for concept group / consolidated concept: organization

Group the selected concepts: Group

Consolidate the selected concepts into a single concept: Consolidate

(Select concepts from below)

Figure 32. Interface for semi-automatic grouping and consolidation of concepts.



Concept Groups

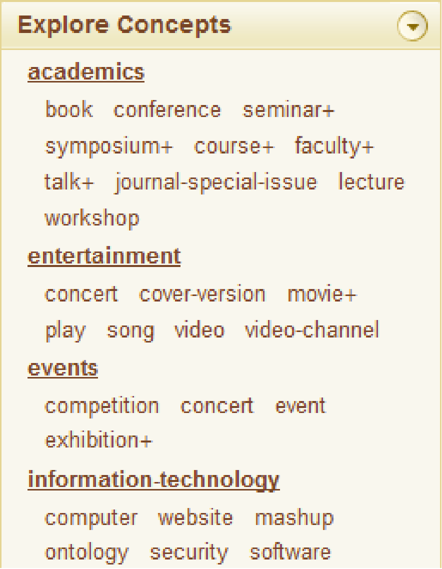
academics [edit | delete]

- book
- conference
- seminar+
- symposium+
- course+
- faculty+
- talk+
- journal-special-issue
- lecture
- workshop

entertainment [edit | delete]

- concert
- cover-version
- movie+
- play
- song
- video
- video-channel

Figure 33. Named concept groups.



Explore Concepts

academics

book conference seminar+
symposium+ course+ faculty+
talk+ journal-special-issue lecture
workshop

entertainment

concert cover-version movie+
play song video video-channel

events

competition concert event
exhibition+

information-technology

computer website mashup
ontology security software

Figure 34. Interface for browsing grouped concepts.

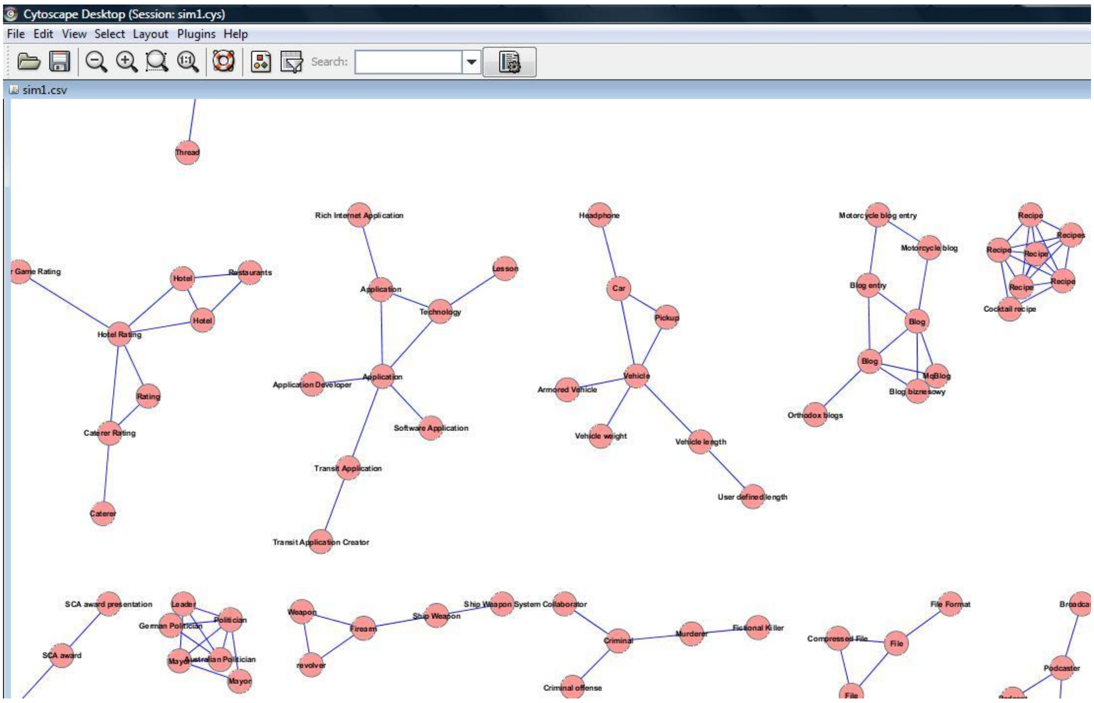


Figure 35. Visualization of similar concept groupings using Cytoscape.

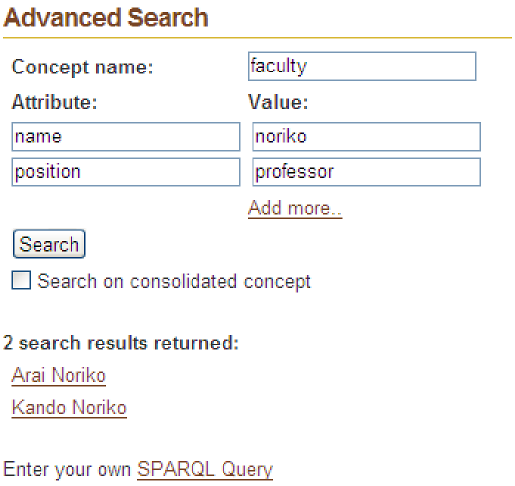
These groups of concepts may further be connected up into a single network with the help of relations in WordNet or other semantic networks like the ConceptNet (Liu & Singh, 2004) or an upper ontology like OpenCyC.[66]

In this way, concepts can be grouped and organized collaboratively by the community along with some automatic assistance. This results in an informal organization of the user-defined concepts which evolves according to the needs of the community.

[66] http://www.opencyc.org/

3.9.6 Structured search

The system provides a structured search interface, as shown in Figure 36, to search linked data instances of a concept by specifying attribute, value pairs as criteria. When the search is done over a consolidated concept, all the constituent concept instances are also searched. The query terms are unfolded from the consolidated concept attributes to the aligned attributes of the constituent concepts as described in Section 3.6 (theorem 3). The system also provides a SPARQL query interface, as shown in Figure 37, for open external access. The SPARQL query results can be obtained in HTML or XML format. Applications can parse the queries results in XML format and use them for further automated processing.



Advanced Search

Concept name: faculty

Attribute:	Value:
name	noriko
position	professor

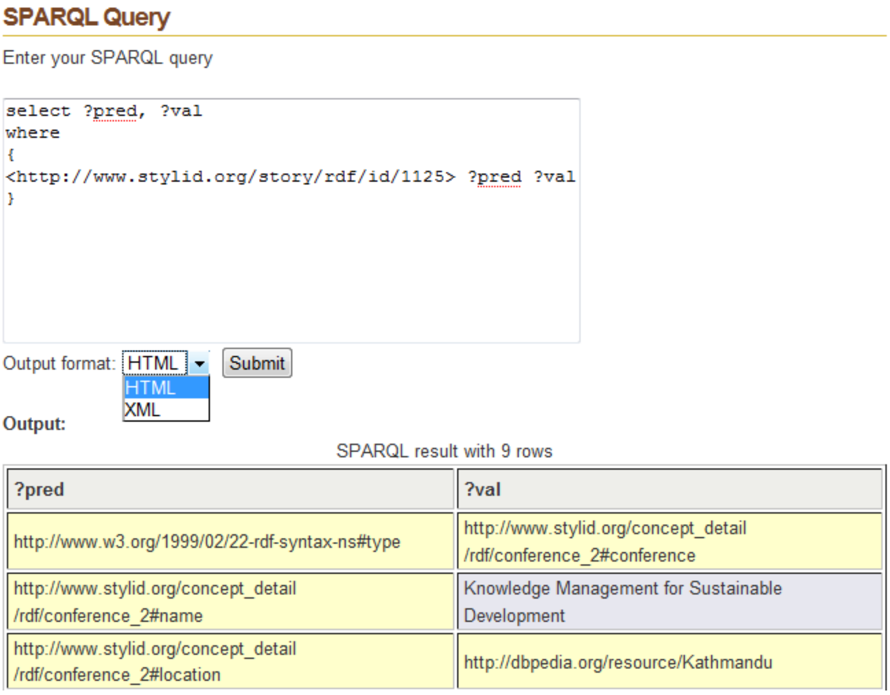
□ Search on consolidated concept

2 search results returned:

Arai Noriko
Kando Noriko

Enter your own SPARQL Query

Figure 36. Structured search interface.



SPARQL Query

Enter your SPARQL query

```
select ?pred, ?val
where
{
<http://www.stylid.org/story/rdf/id/1125> ?pred ?val
}
```

Output format: HTML

Output:

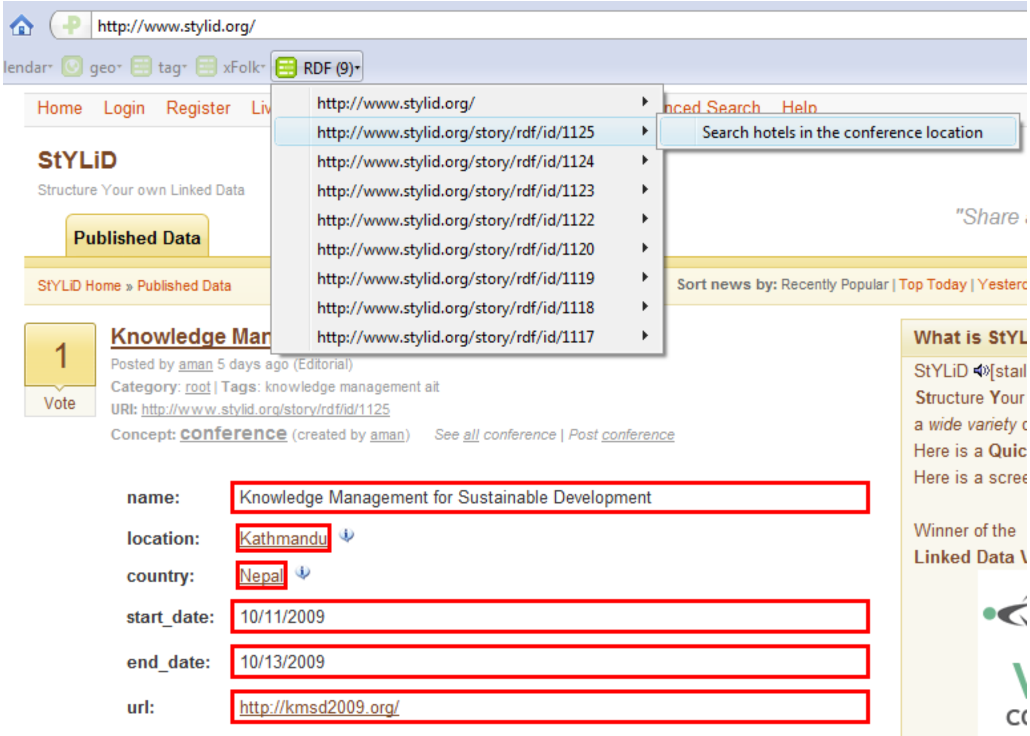
SPARQL result with 9 rows

?pred	?val
http://www.w3.org/1999/02/22-rdf-syntax-ns#type	http://www.stylid.org/concept_detail /rdf/conference_2#conference
http://www.stylid.org/concept_detail /rdf/conference_2#name	Knowledge Management for Sustainable Development
http://www.stylid.org/concept_detail /rdf/conference_2#location	http://dbpedia.org/resource/Kathmandu

Figure 37. SPARQL query interface.

3.9.7 Embedding machine readable data

Besides serving RDF when URIs are dereferenced, the system also embeds machine understandable data in the HTML posts using RDFa. Many useful RDFa tools and plug-ins are available[67] and we may expect more in the future. The use of RDFa has also been demonstrated by works on semantic clipboard (Reif et al., 2006; Möller et al., 2007) which would allow users to copy structured data into useful desktop applications. Users with some programming knowledge may even code small scripts with the Operator[68] browser extension to create useful operations for different types of data. Operator is an extension for Firefox that adds the ability to interact with semantic data embedded in web pages. The Figure 38 shows a custom-made Operator plugin that provides the operation "Search hotels in the conference location" for conference instances. The parts of the HTML page containing embedded RDFa data are also shown highlighted.


http://www.stylid.org/
lendar
geo
tag
xFolk
RDF (9)
Home
Login
Register
Liv
nced Search
Help
http://www.stylid.org/
http://www.stylid.org/story/rdf/id/1125
Search hotels in the conference location
http://www.stylid.org/story/rdf/id/1124
http://www.stylid.org/story/rdf/id/1123
http://www.stylid.org/story/rdf/id/1122
http://www.stylid.org/story/rdf/id/1120
http://www.stylid.org/story/rdf/id/1119
http://www.stylid.org/story/rdf/id/1118
http://www.stylid.org/story/rdf/id/1117
StYLiD
Structure Your own Linked Data
Published Data
"Share
StYLiD Home » Published Data
Sort news by: Recently Popular | Top Today | Yesterd
1
Vote
What is StYL
StYLiD [stai
Structure Your
a wide variety
Here is a Quic
Here is a scree
Winner of the
Linked Data

Knowledge Man

Posted by aman 5 days ago (Editorial)

Category: root | Tags: knowledge management ait

URI: http://www.stylid.org/story/rdf/id/1125

Concept: conference (created by aman) See all conference | Post conference

name: Knowledge Management for Sustainable Development

location: Kathmandu

country: Nepal

start_date: 10/11/2009

end_date: 10/13/2009

url: http://kmsd2009.org/

Figure 38. Providing operations on embedded data using custom Operator script.

3.9.8 Effective usage of the system

As discussed in the application scenarios the system may be used for general purpose or within specific domains and communities. The basic workflow consists of user actions like defining new concepts, posting data instance, grouping related concepts and consolidating similar concepts by aligning their schema. Although help manuals and some initial training may be useful, we want that zero or minimal training should

[67] http://esw.w3.org/topic/RDFa
[68] http://www.kaply.com/weblog/operator/

be required. Users are also assisted automatically by the system to some extent. The usage of the system also depends upon domain specific requirements and nature of the user community. The system gives freedom to the users in order to accommodate personal requirements and perspectives. It can be self-regulated though some moderation may be needed for control. As existing social applications have demonstrated, we can still expect reasonable contributions from users and meaningful knowledge structures to emerge.

3.9.9 Technological details

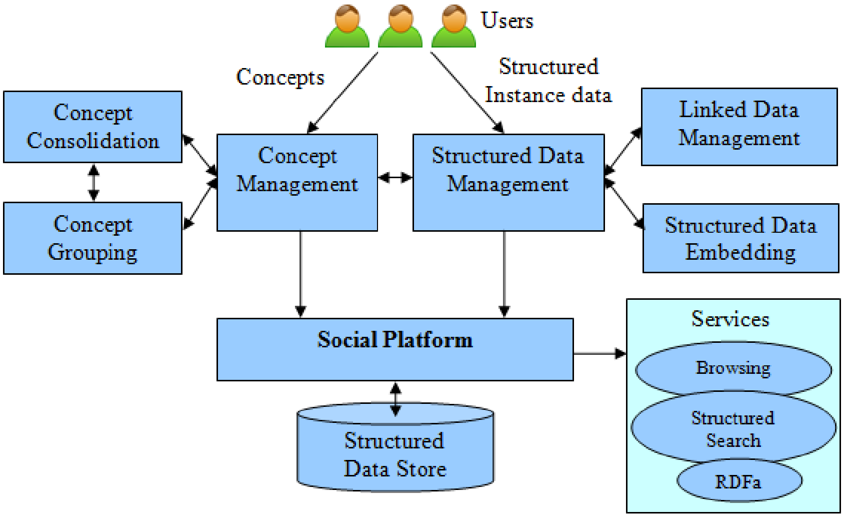


Figure 39. Implementation architecture.

Figure 39 shows the implementation architecture of StYLiD. It is built upon a social software platform for harnessing user contributions. It consists of the following functional components.

Social platform. The social platform provides all the basic features such as content management, assessing popularity of contents, user management, social networking and communication among users. StYLiD has been built upon Pligg,[69] a popular Web 2.0 content management system. It is an open source social software with a long list of useful features and a strong community support. Pligg has an extensible plug-in architecture which allows us to extend it for structured data and semantic capabilities. Further, unlike other social bookmarking platforms, it also supports extra data fields besides the bookmarked URL. Pligg has been built on PHP and MySQL.

Concept management. The concept management component enables the users to define their own structured concepts. The component handles the different versions of concepts defined by different users.

Concept consolidation. The concept consolidation component consolidates multiple versions of a concept defined by several users. The schema alignment component is also embedded in this. It maps the different versions by aligning attributes and provides a unified interface for the consolidated concept.

[69] http://www.pligg.com/

Concept grouping. This component is responsible for concept schema similarity calculation and semi-automatic grouping of similar concepts. It is interlinked with the concept consolidation unit as similar concepts can possibly be consolidated.

Structured data management. The structured data management component gathers the instance data contributions from users based on the concept schemas.

Linked data management. This component is responsible for opening data to the Semantic Web using linked data principles. This component handles URI management by assigning each of the concepts and instances a unique dereferenceable URI. Structured data items are linked using the URIs.

Structured data embedding. The structured data embedding component embeds structured data in HTML posts. The RDFa format has been used for this purpose. RDFa is W3C supported and a comparison with other embedded formats[70] indicates that it is a reasonable choice.

Structured data store. All the concepts and structured data contributed by users are stored in the structured data store coupled with the social software. The structured data instances are stored as RDF triples in a MySQL database. Concept schemas are represented using the RDFS[71] vocabulary which provides enough expressive power for our purpose. In fact, RDFS is recommended instead of OWL for keeping the concept definitions flexible and not constrained. We have used the RDF API for PHP (RAP)[72] as the Semantic Web framework which is a programming interface over the RDF data store.

Services. This component handles services to utilize the structured data like structured browsing, search and query and operating on the embedded RDFa data.

[70] http://bnode.org/blog/2007/02/12/comparison-of-microformats-erdf-and-rdfa
[71] http://www.w3.org/TR/rdf-schema/
[72] http://www.seasr.org/wp-content/plugins/meandre/rdfapi-php/doc/

3.10 Summary

This chapter discussed about concepts and established that concepts cannot be defined perfectly and uniquely. Concepts are vague representations of categories and depend on personal perspectives, knowledge and level of expertise. Hence, multiple conceptualizations are bound to exist. However, data integration and schema alignment methods can help in relating and combining such multiple conceptualizations. An approach for community-driven knowledge base creation by loose collaboration was proposed. Multiple local knowledge bases with different conceptualizations can co-exist and can be combined to form a global knowledge base. The proposed approach allows people to create their own concepts freely for sharing different types of data. The multiple concept definitions can be consolidated to form unified concept definitions. A theoretical formalization of concept consolidation was presented. Concepts can further be grouped and organized facilitating the emergence of lightweight ontologies in a bottom-up fashion. The StYLiD system implementing this approach was described in detail. Aspects of user motivation to contribute structured data and some application scenarios were also discussed.

The proposed approach addresses the specific problems pointed out in Section 2.4.4 as summarized below briefly.

- *Complexity and learning curve*. The social platform offers a simple interface enabling ordinary people to contribute structured contents. The flexible and relaxed interface enables free contribution.
- *Difficulty of concept definition and ontology creation*. Concept definitions come from the community and partial definitions are combined to form rich definitions. Relaxing constraints also keeps the definitions flexible and easy. Lightweight ontologies emerge semi-automatically as common vocabulary to structure and share data by various bottom-up processes.
- *Existence of multiple conceptualizations*. Multiple conceptualizations are maintained and, at the same time, consolidated into a common unified conceptualization.
- *Difficulty of collaboration and consensus*. Global consensus is not necessary and direct collaborative interaction is not needed. People may maintain their perpectives independently.

4. Structured Data Dissemination in Communities

As discussed in Section 2.2.1, the social web serves as a good infrastructure for dissemination of information in communities. It can serve both centralized and decentralized modes of information dissemination.

Centralized vs. decentralized approach. A centralized system or service can serve as a convenient and persistent access point. However, a single system cannot meet all the requirements of different individuals and organizations. We cannot expect all to use the same centralized system. There are numerous autonomous organizations spread worldwide having different information systems. Further, a centralized system has the risk of being a central point of failure. The web as a whole is a decentralized architecture although most of the existing systems provide their own centralized services on the web.

Decentralized structured data dissemination. In the social web, the disseminated information is usually unstructured or has limited structure. However, existing social web technologies can be extended to disseminate structured information. The Semantic Web extends the decentralized architecture of the web to publish structured information. Semantic standards also ensure interoperability which is crucial in a decentralized scenario. The semantic blogging systems serve as decentralized publishing systems. Structured data can also be embedded in the information feeds provided by existing platforms. Decentralized information sharing can also be realized over a peer-to-peer network architecture. The NEPOMUK social semantic desktop framework proposes using P2P networks for decentralized information sharing (Groza et al., 2007). Bibster (Haase et al., 2004) is a peer-to-peer application for sharing bibliographic metadata.

JeromeDL(Kruk et al., 2005) is a digital library system enhanced by semantic web technologies supporting various bibliographic standards. Each person can gather bookmarks, post comments and annotations. It introduces the notion of semantic social collaborative filtering for providing relevant recommendations. Information collections of other people can be linked and drawn into one's own collection. Collections within a friendship neighborhood can be drawn based on expertise level of the owners. JeromeDL is basically a centralized library system. Although interaction with other digital libraries in a peer-to-peer network is possible by using special protocol, it is limited to digital library systems.

Hence, a new approach for decentralized information sharing across social semantic systems was proposed. A semantic blogging system called SocioBiblog was implemented to enable sharing of bibliographic information in a decentralized social networks of researchers. A particular StYLiD installation is centralized. However, as it provides structured data following semantic standards, decentralized information sharing among multiple StYLiD systems can be achieved as demonstrated by SocioBiblog. The proposed approach consists of the following main aspects.

a) Decentralized Publishing and Aggregation

Information sharing is not only about publishing online but also providing relevant information to individuals. Currently, systems for publishing and aggregating information are isolated. However, both functions are essential for effective information sharing in a decentralized scenario. Hence, it would be useful to combine these two counter-parts for online information sharing. Both these capabilities are combined in a single SocioBiblog system as illustrated in the Figure 40. Distributed instances of such systems would be able to both push data into the web and also pull data from the web. Use of standard structured formats will make the data meaningful and facilitate information exchange between different systems.

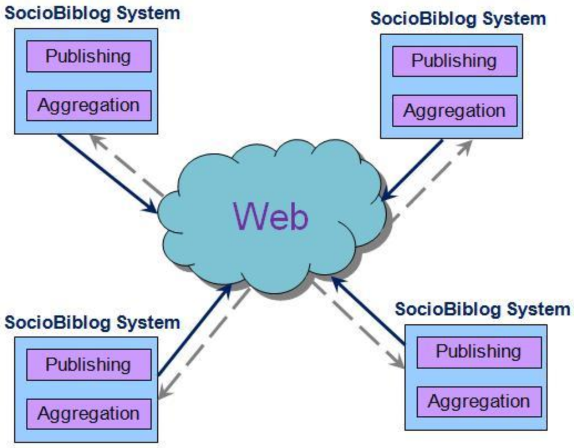


Figure 40. Decentralized publishing and aggregation with SocioBiblog.

b) Social Network based Aggregation

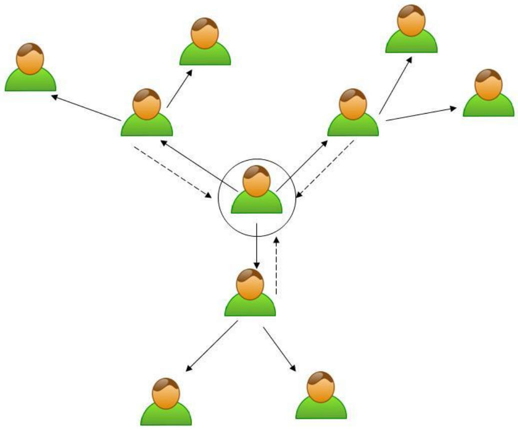


Figure 41. Aggregation of information through social links.

The approach proposes aggregating information from the social links of a person as illustrated in the Figure 41. Social network provides a powerful mechanism for connecting people and disseminating information as pointed out in Section 2.2.1. Any desired target person can be reached within a small number of links. Figure 41 shows up to the second degree of links. Using social links almost all people can be covered in the *small world* for information sharing. Moreover, we can expect to aggregate relevant resources from such social network neighborhood. This is experimentally verified next in Section 4.1. Researchers working in a common area, connected by social network links, have similar interests and are more eager to communicate and share resources. Collecting information through social links and redistributing the information facilitates flow of information in the linked community.

c) Information Source Integration and Metadata-based Filtering

We can aggregate information from multiple distributed sources and integrate these in a homogenous way. An aggregated collection can be filtered by metadata to meet our information requirements. Semantic structure provides fine grained control over information. Selection of appropriate information sources and filtering can be done to suit personal needs and a new customized information source can be constructed. For e.g., in the Figure 42, information sources A, B and C are aggregated and filtered to form a new information channel D. This information source can further be integrated with other information sources (for e.g, D is mixed with another source E and filtered). Morbidoni et al. (2008) have proposed Semantic Web pipes to remix structured data in several semantic formats. SocioBiblog can serve as such Semantic Web pipes for bibliographic information.

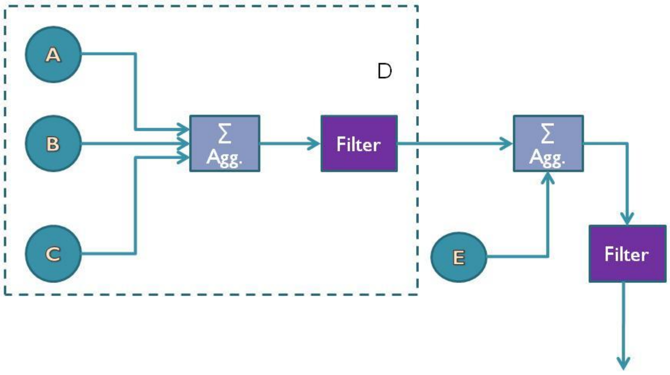


Figure 42. Integration and mixing of information feeds.

4.1 Significance of Social Information Sharing

To verify our hypothesis that relevant information can be obtained through social network links some experiments were performed on the co-authorship network of researchers. It is difficult to obtain the perfect social network of researchers. Open standards like FOAF are still not widely adopted and, moreover, the author's identity is usually not being associated with his/her publications. So rather the co-authorship network of researchers was chosen for experimentation.

4.1.1 Experimental setup

The DBLP[73] database was for experiments. The DBL-browser[74] was used to access the whole DBLP Library offline. The data file downloaded was last updated on Aug 22, 2007. The total number of publications was 928,802 with total 562,115 authors.

To measure the relevance of publications of the co-authors of a person, the similarity between his/her publications and those of the co-authors was computed. A **publication model** is created for an author by concatenating all the titles of his/her publications and removing stop-words. This model is then used for textual similarity measurements. The popular TF-IDF similarity measure implemented in the SecondString[75] package was used. The similarity with publications of the co-authors of co-authors, i.e., the second degree of social links, was also computed.

100 authors were chosen randomly such that each had some co-authors and co-authors of co-authors. For each author *X*, all the co-authors are found and for each co-author of *X* (C_X),

1. Calculate the similarity(Sim_1) between the publication models of *X* and C_X
2. For each co-author of C_X (C_{CX})
   - Calculate the similarity(Sim_2) between the publication models of *X* and C_{CX}

The average ($AvgSim_1$), maximum ($MaxSim_1$) and minimum ($MinSim_1$) of the similarities between the author and co-authors (Sim_1) were calculated for each author. Similarly, the average ($AvgSim_2$), maximum ($MaxSim_2$) and minimum ($MinSim_2$) of Sim_2 were also calculated for each author.

To evaluate these similarity measures a **baseline** is setup. The relevance of our results based on co-author links was compared with the results of traditional keyword search. For the same 100 randomly selected authors, the following process was followed to construct the baseline.

1. From the publication model of author *X*, take *N* distinct words (at most) with the highest TF-IDF scores. It is considered that this set models the interest of the author *X*.
2. Search publications relevant to *X* from the entire collection as follows.
   a. Make all possible bi-gram combinations of the keywords.
   b. Search publications containing each bi-gram in the title.
   c. Return the union of all bi-gram searches.

[73] http://www.informatik.uni-trier.de/~ley/db/
[74] http://dbis.uni-trier.de/DBL-Browser/
[75] http://secondstring.sourceforge.net/

3. Create the “keyword-search model” by concatenating all the search result titles and removing stop-words.
4. Calculate the similarity (*Sim_0*) between the publication model of *X* and the keyword-search model.

4.1.2 Observations

Table 3. Statistics about randomly chosen authors.

	Average	Maximum	Minimum	Standard deviation
n_{c1}	7.7	69	1	11.44729
n_{c2}	154.44	2004	1	302.9296
$AvgSim_1$	0.469683	0.965455	0.160604	0.21159
$MaxSim_1$	0.650913	1	0.160604	0.264625
$MinSim_1$	0.297384	0.905487	0.053999	0.191167
$AvgSim_2$	0.055547	0.380373	0	0.063572
$MaxSim_2$	0.201005	0.766154	0	0.151642
$MinSim_2$	0.006634	0.224139	0	0.033984
Sim_0	0.411849	1	0	0.22834

n_{c1}: Number of co-authors (C_X) , n_{c2}: Number of co-authors’ co-authors (C_{CX})

Table 3 shows some observed statistics about the randomly chosen authors. It is observed that $AvgSim_2$ is usually much less than $AvgSim_1$. This indicates that the relevance of publications diminishes rapidly as the degree of social link increases. However, $MaxSim_2$ is relatively high and, in fact, in some cases even higher than $MaxSim_1$. This shows that even at the second degree of links, there may be some highly relevant publications though the average relevance is low. The low standard deviations for the similarity measures indicate consistency of the results. The baseline keyword similarity (Sim_0) shown in Table 3 was obtained using $N = 5$ in the above procedure. The average co-author similarity ($AvgSim_1$) seems to be better than the keyword similarity.

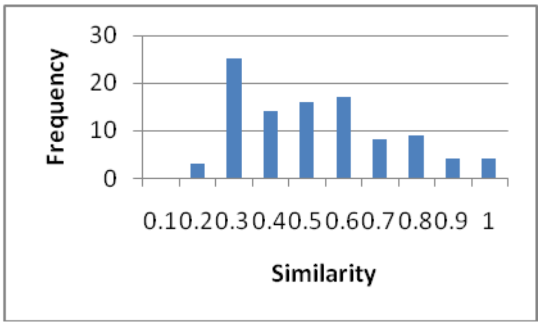


Figure 43. Average co-author similarity ($AvgSim_1$).

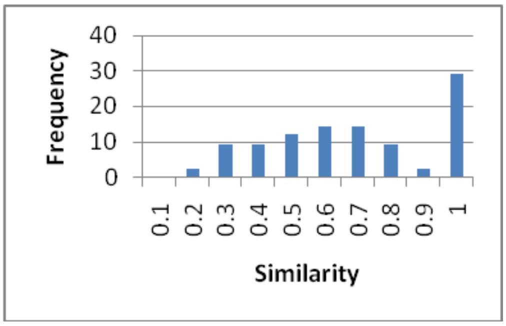


Figure 44. Max. co-author similarity ($MaxSim_1$).

The histograms illustrate some statistics about the co-author similarity measures. Figure 43 shows that most people have the average co-author similarity ($AvgSim_1$) between 0.3 to 0.6. Figure 44 shows that the maximum co-author similarity

(*MaxSim$_1$*) peaks around 0.6 to 0.7. The peak at 1 simply indicates that most people have some co-authors who always write together.

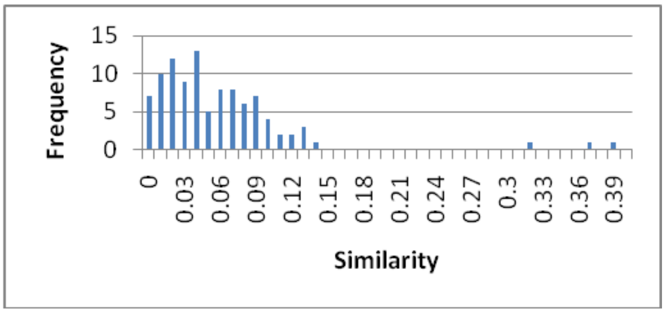


Figure 45. Average co-authors' co-author similarity (*AvgSim$_2$*).

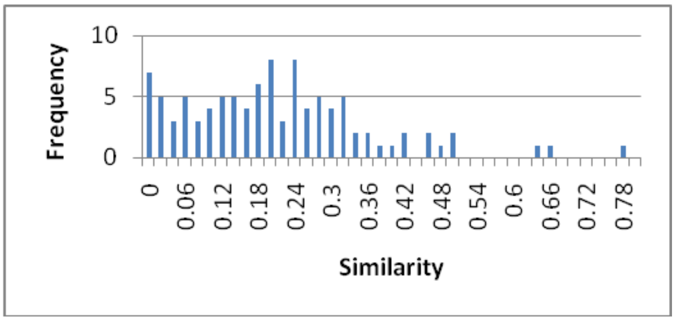


Figure 46. Maximum co-authors' co-author similarity (*MaxSim$_2$*).

Figure 45 indicates that in most cases, *AvgSim$_2$* is around 0.02 to 0.04. *MaxSim$_2$* peaks around 0.2 to 0.24 and 0.02 (Figure 46). However, the maximum similarity goes even as high as 0.66 to 0.78 in few cases. Thus, although the relevance significantly diminishes in the second level of co-authors, some relevant publications can still be obtained.

These co-author based similarities were evaluated by comparing to a keyword search results baseline setup using top 5 keywords (*N*=5 in the above procedure).

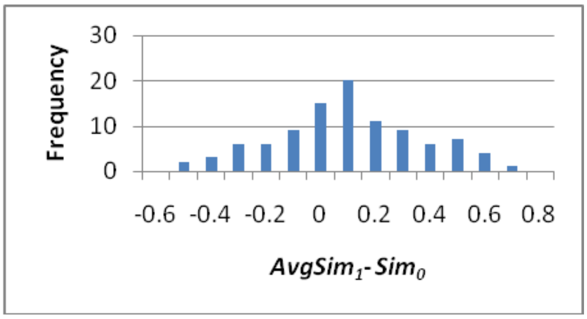


Figure 47. Difference between co-author similarity and keyword similarity (*AvgSim$_1$- Sim$_0$*).

Figure 47 shows that the difference between the average co-author similarity and baseline keyword similarity (*AvgSim$_1$-Sim$_0$*) peaks around 0 to 0.1 indicating that the co-author similarity works similar to or slightly better than the keyword similarity in most cases. *AvgSim$_1$*was greater than *Sim$_0$* for 59 out of total 100 authors.

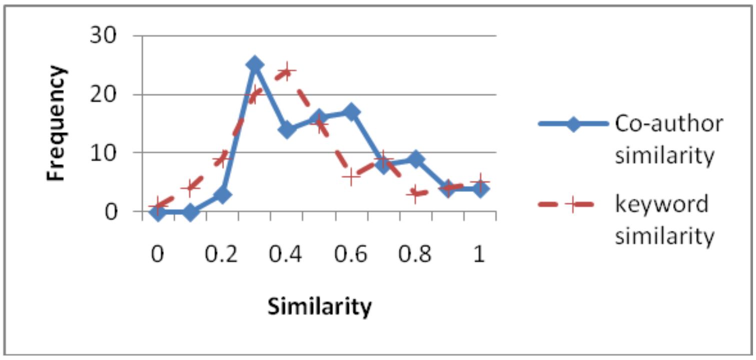


Figure 48. Comparison of co-author similarity(*$AvgSim_1$*) and keyword search baseline(*Sim_0*) (*N* = 5)

Figure 48 compares the histograms of *$AvgSim_1$* (shown in Figure 43) with the baseline keyword similarity (*Sim_0*). *Sim_0* peaks around 0.3 to 0.4 but decreases rapidly towards higher similarities. *$AvgSim_1$* peaks around 0.3 which is slightly behind the peak of the baseline. However, the *$AvgSim_1$* remains greater in frequency than the baseline for higher similarities. This indicates that the co-author similarity is comparable to or slightly better than the baseline keyword-based results.

When *N* = 10 (Figure 49) the baseline keyword search results were even worse. *Sim_0* peaks at quite a low value of 0.2 and falls rapidly. *$AvgSim_1 > Sim_0$* for 78 out of 100 authors. Hence, co-author similarity is much better than the keyword similarity with too many keywords.

These experiments verify that relevant publications can be obtained from one's co-authors. The co-author based results are comparable to or even better than the keyword search results from the entire database of publications. We need not go far in the social network to find relevant publications. Even collecting publications of just the co-authors can yield good results. In fact, similarity diminishes rapidly as the degree of social links increases.

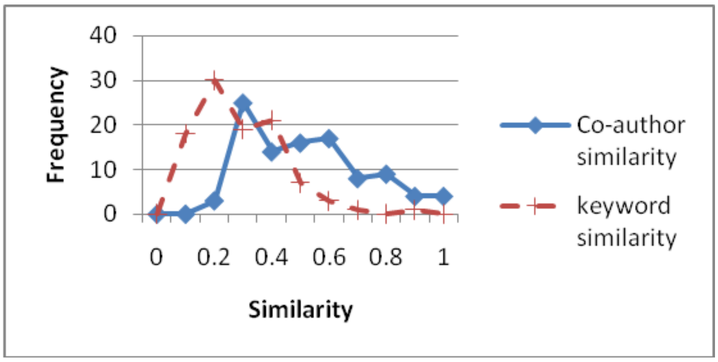


Figure 49. Comparison of co-author similarity(*$AvgSim_1$*) and keyword similarity (*Sim_0*) (*N* = 10)

4.2 Use Case Scenario

Figure 50 illustrates an example scenario. A researcher, ‘A’, publishes information about his publications on his semantic blog. He may enter metadata about his publication. 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 publications from other sites and comment on them. The BibTeX metadata would be scraped from the original site and quoted in the blog entry.

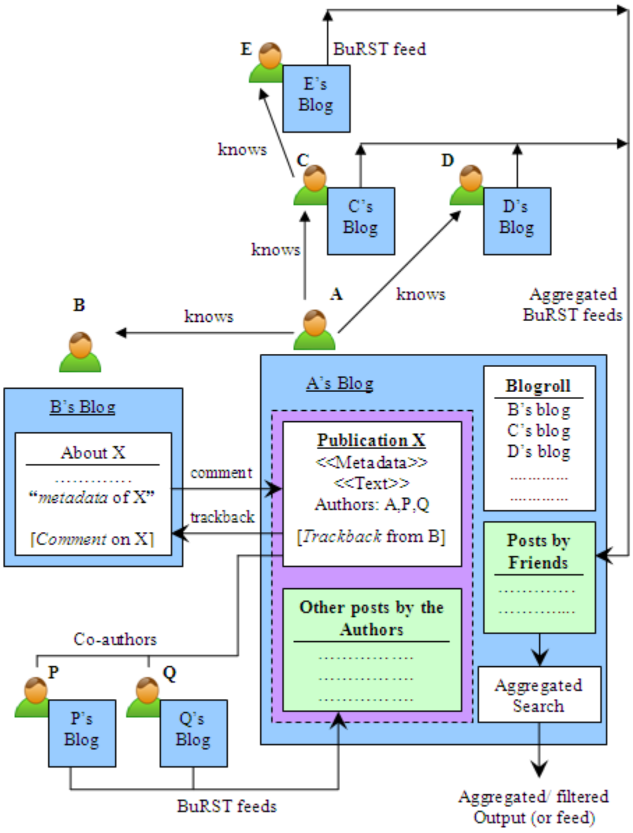


Figure 50. Example scenario for SocioBiblog.

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.

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 with a specific keyword. The aggregated and filtered output thus obtained is again exposed as a new BuRST feed. The user may subscribe to this feed and get desired notifications. The feed can further be integrated with other information sources. For instance, a user may integrate the feed with articles from other related journals.

4.3 *Implementation of SocioBiblog*

4.3.1 System architecture

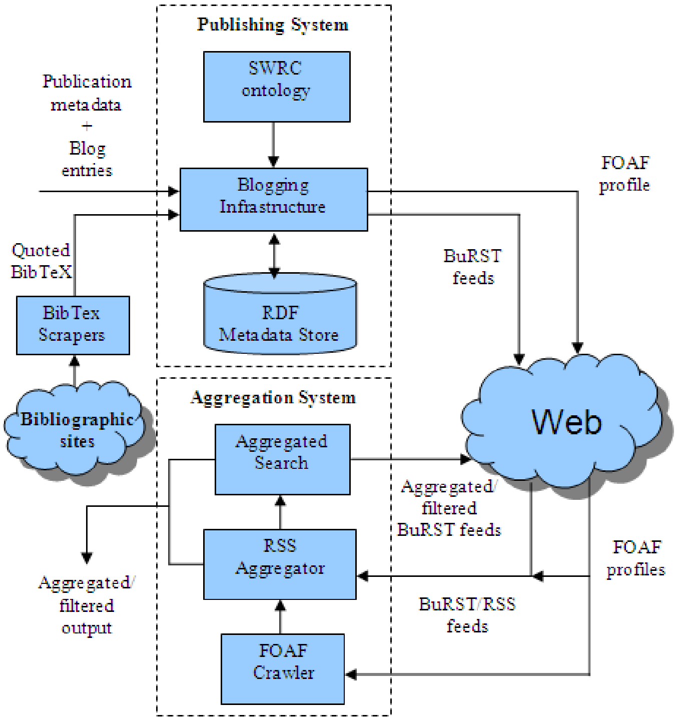


Figure 51. System architecture of SocioBiblog.

Figure 51 shows the architecture of the system. It consists of two sub-systems.

The *publishing system* facilitates publishing blog entries and metadata about publications. It is built over an existing blogging infrastructure. Structured blog entries contain metadata based on the SWRC ontology (Sure et al., 2005). 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 system also publishes the FOAF profile of the blog-owner.

The *aggregation system* utilizes RSS technology to aggregate publications from multiple sources. RSS/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. The aggregated posts are output on the blog. Aggregated search helps in filtering aggregated publications by defining required metadata criteria. The aggregated and filtered BuRST feed thus obtained can be exported again as a new BuRST feed.

Figure 52 illustrates how the system can co-exist and interoperate with existing systems. Existing blogging and publishing systems can usually generate RSS feeds (some systems can generate BuRST feeds too). Aggregation systems exist separately. Our design integrates both the 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.

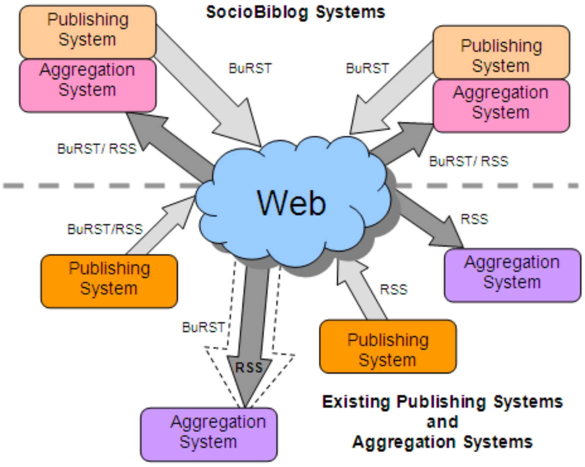
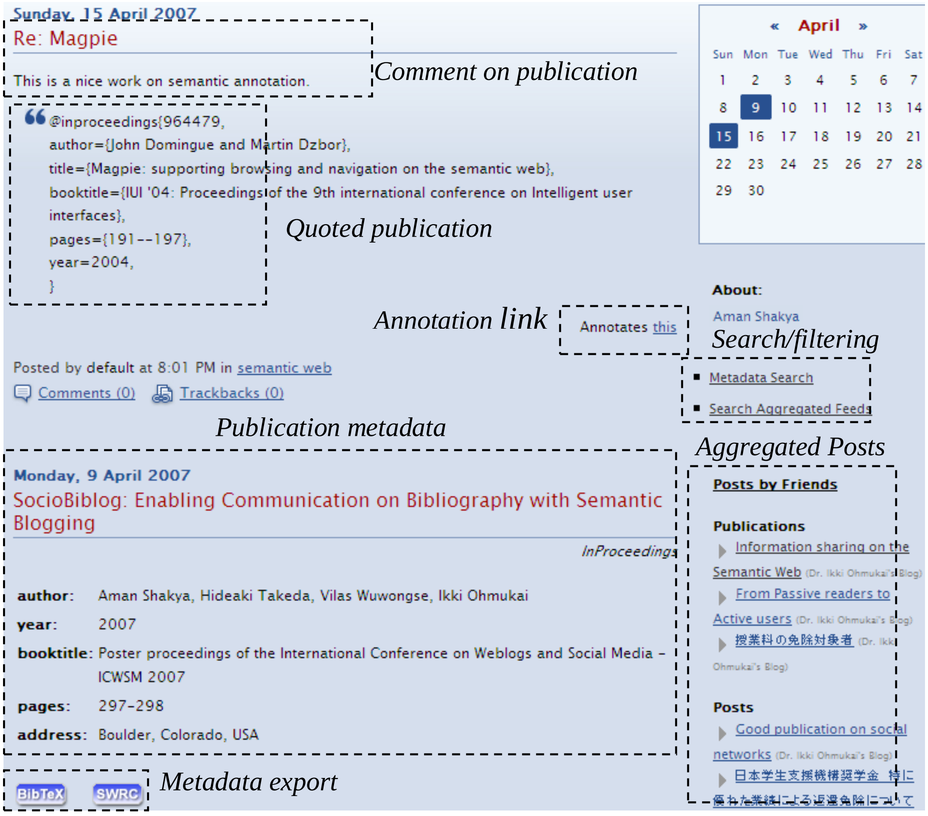


Figure 52. Publishing and aggregation on the current web with SocioBiblog.

4.3.2 Publishing



Sunday, 15 April 2007

Re: Magpie

This is a nice work on semantic annotation.

@inproceedings{964479,
author={John Domingue and Martin Dzbor},
title={Magpie: supporting browsing and navigation on the semantic web},
booktitle={IUI '04: Proceedings of the 9th international conference on Intelligent user interfaces},
pages={191--197},
year=2004,
}

Annotates this

Posted by default at 8:01 PM in semantic web

Comments (0) Trackbacks (0)

Monday, 9 April 2007

SocioBiblog: Enabling Communication on Bibliography with Semantic Blogging

InProceedings

author:	Aman Shakya, Hideaki Takeda, Vilas Wuwongse, Ikki Ohmukai
year:	2007
booktitle:	Poster proceedings of the International Conference on Weblogs and Social Media - ICWSM 2007
pages:	297-298
address:	Boulder, Colorado, USA

BibTeX SWRC

« April »

Sun	Mon	Tue	Wed	Thu	Fri	Sat
1	2	3	4	5	6	7
8	9	10	11	12	13	14
15	16	17	18	19	20	21
22	23	24	25	26	27	28
29	30					

About:

Aman Shakya

- Metadata Search
- Search Aggregated Feeds

Posts by Friends

Publications

- Information sharing on the Semantic Web (Dr. Ikki Ohmukai's Blog)
- From Passive readers to Active users (Dr. Ikki Ohmukai's Blog)
- 授業料の免除対象者 (Dr. Ikki Ohmukai's Blog)

Posts

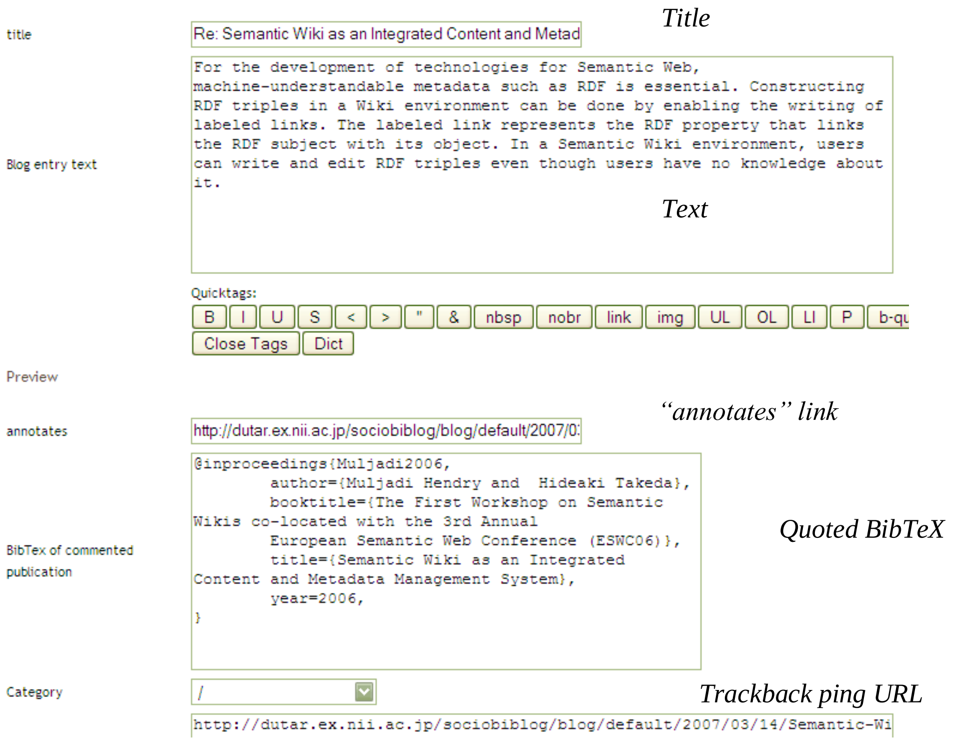
- Good publication on social networks (Dr. Ikki Ohmukai's Blog)
- 日本学生支援機構奨学金 特に優れた業績による返還免除について

Figure 53. SocioBiblog interface.

Publishing of blog entry and metadata

Figure 53 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. We can quote a publication and comment on it. Publication metadata entry is exported in SWRC, BibTeX formats and BuRST feeds. Blojsom[76] 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[77].



title: Re: Semantic Wiki as an Integrated Content and Metad

Blog entry text:

```
For the development of technologies for Semantic Web,
machine-understandable metadata such as RDF is essential. Constructing
RDF triples in a Wiki environment can be done by enabling the writing of
labeled links. The labeled link represents the RDF property that links
the RDF subject with its object. In a Semantic Wiki environment, users
can write and edit RDF triples even though users have no knowledge about
it.
```

annotates: http://dutar.ex.nii.ac.jp/sociobiblog/blog/default/2007/0:

BibTex of commented publication:

```
@inproceedings{Muljadi2006,
        author={Muljadi Hendry and  Hideaki Takeda},
        booktitle={The First Workshop on Semantic
Wikis co-located with the 3rd Annual
        European Semantic Web Conference (ESWC06)},
        title={Semantic Wiki as an Integrated
Content and Metadata Management System},
        year=2006,
}
```

Category: /

```
http://dutar.ex.nii.ac.jp/sociobiblog/blog/default/2007/03/14/Semantic-Wi
```

Figure 54. Blog this interface.

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 that will be discussed in later. 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[78] are used to define relations with people in the blogroll which are mapped into FOAF one-to-one.

Commenting mechanism

“Blog this” bookmarklet. Commenting has been made convenient by providing a 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 Figure 54. The “annotates” link is manifested as

[76] http://blojsom.sourceforge.net
[77] http://www.hpl.hp.com/semweb/jena.htm
[78] http://gmpg.org/xfn/

shown in Figure 53. The link is also added in the BuRST feed to distinguish between quoted entries and the original publication.

BibTeX Scraping. When a publication is bookmarked using *Blog this*, the system tries to scrape out BibTeX information if available. SocioBiblog currently provides scrapers for SocioBiblog instances, the ACM digital library[79] and a generic BibTeX scraper which works for any web page that contains BibTeX snippet (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.

4.3.3 Aggregation

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 Figure 53. Publications and non-publications are separated while aggregating the posts. When a blog entry for a publication is opened, BuRST/RSS feeds of the co-authors of the publication are downloaded and shown alongside. The feed URLs are determined from FOAF profiles of the co-authors. The Flock RSS aggregator[80] has been used for RSS aggregation and extended to process BuRST.

Social Network based aggregation

SocioBiblog aggregates feeds from directly linked friends and also friends of the friends. The aggregator first subscribes feeds from sources listed in the blogroll. The BuRST/RSS feed URL of a friend is obtained from the blogroll or that person's FOAF profile. Then, the system goes one level deeper into the FOAF network to find friends of the friends and adds their feeds to the subscription list as well. The second level of linked friends is traced whenever the blogroll is updated. Discovery of linked sources and aggregation are done in background without affecting responsiveness of the system.

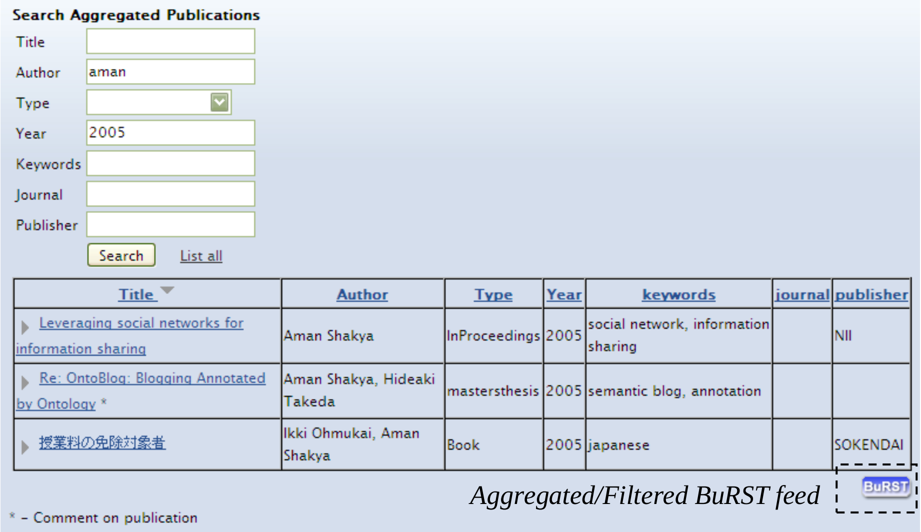
FOAF Crawler. The Elmo scutter[81] has been adapted as a FOAF crawler to find out FOAF links of authors. Elmo provides the interface and options to manage crawling. The crawler traces *rdfs:seeAlso* elements for FOAF links and gathers FOAF profiles in a database. Users may also enter FOAF links of authors while posting publications. Users may search the crawled FOAF database for FOAF links.

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 Figure 55. The system also

[79] http://portal.acm.org/
[80] http://flock.sourceforge.net
[81] http://www.openrdf.org/doc/elmo/users/index.html

searches publication metadata quoted in blog entries. The user can specify values for different metadata fields using the online form. Only the entries satisfying the specified metadata criteria are filtered and included in the result. The metadata filtering parameters can also be directly specified in the request URL. The results can be sorted by clicking on the field headers.



Search Aggregated Publications

Title	
Author	aman
Type	
Year	2005
Keywords	
Journal	
Publisher	

Title	Author	Type	Year	keywords	journal	publisher
Leveraging social networks for information sharing	Aman Shakya	InProceedings	2005	social network, information sharing		NII
Re: OntoBlog: Blogging Annotated by Ontology *	Aman Shakya, Hideaki Takeda	mastersthesis	2005	semantic blog, annotation		
授業料の免除対象者	Ikki Ohmukai, Aman Shakya	Book	2005	japanese		SOKENDAI

* – Comment on publication

Aggregated/Filtered BuRST feed

Figure 55. Searching aggregated publications.

The aggregated and filtered results obtained are again exported as a new BuRST feed. The user can subscribe to this BuRST feed to 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, we can integrate various distributed information sources, filter them and construct new information sources.

4.4 Summary and Lessons Learned

Proper dissemination in communities is important for proper utilization of published information. The web is a highly distributed environment with many independent systems. It is important to provide a decentralized mechanism for information dissemination across such independent systems. Although people may use different online systems they are often well connected by social links which can provide the path for information dissemination. Experimental evidences support the fact that a lot of relevant information can be obtained through such social links. Therefore, an approach for decentralized information sharing through social networks was proposed. The SocioBiblog system was implemented to demonstrate the approach. The system facilitates the flow of bibliographic information in communities by providing both publishing and aggregation capability. RSS aggregation can easily be extended to support structured data. Information feeds can also be integrated, filtered and mixed to obtain desired streams of information. Actually, the approach can easily be adopted by any online system which supports RSS. Disparate systems can interoperate based on common semantic standards for the transported information.

However, some practical difficulties were faced for proper deployment of SocioBiblog. Some of the lessons learnt are as follows.

- Deployment was difficult because extending the blogging platform requires server-side installation. Usually, people do not have the access to the hosting servers and cannot install extensions to the system. It is more convenient for the users to access a centralized online service without requiring any installation or configuration. Nonetheless, if extensions are provided by the public blog providers, many users would benefit without any effort.

- The implementation is blogging platform dependent. Therefore, it is difficult to deploy for people using different blogging systems. People prefer to continue using their existing blogs and social applications rather than using a new system. However, it is not feasible to deploy extensions for every platform. In future, to have a large user base, such extensions may better be provided for popular social platforms like Facebook[82], which already provides an extension API.

- SocioBiblog uses FOAF as an open social networking standard. However, FOAF, although being one of the most popular ontologies, is still not widespread. Most online social networking services are closed and do not allow social links to users of other services. Hence, real online social networks are still disconnected islands. Information can still be transported across these islands using RSS mechanisms but the social links are not in place across systems.

- Finally, the current implementation is for bibliographic information only. However, people want to share a wide variety of data. Although some social sites like LivingSocial[83] provide a number of data types it is not possible to cover all types of data different people are interested in. A system like StYLiD provides a generic solution.

[82] http://www.facebook.com/

[83] http://livingsocial.com/

5. Evaluation

5.1 Evaluation Scheme

The proposed approach consists of a number of aspects including providing a data authoring interface usable for ordinary users, collecting user-defined concepts, consolidating them, grouping and organizing them and enabling emergence of lightweight ontologies. It is not straightforward to directly evaluate the outputs of the approach like the user-generated concepts, structured Semantic Web data, consolidated concepts and the informal ontologies. Nevertheless, the following aspects can be considered for the evaluation of the approach.

1. *Evaluation of usability*. It should be tested whether people from any background can start using the system without any training and how effectively they can use the system considering the various features. As users are an integral part of the approach, it is important to evaluate this aspect. However, ideally this process may be iterative and require a long period of time as in (Pfisterer et al., 2008).
2. *Evaluation of user defined concepts*. It should be checked whether people can define concepts in terms of schemas. The nature of such user-defined concepts from different people should also be studied.
3. *Evaluation of the consolidated concepts*. The applicability of concept consolidation and the validity of the method used should be verified.
4. *Evaluation of the emerging ontology*. Ontology evaluation is a difficult, indirect and imperfect area (Brank et al., 2005). The ontologies, in our case, emerge as informal vocabulary for information sharing from user-defined concepts. So methods available for evaluating formal well-engineered ontologies (Sure et al., 2004; Guarino & Welty, 2000; Gómez-Pérez, 2003) are not applicable. As an indirect way, we can evaluate the processes that enable emergence of the ontology, mainly the consolidation and grouping of concepts. It can be argued that if the methods work correctly and the input is valid then the output should be as expected.

To evaluate the aspects listed above the following methods have been used.

1. *Experiment on usability*. To evaluate the usability of the system it was tested with a number of people by designing several experimental tasks. The same tasks were also performed with an existing system to provide an evaluation baseline.
2. *Experiment on conceptualization*. This experiment was designed to observe how different people can define concepts, by assigning them some conceptualization tasks. The applicability of concept consolidation on these user-defined concepts was also tested.
3. *Experiment on existing data*. Some experiments were also performed on an existing dataset of user-defined concepts. This helped in making observations about user-defined concepts and testing the methods for concept consolidation and grouping.

4. *Practical applications*. Finally, the system has also been used for some real-world applications providing evidence of the applicability of the overall approach.

These experiments and means of evaluation are described in the following sections.

5.2 Experiment on Usability

Experiments were performed with some invited users. The purpose of these experiments was to check the usability of various features of the system and to study the user behavior regarding the various aspects of the system. Specific tasks were designed to cover various capabilities of the system and the users were asked to perform these tasks. For evaluation purpose, users were asked to perform the same tasks using the Freebase system too, as a baseline for comparison.

The basic hypothesis of the experiment is that "StYLiD is more usable than Freebase for the given tasks". Freebase was chosen for comparison considering the following reasons.

- Freebase is functionally more similar to StYLiD than any other system we were aware of. Like StYLiD, Freebase also allows users to define their own schemas and input structured data instances, data instances can be interlinked, etc.
- Freebase, with interactive interfaces, seems to be easier to use compared to other systems and does not seem to require technical knowledge or special training. Alternately, the semantic wikis are more difficult to use and need some training.
- Freebase is also meant for public use like StYLiD. It is available as an online service for free. No installation is required to use it.

5.2.1 Experimental task design

The tasks for the experiment were mainly designed keeping the features of StYLiD in mind. Few tasks or some specific instructions are not directly applicable for Freebase. So the tasks were modified or omitted accordingly for Freebase.

Task 1 (Structured data authoring)

In this task, the user was asked to input a given structured data instance for a concept that already exists in the system. As given in the Appendix A (Task 1), the user was asked to input data about "The Beatles" as an instance of the "band" concept. To test the various capabilities of the system related to entering instance data, the example instance was designed so as to include all these features an instance can have. The following features were included (as shown in the Appendix).

Linking to internal instances. The members of the band were to be picked up from the singers already in the system. These were instances of the "singer" concept (though this was not mentioned explicitly). Picking up the members would link the band instance to the singer instances in the system. All the given singers were already in the system.

Entering multiple values. Multiple attribute values had to be entered for 3 attributes – members, films and past_members.

Entering Wikipedia URI. The origin of the band, “Liverpool, England” was to be linked to the Wikipedia page of Liverpool.

Entering arbitrary URI. A given URI (http://dbpedia.org/resource/Brian_Epstein) was to be entered for the manager of the band “Brian Epstein”.

Picking up from enumerated values. Value for the genre attribute could be picked up from a drop-down of enumerated possible range values.

Some features were omitted for the same task for Freebase, as seen in the Appendix. Entering the URIs, from Wikipedia or arbitrary, were omitted. In Freebase, there is no direct way to link attribute values to external resources.

Task 2 (Structured concept schema creation)

In this task, the user had to enter a given concept into the system. The concept did not already exist in the system. As shown in the Appendix A (Task 2), the “Concert” concept had to be entered with given schema. A list of attributes along with descriptions had to be entered. The following features for concept schema were tested.

Specifying or suggesting the range concepts for values. The user had to specify that the performer attribute of the “Concert” may be an instance of “band” concept (which already exists in the system and used in task 1). In Freebase, the type of an attribute may be specified as a single existing “type”. In StYLiD, multiple concepts may also be suggested as possible range. This capability was also tested. The user had to suggest that the “organizer” attribute of the “concert” may be an instance of “organization” or “band” concepts. (Both the “organization” and “band” concepts already existed in the system). However, this is not directly possible in Freebase. So this requirement was dropped for the same task for Freebase.

Enumerating range of literal values. For the “type” attribute, some possible values (rock, classical, jazz, pop) had to be enumerated. This is also not directly possible in Freebase and so was omitted from the task for Freebase.

Task 3 (Modifying and reusing an existing concept)

This task was used to test the capability of StYLiD that allows users to reuse an existing concept and modify it to create a new concept. However, this is not possible in Freebase and so was omitted for Freebase.

The user was asked to enter a “singer” concept with given attributes, as shown in the Appendix A (Task 3). The users were not informed directly that the “singer” concept already existed in the system (though cautious participants would have noted this from task 1 which used the “singer” concept).

The following features were again tested with this task.

Specifying or suggesting the range concepts for values. The user had to specify that the “member-of” attribute of the singer may be instance of “band” or “organization” concepts as possible range. Similarly, the “live-performances” may be instance of the “concert” concept (which was defined by the user in Task 2).

Enumerating range of literal values. The user had to enumerate some given values for the “genre” attribute (rock, pop, classical, jazz, country).

Most of the schema definition was already present in the existing “singer” concept which was as follows.

- name [description: name of the singer]
- nationality
- genre [enumerated range: rock; pop; classical]
- member-of [range concept: band]
- years

Hence, the user might easily adapt the existing concept by simply modifying and adding some attributes.

Task 4 (Updating one’s own concept)

This was a short task in which the user simply had to modify the “singer” concept defined in Task 3 to add two new attributes. When a user tries to modify his/her concept, the system offers two possibilities, either to “modify the existing concept” or to “create a new version of the concept”. The task was mainly intended to check the user response to these options. This check is not applicable for Freebase as a new version of an existing concept cannot be created. So this task was also omitted for Freebase.

Task 5 (Structured concepts and instances authoring)

In this task, the user had to input a given structured data instance of an “album”. The “album” concept did not exist in the system. However, the users were not explicitly told so that they would need to figure out by themselves that the concept should be created first. The instructions for this task were exactly the same for both the systems (as shown in Appendix A, Task 5).

Task 6 (Searching)

In this final task, the user had to search all the movies directed by “Martin Scorsese” which had “Leonardo DiCaprio” in the starcast. The concept and attribute labels were slightly different in StYLiD and Freebase. So the instructions were worded to match these labels in the respective systems, as shown in the Appendix A (Task 6).

5.2.2 Experimental setup

Participants were invited by sending an email to everyone. It was stated that absolutely no prerequisite and no training or learning would be needed to participate in the experiment. People from any background were encouraged to participate. To prevent the participants from trying the system before the experiment, the names of the systems and nothing about the systems were mentioned. It was also stated in the invitation that it was preferable that the participants do not know about the systems or use them beforehand.

To motivate people to participate in the experiment, participants were provided with appropriate awards. Further, the participants were also assured of privacy and anonymity through a privacy policy and signed agreement. The experiment session was about 1.5 to 2 hrs long. Breaks were also allowed if needed, not to pressure the users with all the tasks. The experiment was entirely conducted in English.

A separate installation of StYLiD was made on a different server for the experiment[84]. The data from *http://www.stylid.org* was imported to this installation to populate the system with real data. The concepts and instances needed for the experimental tasks were also entered.

For each participant, a separate user account was created and the user was signed in before starting the experiment. For Freebase, a new base called "Experiment" was created for the experimental tasks. A single user account was used for all the participants to avoid many dummy users in the real online system.

Each participant was asked to fill some details about themselves in the form shown in the Appendix B (Participant Details). The systems were briefly explained. Then, the tasks were assigned sequentially one at a time. The participants were asked to do the tasks on their own as far as possible. However, if they got stuck or started going totally wrong, they were hinted or interrupted accordingly to keep them in track and to run the experiment smoothly. To make them feel comfortable, the participants were not watched constantly. However, it was intermittently checked whether they were stuck or going wrong way. Before starting an experiment session, the data from previous sessions were deleted and the systems were reset to the original state.

The participants were told that we were working on both the systems to prevent any bias in their response. They were not told that we were working on StYLiD only and not Freebase until the end of the experiment. Furthermore, it is possible that after performing a task on one system, the user would learn from this and the experience would affect the use of the next system for the same task. To avoid this effect in the experiment, the order of StYLiD and Freebase was switched for the tasks alternately for each new participant. Hence, any such learning effect would be canceled out in the overall evaluation.

After finishing the experiment each participant was also asked to fill the final questionnaire shown in the Appendix B (Final Questionnaire). This form was used to learn few things about the participant's background. It was kept at last so that the participant does not feel intimidated before the experiment. The participants were also asked whether they had any final overall comments or suggestions.

5.2.3 Means of observations

The following means were used to make observations for each task of the experiment.

Questionnaires. After each task, the participant was asked to fill the Task-specific Questionnaire, shown in the Appendix B. The 5 possible responses on the scale were

[84] http://sicily.ex.nii.ac.jp/stylid/

later assigned scores from 0 to 4, 0 for the worst case and 4 for the best. Thus, for the first question about confidence the scores are 0 to 4 from left to right (very low to very high confidence). For the second question about ease, the scores are 4 to 0 from left to right (very easy to very difficult). This questionnaire also collects notes on difficulties, comments and suggestions from the users regarding the task.

Then, the participant was also asked to fill the Task-specific Comparative Questionnaire (shown in Appendix B), comparing StYLiD and Freebase for the task, only if the task was applicable for both the systems (tasks 1, 2, 5 and 6).

Screen video logs. While the user performed each task, the computer screen was captured as a video using a screen capturing software. The users were informed before the experiment that the screen videos would be captured. However, the capturing worked in the background causing no disturbance to the user actions.

User behaviour and comments. Notes were also taken about the behavior exhibited by the users. Any comment or question made by the users was also noted.

System Usability Scale. After finishing all the tasks, the participant was asked to fill the System Usability Scale (SUS), shown in the Appendix B, for both systems. The SUS (Brooke, 1996) is a likert scale designed by the Digital Equipment Corporation. It provides a broad measure which can be used for global assessments of systems usability applicable across a range of contexts. It serves as a “quick and dirty” method for low cost assessments of usability in industrial systems evaluation. The SUS scale has also been used by Pfisterer et al. (2008) to evaluate the Semantic Mediawiki.

Action measures. The number of interruptions required for providing assistance or hints was noted. It was also noted why each interruption was required. The screen videos logs were observed later after the experiments to make notes about the user actions in detail. For each task, the number of errors made by user was noted. Only the mistakes retained till the end of the task were considered as errors. A mistake was not considered as an error if the users corrected it later either by realizing by himself or by the system constraints. The time roughly taken for each task was also noted from the video.

5.2.4 About the participants

There were total 15 participants. They were from 10 different countries, 8 male and 7 female. The age range was from 22 to 43 years, average 28.3 years. There were 4 PhD students, 4 master’s students, 4 internship students, 2 post doctoral researchers and 1 hotel staff. The participants were from various fields of studies - 1 from experimental psychology, 2 from international relations studies, 2 from public policy studies, 1 from hotel services, 1 from telecommunication, 2 from computer networks, 1 from HCI, 1 from AI, 1 from multimedia information retrieval, 1 from computer graphics and 2 from web technologies. Hence, 6 participants were totally from non-IT background. The participants had a wide range of interests.

Most of them knew about Wikipedia except one. Only 3 people knew about StYLiD before the experiment. Only 1 person knew about Freebase. 8 had heard about the Semantic Web and only 6 out of them knew what it really is. 9 of them had done some database design. Almost all participants said that they do not use help when using a new online system. Only 2 said yes and one said sometimes when needed.

8 of the participants used StYLiD first and then Freebase for each task. The remaining 7 used Freebase first and then StYLiD.

5.2.5 Results

a) SUS scores

As shown in Table 4, the average SUS score for StYLiD is 69.67% (ranging from 7.5 to 95) and that for Freebase is only 39.33% (ranging from 17.5 to 80). There was no notable effect on the results either if StYLiD or Freebase was used first by the participant. It may be noted that the participant number 10 to 15 were totally from non-IT background. It is seen the background of the participants does not make any big difference in the usability scores for the systems. It seems that the response varies at individual level only and not much by the background of the user. Anyway most of the users distinctly ranked StYLiD over Freebase.

Table 4. Total SUS scores given by participants.

Participant	SUS score for StYLiD(%)	SUS score for Freebase (%)	System used first
1	42.5	30	Stylid
2	80	52.5	Stylid
3	7.5	17.5	Freebase
4	90	35	Stylid
5	75	27.5	Freebase
6	87.5	32.5	Stylid
7	87.5	17.5	Freebase
8	90	35	Stylid
9	55	40	Freebase
10	65	20	Freebase
11	57.5	40	Stylid
12	92.5	25	Freebase
13	70	77.5	Stylid
14	95	80	Freebase
15	50	60	Stylid
Average	69.67	39.33	

The average score for each question in the SUS questionnaire, normalized from 0 to 4, for both systems is shown in Table 5 (0 is the worst and 4 is the best case). We can see that StYLiD has a better score than Freebase for each of these questions. The questions can be seen in the Appendix B (System Usability Scale).

Table 5. SUS question scores.

SUS question	StYLiD average score	Freebase average score
1	2.47	1.27
2	2.8	1.47
3	2.73	1.27
4	2.93	2
5	2.33	1.8
6	2.87	2.07
7	2.73	1.07
8	2.73	1.67
9	3	1.33
10	3.27	1.8

b) Results from each task

The results from each task are described in the following text. The detailed results, including the scores from each participant, are given in the Appendix D. Table 6 shows the average scores for each task for both the systems.

Table 6. Average evaluation scores for all the tasks.

Task	Confidence		Ease		Time (in mins)		Errors		Assistance	
	S	F	S	F	S	F	S	F	S	F
1	2.8	1.87	3.07	1.73	10.23	12.07	0.93	0.47	1.53	3.93
2	2.93	2.27	3.13	2.33	7.43	11.4	1.13	0.4	1.13	2.33
3	3.2		3.33		5.33		1		0.33	
4	3.53		3.73		2		0		0.13	
5	3.2	2.33	3.4	2.07	5.87	10.4	0.87	1	0.53	2
6	2.8	1.53	2.4	1.47	4.93	7.07	0	0	2	2.4

Note: S stands for StYLiD and F stands for Freebase

Task 1 (Structured data authoring)

The average confidence score for StYLiD (2.8) was higher than that for Freebase (1.87). The average score for ease of use was also higher for StYLiD – 3.07 for StYLiD and 1.73 for Freebase. The average time required for this task was somewhat lower for StYLiD (10.23 mins.) than for Freebase (12.07 mins.). However, the average number of errors was slightly higher for StYLiD (0.93) compared to Freebase (0.47). This may be due to the relaxed and tolerant interface of StYLiD unlike the strict and constrained interface of Freebase allowing only perfect data. The number of assistance required for Freebase (3.93) was much higher than that for StYLiD (1.53).

10 people felt more confident with StYLiD than Freebase, 4 marked Freebase and 1 felt almost the same with both the systems. 13 people felt that StYLiD was easier to use and only 2 said that Freebase was easier for this task.

Task 2 (Structured concept schema creation)

The average confidence score for StYLiD (2.93) was higher than that for Freebase (2.27). The average score for ease of use was also higher for StYLiD (3.13) than for Freebase (2.33). The average time required for this task was lower for StYLiD (7.43 mins.) than for Freebase (11.4 mins.). However, the average number of errors was again slightly higher for StYLiD (1.13) compared to Freebase (0.4). The number of assistance required for Freebase (2.33) was higher than that for StYLiD (1.13).

11 people felt more confident with StYLiD, 3 marked Freebase and 1 felt almost the same. 12 people felt that StYLiD was easier to use and only 3 said that Freebase was easier for this task.

Task 3 (Modifying and reusing a concept)

This task was only for StYLiD. The average scores for confidence and ease of use were quite high, 3.2 and 3.33 respectively. The task took 5.33 minutes in average with 1 error in average and little assistance required (0.33 in average).

Task 4 (Updating own concept)

This task was also for StYLiD only. The confidence and ease of use score very high in average, 3.53 and 3.73 respectively. It required only 2 mins. in average. There were no notable errors and almost no assistance required (0.13).

Task 5 (Structured concepts and instances authoring)

The average confidence score for StYLiD (3.2) was higher than that for Freebase (2.33). The average score for ease of use was much higher for StYLiD (3.4) than for Freebase (2.07). The average time required for this task was lower for StYLiD (5.87 mins.) than for Freebase (10.4 mins.). The average number of errors was quite low and almost the same for both StYLiD (0.87) and Freebase (1). The number of assistance required for Freebase (2) was much higher than that for StYLiD (0.53).

13 people felt more confident with StYLiD, 1 marked Freebase and 1 felt almost the same. 13 people felt that StYLiD was easier to use, 1 said that Freebase was easier for this task and 1 said it was almost the same.

Task 6 (Searching)

The average confidence score for StYLiD (2.8) was higher than that for Freebase (1.53). The average score for ease of use was slightly higher for StYLiD (2.4) than for Freebase (1.47). We see that both the systems scored lower for this task compared to the previous tasks. The average time required for this task was lower for StYLiD (4.93 mins.) than for Freebase (7.07 mins.). However, the time required is not perfectly valid because some people could not complete this task by themselves. The number of assistance required for Freebase (2.4) was slightly higher than that for StYLiD (2). We cannot clearly define what is to be considered as an error in case of searching so errors were not counted.

11 people felt more confident with StYLiD, 1 marked Freebase and 3 felt almost the same. 10 people felt that StYLiD was easier for this task, 1 said that Freebase was easier and 4 felt almost the same with both the systems. Hence, searching still seems to be difficult in both the systems and even more difficult in Freebase.

Table 7 shows the aggregated results from tasks 1, 2, 5 and 6 applicable to both the systems. The overall average confidence score for StYLiD (2.93) is much higher than for Freebase (2). The overall average ease of use score is also much higher for StYLiD (3) than for Freebase (1.9). The average total time is the sum of the average time taken of all these tasks. The average total time taken for StYLiD was much lower (28.47 mins.) than for Freebase (40.93 mins.). The average total error is the sum of all average errors for these tasks. It is seen that StYLiD has more errors (2.93) than Freebase (1.87) in average. The average total assistance is the sum of the average number of assistance required for all these tasks. Overall, Freebase required 10.67 assistances in average which is about twice than that for StYLiD (5.2).

Table 7. Aggregated results from the tasks.

Task	Confidence		Ease		Time (in mins)		Errors		Assistance	
	S	F	S	F	S	F	S	F	S	F
1	2.8	1.87	3.07	1.73	10.23	12.07	0.93	0.47	1.53	3.93
2	2.93	2.27	3.13	2.33	7.43	11.4	1.13	0.4	1.13	2.33
5	3.2	2.33	3.4	2.07	5.87	10.4	0.87	1	0.53	2
6	2.8	1.53	2.4	1.47	4.93	7.07	0	0	2	2.4
Average Total	2.93	2	3	1.9	28.47	40.93	2.93	1.87	5.2	10.67

Note: S stands for StYLiD and F stands for Freebase

Results for non-IT participants

The results were also analyzed separately for the non-IT participants. As stated before, 6 participants did not have any IT background (participants 10-15 in Table 4). The average SUS scores, from these participants only, are 71.67% for StYLiD and 50.42% for Freebase. The scores agree very much with the total SUS scores and StYLiD has better score. It can also be noted that even the participants without any IT background rated the systems quite high, even better than the overall rating.

The aggregated results from these non-IT background participants for each of the tasks are summarized in Table 8. The results are quite similar to the total results (Table 6 and Table 7) and exhibit similar comparative trends. The users felt more confident with StYLiD and found it easier to use for the given tasks. They also required less time using StYLiD for the same tasks. However, the time required by the non-IT background participants was bit higher than the overall observation. Some more errors were made with StYLiD than Freebase, as observed overall. Also much more assistance was needed for Freebase than StYLiD. The assistance needed for the non-IT background participants was naturally bit higher than overall.

The results indicate that the background of the participants do not make much difference in using systems like StYLiD or Freebase. If the system does not require any technical informatics knowledge, the ease of using it does not depend much on the background of the user. Rather it may vary from person to person by other factors like how much he/she uses web applications, general IQ, patience while using a new system, etc. But it should be mentioned that it was bit difficult to explain the tasks and the idea of structured data to people without IT background. However, once explained they could do the tasks almost as good as the participants from the IT background. More experiments in the future with more participants may present some stronger revelations in this matter and also along other aspects that may influence the use of the system.

Table 8. Results for non-IT participants.

Task	**Confidence**		**Ease**		**Time (in mins)**		**Errors**		**Assistance**	
	S	F	S	F	S	F	S	F	S	F
1	2.17	1.67	3	2	12.08	13.33	0.67	1	2.17	4.5
2	2.5	2.5	2.67	2.33	9.08	12.83	1	0.5	1.67	3.83
3	3.17		3		6.08		1.17		0.17	
4	3.17		3.5		2.5		0		0.33	
5	3.33	2.17	3.33	2.17	4.67	10.42	1.17	1.17	0.83	3.33
6	2.33	1.67	2.33	1.5	5.5	8.17	0	0	2.83	3.33
Average Total	2.58	2	2.83	2	31.33	44.75	2.83	2.67	7.5	15

5.2.6 Observations

Following are some main observations noted based on the detailed study of the screen video logs, comments explicitly provided by the participants and behavior of the users while performing the tasks. More significant issues have been listed first. Simple usability issues that can be easily addressed by changed in the system interface have been listed as usability notes.

a) Structured data authoring (Tasks 1 and 5)

StYLiD

- Most people do not distinguish between URL of the page and URI of the structured data object. Some users copied the entry URL from the pop-up of suggested instances. One user entered the URL www.beatles.com as URI.
- Many users had confusion about the entry title and title/name of the data object.
- Users tend to type in multiple values in one line, separated by comma. For e.g., many users typed is all genre values in one line, instead of entering them as separate values. Some users started to type in all the band members in one line.

Usability notes

- Picking up existing instances from the system was not obvious for most users. Some users commented that it was difficult and time consuming to pick up instances from a separate pop-up. They suggested that the system should suggest existing instances while the user types in directly. An auto-complete mechanism can be implemented to solve this issue.

- Linking to Wikipedia was not obvious for some users at first. However, once they figure it out, they could do it easily. In fact, some users entered Wikipedia URIs for attributes even if not instructed for the task.
- The drop down list of enumerated values can be improved, especially for entering multiple values.
- The feature for filtering the data instances in the suggestion pop-up seems to be useful. Many users used it to filter the list of singers.

Freebase

- One main difficulty in Freebase was that it only shows few fields to be entered at the first. After entering these, the user has to open the topic, find the 'empty fields' button and add the remaining fields. This was difficult for most of the users. Users preferred StYLiD because all fields could be posted at once.
- The confusion about the title of the topic and title/name of the instance was also observed in Freebase.
- The auto-complete was handy but some users found it difficult to choose when multiple things exist with the same name. Some users said that it is easy to make mistake in such case and also it makes the system bit slow.
- Freebase automatically pulls contents from its database scraped from Wikipedia and adds to the user's post if the user selects the title form auto-complete. This was confusing and unexpected for most of the users.

Usability notes

- It was difficult to get started for most. Most of the users did not know how to start inputting instance data. The links and buttons are not easy to find.
- There are many confusing labels for different purposes, for eg, “add new”, “add more”, “add some”, “add it!”, "add types to include", “import list”, etc.
- Links for adding description and web links are not spotted easily as they are spread out on the page.

b) Creating and modifying concepts (Tasks 2, 3, 4 and 5)

StYLiD

- The notion of a concept was not obvious for some users. They were confused about the attribute labels and descriptions while entering the concept.
- It was observed that most people try to submit new data directly without checking whether the concept exists or not. People tend to enter the data directly in one step, without creating the concept first. Some people tried to enter the instance data directly in the interface to create the concept schema.
- Specifying the value range was not obvious for many. Many users did not understand that concepts can be selected as range. Many users typed in range concepts as enumeration or description. It was not obvious for users that

specifying the range concepts would help in linking data instances. However, once explained most people could do this properly. Also some people did not expect that multiple concepts may be selected for the range. Some were not sure if they should specify range for all attributes.

- Some users had confusion about specifying concepts as range and enumerating literal values. Some users tried to find concepts for values in enumeration list.
- Most people took the advantage of the capability to modify an existing concept to create a new version, although most users were bit confused when the system says that concept already exists. Few users defined their own concept from scratch. Some users explicitly suggested that they should be able to find out if a concept already exists or not, even with a different name or a similar concept.

Usability notes

- An auto-complete may be better for selecting the range concepts too.
- The concept collection tabs "my concept collection" and "concepts created by me" were used by many people suggesting that they are useful. However, some users are bit confused when the concepts are in one tab and not in another. One user even suggested using color codes to differentiate the different types of concepts.
- Many people list the enumeration values separated by comma, not semi-colon, though it is instructed in the input form to use semi-colon. People do not seem to read instructions provided in the interface so well.

Freebase

- It was difficult to create schema in Freebase for most users because of the strict constraints to be specified. A user cannot save a property without specifying the types properly.
- Choosing proper data types was not easy for many. Different users selected different data types as range for the same property. Users were also not clear about specifying the type formats in detail and this created problems while entering instance data. For e.g., many users had difficulty in entering the date with month and year only. Some users had to go back to the schema definition and change the format for the date/time type or change it into text type.
- Constraints like "restrict to one value" were also not easy to decide.
- One participant tried to specify the range of 'artist' as both 'band' and 'singer' types. However, this is not directly possible in Freebase, unlike StYLiD.

Usability notes

- The meanings of many of options that are shown in the interface for creating properties and specifying types are not clear. For e.g., "short text/ machine readable text", "disambiguator", "horizontal/vertical list" etc.
- The use of reserved words like 'type' created problem for many users. Users commented that such things should be handled transparently by the system.

c) Searching (Task 6)

StYLiD

- It seems that people are very accustomed to keyword search. Most people attempted to perform the Task 6 with various keyword searches first. People are used to simple keyword search than advanced structured search.
- After realizing that keyword search is not enough, some users browsed the concept cloud and listed instances of the 'movie' concept.
- Most users tried to filter the list by keyword again.
- Some users tried to open the movie director's page. However, it leads to the Wikipedia page.
- The system only produces results with exact sub-string matches. So no results were produced when the users made mistakes. Some users explicitly suggested that such inexact searches should be possible.
- A participant even suggested that something like a faceted search may provide better insight into the existing contents.

Usability notes

- Most of the participants had difficulty in spotting the "Advanced search" because it was located just after the "logout" link and without much space in between. Some users suggested that placing it near the keyword search would make it more visible.
- Many users commented that they should not be required to know the structure of the concept for performing the search. The system should automatically suggest the proper attributes. An auto-complete mechanism would make it easy.
- Many users did not type in complete words while searching. An auto-complete would be helpful here too.

Freebase

- It was also observed with Freebase that people start with keyword search. Some users also tried to search for the proper "base" by keywords.
- Most participants had to be guided to open the 'film' base where concepts related to films are located.
- An interesting observation was that most users opened the list of directors rather than the film list although the task was to find out a list of films with the given director and actor. Most users tried to filter this list directly by keywords.
- For most participants, it was very difficult to see how to filter by different properties, e.g., director, actor at the same time. Adding filters was very difficult in Freebase interface.
- For many people it was not obvious that the filters can be added one at a time and it will narrow down the list successively.

- The auto-complete worked for exact matches only. Inexact searches by the users failed.
- Some users commented that something like a faceted search would be better for Freebase too.

Searching was more difficult in Freebase than in StYLiD. In fact, 7 users out of 10 had to give up for task 6 in Freebase.

d) General observations

Besides the above task-specific observations, other general observations were also noted and the users were also asked if they had any overall final comments.

- Many participants were not familiar with such online systems to post structured data. One user commented that both the systems were bit complex compared to usual online applications. She said that she would expect to do everything by herself without any help from start.
- Some users commented that things should be quicker. Some said that it may be cumbersome to input data like this based on schema.
- Most participants stated that Freebase interface was too complex than needed. However, most liked the auto-completion feature in Freebase. Some users explicitly commented that they preferred StYLiD to Freebase because the interface was simpler, the work flow steps were clearer and easier to understand. The workflow was not clear in case of Freebase.
- One participant said that StYLiD seems to be easy to understand even without any manuals.
- Almost none of the users opened the help although both of the system had elaborate help manuals. Only one user in case of StYLiD and another in case of Freebase opened the help when they were really stuck. However, they were not patient to find out the solution and closed the help. Some users suggested that help should be provided "on the spot" when needed. In fact, there are some on the spot help icons but people rarely clicked on them.

Besides there were few other usability issues like follows.

- In case of Freebase, some users were not sure whether the task was done or not because there is no submit or confirm button as in StYLiD.
- Some users commented that small fonts are difficult to see in case of both systems.

5.2.7 Discussion

The quantitative results from SUS scores, task-specific questionnaires and observations made above all indicate that StYLiD is more usable than Freebase. Users felt more confident with StYLiD, found it easier to use, required lesser time to perform the given tasks and required only minimum assistance compared to Freebase. However, StYLiD allowed some more errors due to its relaxed interface. Most of the users were not acquainted with the systems. Thus, users are able to use StYLiD without any training, except for few minor interface issues that can be fixed easily.

People seem to prefer StYLiD because of simple focused interface designed with clear workflow steps. It seems that StYLiD is quite easy to use even without any help manual. In fact, it was observed that people rarely use any help or manual for using such online applications. The background of the participant does not seem to have significant effect on the usability. Rather the individual's confidence and familiarity with web applications seems to make it easy. Nonetheless, it was bit difficult to explain the tasks to users from non-IT background as they are not used to any jargons. However, once the tasks were understood, they could proceed almost as well as the people from IT background.

Hence, the hypothesis of the experiment that StYLiD is more usable than Freebase for the given tasks is supported. However, it should be stressed that this should not be taken as an overall evaluation of Freebase. Freebase has many other attractive features not covered by the experimental tasks. The comparison is only valid for the given tasks of interest.

Some lessons learnt from the observations are summarized here.

- Many people are not familiar with or are not used to such systems for sharing structured data by defining concept schemas.
- Especially, specifying attribute value range is not obvious for many. Moreover, it is difficult to specify strict data types as in Freebase. Many times people are not sure what the data type should be.
- Things like title of the entry and name of the data object, URL or URIs can cause confusion. This issue is often referred to as the "URI crisis" in the Semantic Web community (Oren et al., 2006). It seems that the title also causes a "title crisis".
- People tend to enter all the data directly in one step. Thus, it would be better to have a combined interface to define a new concept and input the instance at the same time. People tend to type in data freely whenever they can. Mechanisms like auto-complete seem to be helpful although it has some limitations. People want quicker ways to input data.
- It was also observed that people would reuse or modify an existing concept rather than starting from scratch.
- It was observed that people are very accustomed to simple keyword search and not so used to structured search or filtering lists. Faceted browsing may also be an intuitive way. The ability for inexact matches also seems to be essential. Hence, proper combination of traditional keyword search with ranked results and structured search or facets may be more effective.

5.3 Experiment on Conceptualization

This experiment is about conceptualizations done by different people on the same thing. This experiment was conducted to study the following questions.

- Can people express their conceptualization in terms a schema with attributes and values?
- How different people conceptualize and model the same thing?

- Can we consolidate independent conceptualizations to form richer consolidated conceptualizations?

The basic hypothesis is that multiple conceptualizations by different people for the same thing can be consolidated.

5.3.1 Experimental design

The participants were given short text passages to read and asked to list down facts about the thing each text is about. They were explained some examples as shown in the Appendix C (Conceptualization Task). They were asked to list down important facts in the form of attributes and values, as in the examples. They were provided blank tables, as shown in the Appendix C (Table for Representing Conceptualization). After reading each passage, or while reading, they had to fill up the provided sheets with attributes and values from the text.

This task can also be considered as manual annotation of the text, as in the annotation experiment with Semantic MediaWiki (Pfisterer et al., 2008). However, in this case, we do not have any given background ontology. The participant has to think of the possible attributes by himself.

About the texts. All the participants were given the same 6 short text passages, in the domain of travel in Japan (included in the Appendix C). These included 2 passages about hotels, 2 about temples and two about museums in Japan. The texts were taken from the websites http://www.japan-guide.com and http://jp.hotels.com (retrieved on February 11, 2009).

5.3.2 About the participants

6 people participated in this experiment, all different from participants of the previous experiment. The participants were aged between 24 to 34 years (average 27.17 years). They were from 5 different countries and all were fluent in English. 3 were PhD students and 3 were Masters students. Only 2 knew what the Semantic Web is though 4 had heard of it. 4 of them had done some database design. They had different fields of specializations and had varied interests. 1 was studying electronics, 1 telecommunication, 1 mathematics and 3 about information and computer science. The participants were all living in Tokyo, Japan. So they would feel comfortable with the texts in domain of travel in Japan.

5.3.3 Results

The participants required about 57 minutes in average (from 45 to 80 minutes) to complete this experiment (for total 6 passages of text, i.e., about 10 minutes per text). This includes the time to read, conceptualize and to write down all the attribute and values on the provided sheet using pencil.

It can also be noted that all the participants named all the concepts with the basic level concept names like temple, hotel and museum. They chose the basic level concept name “temple” and not “Japanese temple”. This supports the theory of basic level of concepts described in Section 3.1.

The participants defined about total 46.33 attributes in average (ranging from 38 to 59) as shown in the table below.

Table 9. Conceptualization by different participants.

participant	Time required (in mins.)	No. of attributes	Missed no. of consolidated attributes
1	45	47	7
2	50	54	6
3	45	38	16
4	60	39	15
5	80	59	3
6	60	41	9
Average	56.67	46.33	9.33

Consolidation

The attributes defined by the different participants for each text instance were then consolidated manually. Corresponding attributes were aligned and unified as consolidated or global attributes. Total 47 such consolidated attributes were formed. These together form the global consolidated schemas.

None of the users defined all attributes corresponding to the attributes in the consolidated schema. Only when the attributes defined by all are consolidated, the complete global schema is formed. Table 9 shows the number of attributes that each participant missed from the consolidated schema. On the average 9.33 attributes were missed by each participant. Minimum 3 to maximum 16 attributes were missed.

This shows that none of the participants have the same conceptualization over the same given texts. The participants have different conceptualizations modeled with different number of attributes. Only when all these conceptualizations from all the participants are consolidated, the full consolidated schema is formed. The conceptualization by each participant only covers a part of this consolidated schema, missing some attributes. Participants defining more attributes miss less from this global schema and those defining less attributes miss more from the global schema.

The consolidated attributes

The different types of alignment relations found in the consolidated attributes are shown in Table 10.

Table 10. Different types of alignments found.

Alignment relation	**Number**
Equivalent	16
almost equivalent	8
Composite	3
Similar	3
mixed (unifiable)	9
mixed (complex)	7
Unary	1
Sum	47

16 alignments, a majority of the alignments, were found to be equivalent relations. 8 alignments were almost equivalent with slight differences in coverage, though the intension is roughly the same. For e.g.,

Attribute	*Value*
a.Exhibition_items	historic buildings from Tokyo area
b.Exhibition_types	historic buildings
a.Period	Meiji Period (1868-1912) or more recent times
b.exhibition_items_period	Meiji Period (1868-1912)

The prefixes a and b before the attribute names denote the different schema created by different persons.

There were 3 composite alignment relations. By composite we mean that when two or more attributes of one schema is combined it can be aligned to a single attribute in another schema. For e.g.,

Attribute	*Value*
a.name	Kiyomizudera
a.english meaning	Pure Water Temple
b.name	Kiyomizudera (Pure Water Temple)

3 alignments were similar with similar intension for the values. For eg,

Attribute	*Value*
a.type	open air museum
b.type	Historical museum

9 of consolidated attributes had mixed relations that could not be mapped directly. However, these attributes are unifiable. All the attribute values are related and they are about one thing only when they are combined. For eg, the participants modeled different information about the access to a museum. Some mentioned the distance from station, some mentioned access_routes, train line information, nearest stations, time required from the nearest station, and so on. However, all these information are

complementary and can be combined to form the complete information about the access to the museum.

7 other consolidated attributes also had mixed relations that could not be mapped directly but these were more complex. People may put emphasis on different aspects and make different conceptualizations. For e.g., some modeled all the attractions of a temple by a single “attractions” attribute, some considered the interesting features, the scenic views, the food as attributes, some modeled the deity in the temple in more detail and so on. Some even had nested conceptualizations to describe few things in detail. Similarly, the facilities of a hotel were differently modeled. Some separated things like internet and TV as facilities and other things like pool, bowling alley, golf as recreation. One person even modeled internet and TV as attributes with binary values Yes/No. Some put the restaurants as the hotel facilities while others separated it into the food and drinks attribute. These attributes are very flexible and it seems that people have wide number of ways of conceptualizing with different perspectives and emphasis. There is no distinct way of modeling and easy way of consolidating few attributes like these.

Finally the 1 unary alignment relation includes only a single attribute defined by a single person and no others defined a corresponding attribute.

Attribute label similarity

If we see the attribute labels in the aligned attributes, we observe different types of label similarities as shown in Table 11. Only 9 consolidated attributes have the same labels in the constituent attributes (e.g., a.name, b.name). 16 have similar attribute labels mapped (e.g., a.founded, b.established, c.construction). 21 consolidated attributes consist of mixed types of labels including similar, different and related labels (e.g., a.parts, b.exhibition_villages, c.sections, d.types_of_buildings).

Table 11. Attribute label similarity.

Type of similarity	No. of consolidated attributes
same	9
similar	16
mixed	21

This shows that attribute labels need not be necessarily the same or similar to be aligned with other attributes. Although most attributes could be aligned, as equivalent or compatible, the attribute labels were mostly of mixed variety or similar to some extent.

5.3.4 Discussion

We can draw some conclusions from this experiment as follows. These answer the questions posed at the beginning of this section.

People are able to conceptualize and express things structured in a schematic way in terms of attributes and values. They can express such conceptualizations within reasonable amount of time without much difficulty.

Different people conceptualize the same thing in different ways. In a particular conceptualization, each different person may have at least some unique aspects to add. A single person usually does not consider all the possible aspects for conceptualizing the thing. Many complementary attributes can be found in the conceptualizations and these can be simply added up consistently to form the complete global conceptualization.

It is possible to combine all the conceptualizations of the same thing by different people and form a rich consolidated conceptualization. Most of these conceptualizations overlap significantly. In fact, almost all attributes overlap with at least some attribute in someone else’s conceptualization. Most of the overlapping parts can be aligned. Most of these can be aligned by simple equivalence relation or almost equivalent. Some relations may be more complex but still can be mapped or unified. The granularity or depth of conceptual modeling may differ. Some people conceptualize in detail with more attributes while others may conceptualize the same thing in a shallow way with fewer attributes. The attributes labels of aligned attributes may be same, similar, related or even completely different.

Few related attributes may be complex showing different possible ways to conceptualize. It may be difficult to map or unify these few attributes reflecting the unique perspectives of independent individuals. However, it was observed that most parts of the conceptualizations overlap and can be consolidated to form a shared global conceptualization, in spite of the fact that the individual conceptualizations were done independently. Hence, the hypothesis that multiple conceptualizations can be consolidated is valid.

5.4 Experiments on Existing Data

Some experiments were conducted using data from Freebase[85] (Bollacker, Tufts, Pierce, & Cook, 2007), a community-driven open database of world's information. These experiments were mainly intended to make some observations about the user-defined concepts, verify our views about multiple conceptualizations and to validate the automatic techniques used for concept consolidation and grouping.

5.4.1 About Freebase and the dataset

Freebase has a large collection of user-defined types and data, freely available. Concepts are called "types" in Freebase and each type is associated with a schema. Freebase data can be queried using the Metaweb Query Language (MQL) which was made by Metaweb Technologies[86], the company which built Freebase. Freebase provides a convenient API and interface for querying.

The following steps were followed for experimenting on the Freebase data.

1. Retrieve all user-defined concepts by querying
2. Clean the collection
3. Perform concept consolidation
4. Group similar concepts and visualize them

Details are described below.

Freebase User-defined Types

All user-defined types were retrieved from the system (as on May 20, 2008). This was done by querying Freebase directly through the provided API. Only user-defined types, contained in the user's domain, were considered and not standard Freebase types, as we are interested in how different people model their data. The types in user-defined spaces were filtered by considering only the type IDs starting with "/user/". The query results are produced by Freebase in JSON[87] format which can be easily parsed.

Cleaning. Many unwanted user-defined types had to be filtered out including types created by the users for test purposes, spam, etc. (although Freebase provides a sandbox for test purpose a lot of test data was also found in the main site). String matches and regular expressions were used to filter out obvious unwanted types.

5.4.2 Observations about the data

There were 2120 total types without any cleaning. Cleaning resulted in 1,852 types. Types with no instance were further filtered out leaving a total of 1,412 types. There were total 500 users who defined at least one concept. This considerable number of

[85] http://www.freebase.com/
[86] http://www.metaweb.com/
[87] http://www.json.org/

types defined by many users, from a system in its initial stages, indicates that there is, in fact, a wide variety of data types different users are interested in and that users are willing to define their own concept schemas. It was also observed that most people define very few concepts (only 1 to 5), as illustrated in the histogram below (Figure 56), but altogether it amounts to so many concepts. This makes up the long tail of information types that small groups of people are interested in. Only very few users define more concepts (maximum number of concepts was 144, probably bulk imported which is possible in Freebase)

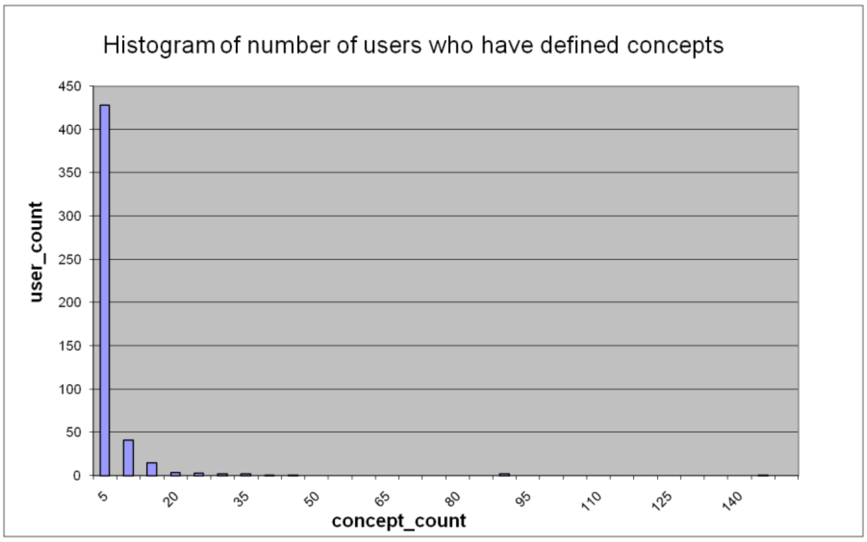


Figure 56. Histogram of the number of users who have defined concepts.

The number of instances of concepts ranged widely from 1 to 29,146 (average 78.17). However, the histogram of instance counts show that most concepts have below 50 instances (Figure 57). There are only very few user accounts with very large number of instances (these also must have been bulk imported because it is unlikely that a user would input so many instances manually).

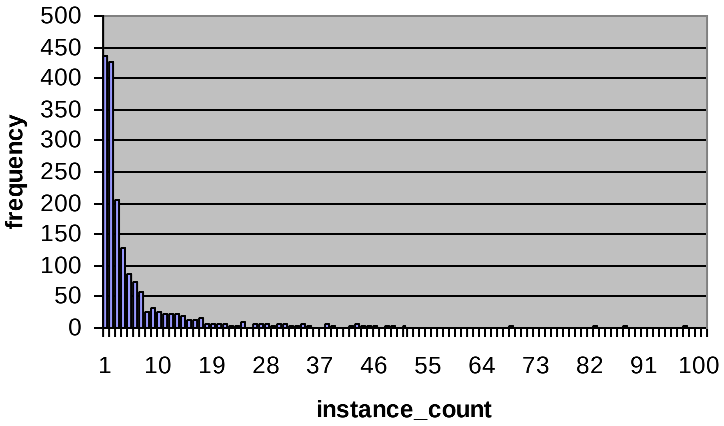


Figure 57. Histogram of instance counts.

5.4.3 Concept consolidation

The proposed method of concept consolidation was tested by automatically consolidating user-defined Freebase types. All possible pairs of concepts were compared and user-defined types with the same name, slight morphological variants or synonyms were determined. This was done by selecting the pairs with the WordNet-based similarity measure of 1.0. It was assumed that these represent the same concept and, hence, can be consolidated.

From the 1,412 concepts, 57 such groups were found. There may also be other sets of same concepts, but differently named, that were not considered. Nonetheless, the observations already support our view that different people define the same concept in their own ways. Even up to 6 versions of the same concept were found. There were also a few cases where the same user defined multiple versions of a concept.

Schema Alignment

Then, for each consolidated group of concepts found above, all pairs of the schemas were aligned automatically. A consolidated concept is formed for each group of concepts. The aligned set of attributes forms a consolidated attribute. For example, the sets of concepts {Recipe (r_1), Recipe (r_2), Recipes (r_3) ….} are consolidated to form a single consolidated concept "Recipe" with the following set of consolidated attributes, derived from each of the concepts (r_1, r_2, r_3 represent 3 different sources).

- **{**r_1#ingredient, r_2#ingredients, r_3#materials**}**
- **{**r_1#steps, r_2#instructions**}**
- r_3#directions
- r_2#tools_required
- r_3#taste
- r_3#author ……

aligned attribute sets

The same Alignment API implementation used in StYLiD, based on WordNet-based similarity between the labels, was used for the purpose. A similarity threshold of 0.6 was used in the alignment API implementation for determining attribute alignments. The threshold has been chosen by trial to achieve a good coverage with precision. 44 sets of aligned attributes (forming the consolidated attributes) with total 51 alignment relations were found from the consolidated concepts.

Most of these alignments were found to be reasonable along with about 6 wrong alignment relations (i.e., 45 right relations or a precision of 88.24%). A PhD student with informatics background helped in the evaluation judgments. The sets of aligned attributes were made into a list. The judge was asked to mark the attributes in each set that do not seem to match with the other attributes in the set, if any. The attributes in each set were considered to be connected by similarity relations.

Then, all the correct alignment relations were listed manually, though it is difficult to point out all and some may be subjective. Total 67 alignment relations were found. Hence, the recall is about 67.16%. So we see that even with simple methods for schema alignment, the precision can be quite good with a satisfactory recall. Certainly, more alignments with better accuracy can be achieved with more sophisticated techniques and human involvement.

Table 12 shows some more statistics about the consolidated concepts found. It shows that there were 2 to 6 constituent concepts in a consolidated concept and upto 5 aligned sets of attributes. The concepts had a good number of instances overall indicating that the concepts were not just dummy concepts made for testing.

Table 12. Statistics about the consolidated concepts.

	No. of constituent concepts	No. of aligned set of attributes	No. of instances
maximum	6	5	1472
minimum	2	0	2
average	2.25	0.77	67.54

5.4.4 Concept grouping

Using the process described earlier in Section 3.7, groups of similar user-defined concepts were determined. Concept groupings were formed using different values for the concept similarity threshold. Concepts were considered to be related if the similarity was above the threshold. Each of these groups of concepts was shown to the same human judge. He was asked to mark if any concept does not seem to belong to the group. The similarity relations were considered as concept pairs and the judge was asked if each pair was similar or not. The total number of nodes and relations in each cluster, total number of correct nodes and total number of correct relations were noted. Observations are summarized in Table 13 below (for $w_1 = 0.7$ and $w_2 = 0.3$ in Equation 1).

The different values for the threshold were used to observe its effect in finding relations and groupings. The shown values are not optimal ones, rather these help in observing the trend. Also, the results are not meant to be perfect due to the subjective nature of the decisions. However, it gives us a general idea and, moreover, relative observations helps us to analyze the effect of changing the parameters. Appropriate values for the threshold and the weights may be determined iteratively.

Table 13. Concept grouping results for different thresholds ($w_1 = 0.7$, $w_2 = 0.3$).

Threshold	Groups found	Nodes covered	Relations found	Correct nodes (%)	Correct relations (%)
0.5	177	639 (45.25%)	608	87.95	79.44
0.8	108	275(19.48%)	194	95.64	94.33

The observed results show that groups of similar concepts can be formed with good precision. Further, it suggests that with tighter threshold more precise groups can be formed but the coverage of nodes and relations decreases, forming lesser number of groups. With a lower threshold, the coverage can increase significantly while the precision still remains quite good. Thus, the threshold can be tuned to produce appropriate level of groupings.

Table 14 shows the effect of varying the weights for *NameSim* and *SchemaSim* (w_1 and w_2 respectively) in Equation 1 (Section 3.7.1). Different combinations of weights were used keeping a constant threshold of 0.8. Grouping with schema similarity alone ($w_1 = 0$, $w_2 = 1$), ignoring concept name similarity, produced poor results. Only 42.65% nodes were grouped correctly. The coverage was also low (9.63%). This is because even dissimilar concepts may sometimes have similar schemas. On the other hand, using the concept name similarity alone ($w_1 = 1$, $w_2 = 0$), ignoring the schema similarity, already produces quite good results. There were 78.97 % correct nodes and 83.43% correct relations. The coverage was also much higher (16.5%) than the previous case. Hence, the concept name similarity plays a much more significant role than the schema similarity in calculating overall schema similarity.

However, weighted combination of both the similarity measures produced better results than the cases where either was used alone. For $w_1 = 0.7$ and $w_2 = 0.3$, the number of groups found, total nodes and relations covered, percentage of correct nodes and relations were all higher than the other two cases. The result is slightly better than the case when name similarity is used alone. The optimal weight parameters may be determined by testing iteratively. However, it should be admitted that the process of evaluating correct nodes and relations by human is tedious and imperfect. Hence, only few important combinations were tested.

Table 14. Concept grouping results by varying weight parameters (threshold = 0.8).

Parameters Threshold = 0.8	**Groups found**	**Nodes covered**	**Relations found**	**Correct nodes (%)**	**Correct relations (%)**
$w_1 = 0$, $w_2 = 1$	60	136 (9.63 %)	98	42.65	32.65
$w_1 = 1$, $w_2 = 0$	101	233 (16.5 %)	175	78.97	83.43
$w_1 = 0.7$, $w_2 = 0.3$	108	275(19.48%)	194	95.64	94.33

5.4.5 Discussion

The observations and results from the experiments of user-defined Freebase data can be summarized as follows.

- There is a wide variety of concepts different users are interested.
- Users define their own concept schemas. Most people define very few concepts and contribute few instances.
- Different people define the same concept in different ways.
- Concept consolidation can be done with quite good precision and satisfactory recall even with simple methods for automatic alignment.
- The automatic method used for grouping similar concepts can have good precision. The threshold can be tuned for obtaining appropriate trade-off between the precision and coverage.
- The concept name similarity is more significant than the schema similarity in the weighted combination to calculate the concept similarity.

5.5 Summary of Evaluation

The findings of all the experiments and observations can be summarized in following points.

Usability

- People are able to start using StYLiD with no training or minimum assistance.
- StYLiD is more usable than existing platforms like Freebase.
- The flexible and relaxed interface makes contribution easy although some noise has to be tolerated.

User-defined concepts

- There is a wide variety of concepts different people are interested in and they can define their own concepts.
- People are able to conceptualize things and express in terms of schema.
- Different people conceptualize the same thing in different ways.

Concept consolidation and grouping

- Most user-defined conceptualizations overlap significantly and contribute complementary attributes. It is possible to combine them into a single consolidated conceptualization.
- Most of the overlapping parts of the concepts can be aligned by simple relations like equivalence.
- Even with simple methods for schema alignment satisfactory precision and recall can be obtained while consolidating user-defined schemas.
- With the applied method, similar concepts can be grouped with quite good precision and the threshold can be tuned to produce appropriate level of groupings.

5.6 Some Practical Applications

The implemented systems can be used for various practical applications. StYLiD can be used for purposes including social information sharing, content management and information integration. It is already being used for some real world applications some of which are mentioned below. The structured data and knowledge structures produced can further be used in applications like OntoBlog to organize unstructured contents.

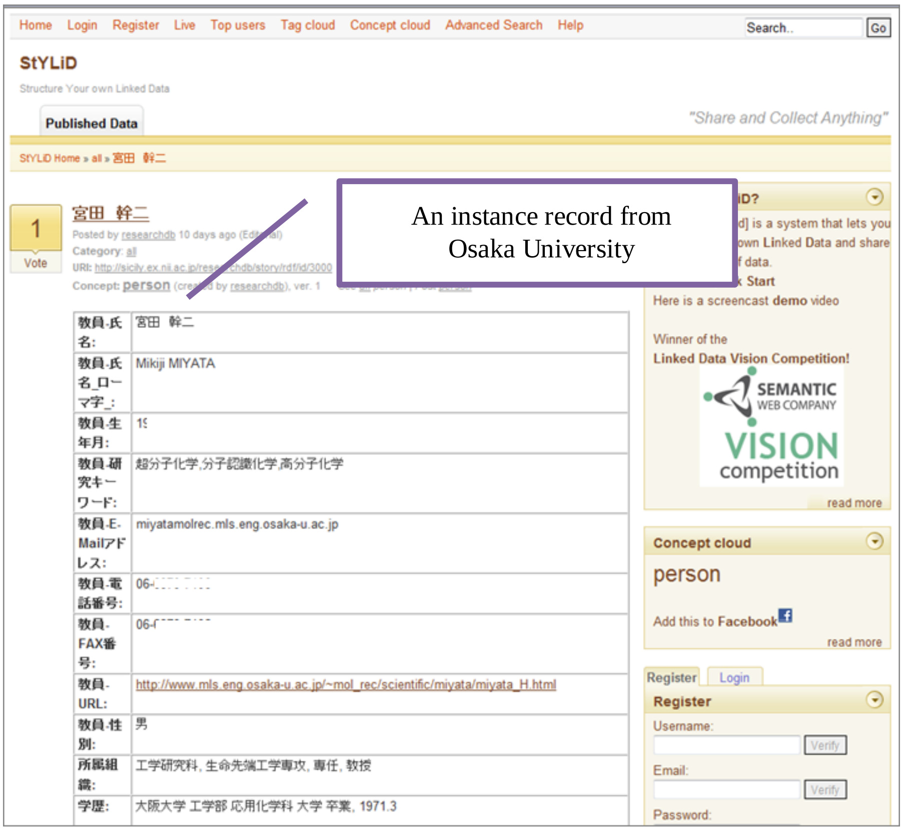
5.6.1 Integration of research staff directories

StYLiD is being used in a project to integrate research staff directories from various Japanese universities (Kurakawa et al., 2009). Most Japanese universities maintain their own staff directories on the web. These directories have been developed independently and use different conceptual schemas for describing information about the staff. The individual sites often provide useful search services. Unfortunately, the search services only work for the particular conceptual schema and retrieves results from the particular university database only. StYLiD can be used to map the conceptual schemas of the different universities and provide a consolidated view over them. Then, federated search over all the directories can be provided though a single unified interface. Hence, the integrated data collection can be accessed uniformly through a single website like a vertical portal. A prototype implementation is available online[88].

Importing the data. Currently, data from the Osaka University and Nagoya University of Japan have been imported into the system. The websites of the universities were crawled and the structured data was scraped from the pages and represented in XML format. The XML format has nested structure at several places due to the nested presentation of data on the web pages. To replicate the schemas in StYLiD, hierarchical parts of the XML schema were flattened. The flattening of nested XML structures was done by concatenating the labels of the child nodes to the parent node successively. For example, if a “Teacher” node has two sub-nodes “Name” and “Sex” in the XML representation, these are flattened to create two attributes “Teacher_name” and “Teacher_sex”.

The flattening was done while importing the XML data to fit into the flat schemas in StYLiD. The data was imported using simple API-like function calls. The imported data was stored as RDF triples. The database consists of 1106 records from Osaka University and 1888 records from Nagoya. Screenshots of the system showing an instance record from Osaka university and Nagoya university are shown in Figure 58 and Figure 59 respectively.

[88] http://sicily.ex.nii.ac.jp/researchdb



An instance record from Osaka University

宮田　幹二

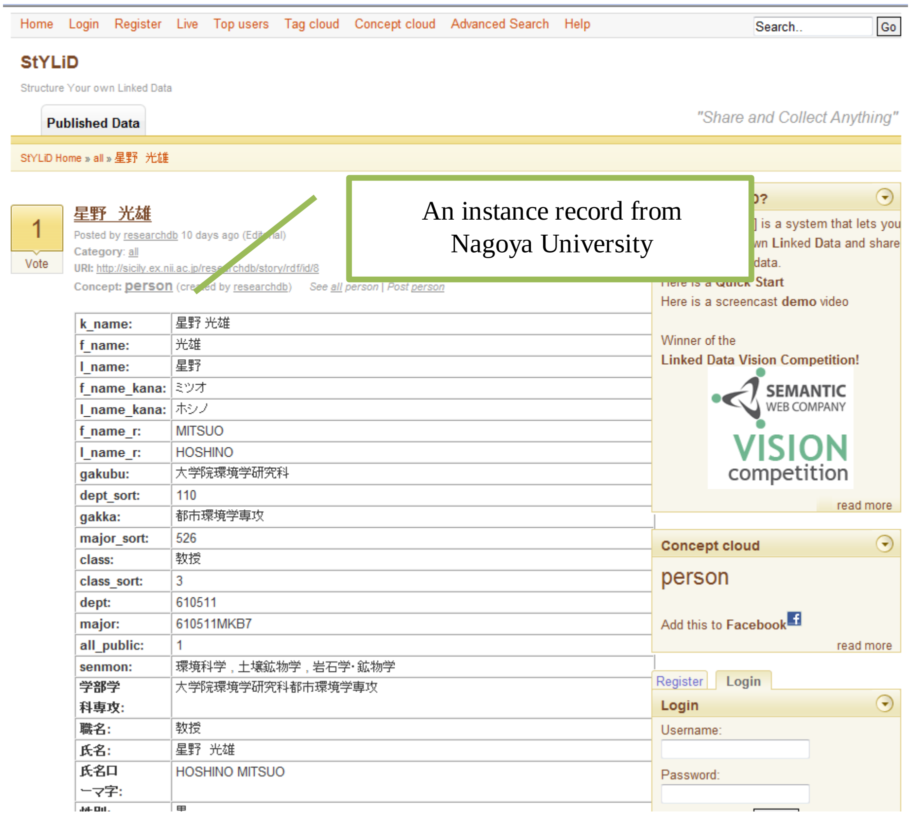
Posted by researchdb 10 days ago (Editorial)
Category: all
URI: http://sicily.ex.nii.ac.jp/researchdb/story/rdf/id/3000
Concept: person (created by researchdb), ver. 1

教員-氏名:	宮田　幹二
教員-氏名_ローマ字_:	Mikiji MIYATA
教員-生年月:	19
教員-研究キーワード:	超分子化学,分子認識化学,高分子化学
教員-E-Mailアドレス:	miyatamolrec.mls.eng.osaka-u.ac.jp
教員-電話番号:	06-
教員-FAX番号:	06-
教員-URL:	http://www.mls.eng.osaka-u.ac.jp/~mol_rec/scientific/miyata/miyata_H.html
教員-性別:	男
所属組織:	工学研究科, 生命先端工学専攻, 専任, 教授
学歴:	大阪大学 工学部 応用化学科 大学 卒業, 1971.3

iD?
d] is a system that lets you
own Linked Data and share
f data.
k Start
Here is a screencast demo video

Winner of the
Linked Data Vision Competition!

Figure 58. A data instance from Osaka University.



An instance record from Nagoya University

星野　光雄

Posted by researchdb 10 days ago (Editorial)
Category: all
URI: http://sicily.ex.nii.ac.jp/researchdb/story/rdf/id/8
Concept: person (created by researchdb)　See all person | Post person

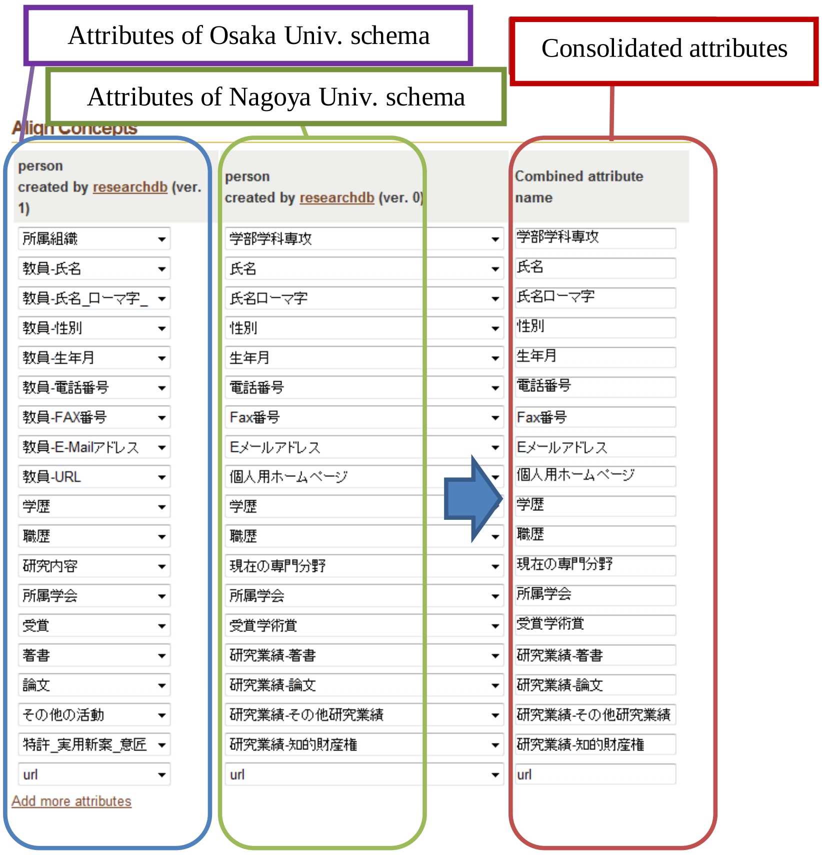
k_name:	星野 光雄
f_name:	光雄
l_name:	星野
f_name_kana:	ミツオ
l_name_kana:	ホシノ
f_name_r:	MITSUO
l_name_r:	HOSHINO
gakubu:	大学院環境学研究科
dept_sort:	110
gakka:	都市環境学専攻
major_sort:	526
class:	教授
class_sort:	3
dept:	610511
major:	610511MKB7
all_public:	1
senmon:	環境科学，土壌鉱物学，岩石学･鉱物学
学部学科専攻:	大学院環境学研究科都市環境学専攻
職名:	教授
氏名:	星野　光雄
氏名ローマ字:	HOSHINO MITSUO

D?
] is a system that lets you
wn Linked Data and share
data.
Here is a Quick Start
Here is a screencast demo video

Winner of the
Linked Data Vision Competition!

Figure 59. A data instance from Nagoya University.

Aligning the data. The separately maintained directories follow different conceptual schemas. These were aligned in StYLiD by creating a consolidated concept representing research staff. The system automatically suggested total 10 alignments. It was observed that all these automatic suggestions were correct. The attribute labels were almost all in Japanese. The alignment module worked well even if it was designed mainly for English text. This may be due to the fact that substrings of the labels matched substantially. These alignments were verified by human and 9 other missing alignments were added resulting into total 19 mappings. Hence, the precision of alignment was observed to be high and the recall above 50%. The completed alignment can be seen in Figure 60 showing the sets of attributes from the schemas of Osaka University and Nagoya University aligned to form the set of consolidated attributes.


Attributes of Osaka Univ. schema
Attributes of Nagoya Univ. schema
Consolidated attributes

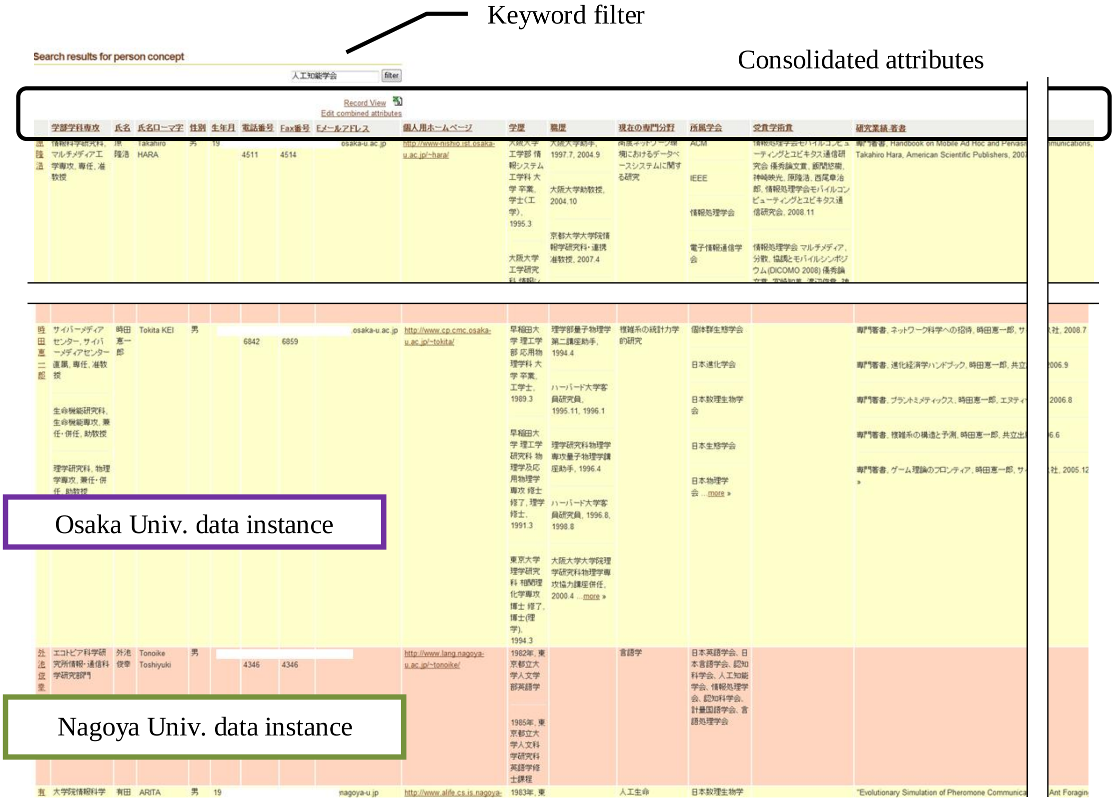
Align Concepts

person created by researchdb (ver. 1)	person created by researchdb (ver. 0)	Combined attribute name
所属組織	学部学科専攻	学部学科専攻
教員-氏名	氏名	氏名
教員-氏名_ローマ字_	氏名ローマ字	氏名ローマ字
教員-性別	性別	性別
教員-生年月	生年月	生年月
教員-電話番号	電話番号	電話番号
教員-FAX番号	Fax番号	Fax番号
教員-E-Mailアドレス	Eメールアドレス	Eメールアドレス
教員-URL	個人用ホームページ	個人用ホームページ
学歴	学歴	学歴
職歴	職歴	職歴
研究内容	現在の専門分野	現在の専門分野
所属学会	所属学会	所属学会
受賞	受賞学術賞	受賞学術賞
著書	研究業績-著書	研究業績-著書
論文	研究業績-論文	研究業績-論文
その他の活動	研究業績-その他研究業績	研究業績-その他研究業績
特許_実用新案_意匠	研究業績-知的財産権	研究業績-知的財産権
url	url	url

Add more attributes

Figure 60. Alignment of concepts from two universities.

Accessing the integrated data. The data from the different sources can now be accessed uniformly with the help of the consolidated concept. Figure 61 shows the integrated data listed in a uniform table view. An instance each from the Osaka University and Nagoya University are shown uniformly represented. This table can be filtered, sorted or exported to spreadsheet applications for desired processing. The integrated repository may also be searched uniformly in a structured way using the consolidated concept schema.



Search results for person concept

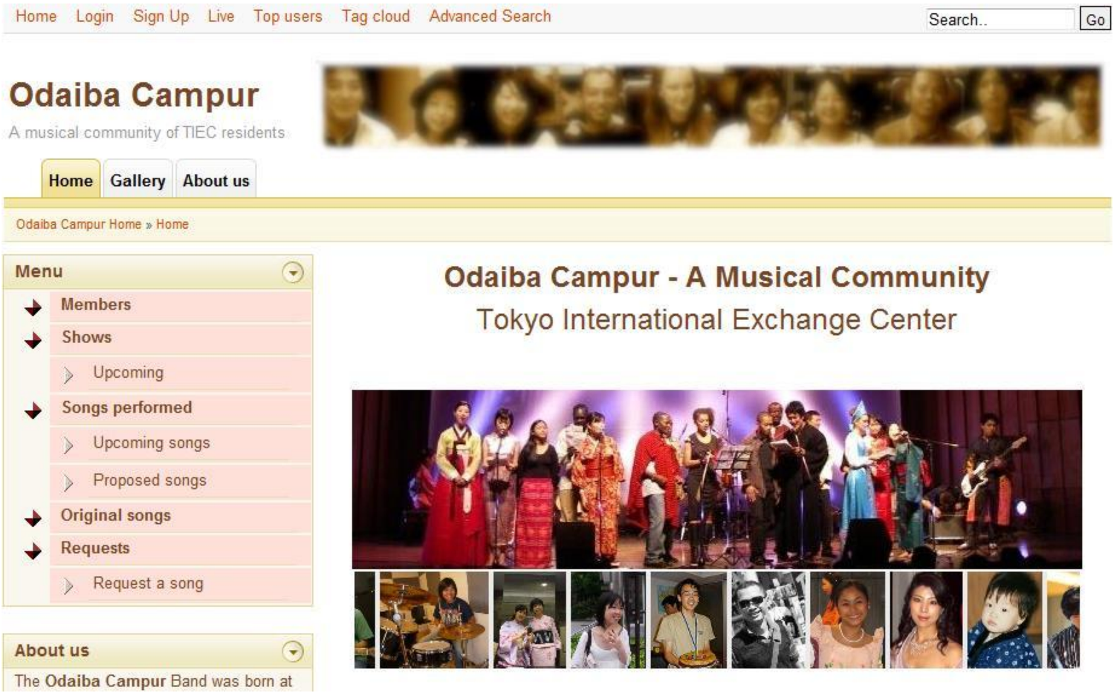
人工知能学会 filter

Record View
Edit combined attributes

	学部学科専攻	氏名	氏名ローマ字	性別	生年月	電話番号	Fax番号	Eメールアドレス	個人用ホームページ	学歴	職歴	現在の専門分野	所属学会	受賞学術賞	研究業績 著書	
原隆浩	情報科学研究科, マルチメディア工学専攻, 専任, 准教授	原 隆浩	Takahiro HARA	男	19	4511	4514	osaka-u.ac.jp	http://www-nishio.ist.osaka-u.ac.jp/~hara/	大阪大学 工学部 情報システム工学科 大学 卒業, 学士(工学), 1995.3 大阪大学 工学研究科 情報シ	大阪大学助手, 1997.7, 2004.9 大阪大学助教授, 2004.10 京都大学大学院情報学研究科・連携准教授, 2007.4	高度ネットワーク環境におけるデータベースシステムに関する研究	ACM IEEE 情報処理学会 電子情報通信学会	情報処理学会モバイルコンピューティングとユビキタス通信研究会 優秀論文賞, 飯間悠樹, 神崎映光, 原隆浩, 西尾章治郎, 情報処理学会モバイルコンピューティングとユビキタス通信研究会, 2008.11 情報処理学会 マルチメディア, 分散, 協調とモバイルシンポジウム(DICOMO 2008) 優秀論文賞, 宮崎知美, 渡辺俊貴, 神	専門書, Handbook on Mobile Ad Hoc and Pervasi Takahiro Hara, American Scientific Publishers, 200	mmunications,
時田恵一郎	サイバーメディアセンター, サイバーメディアセンター直属, 専任, 准教授 生命機能研究科, 生命機能専攻, 兼任・併任, 助教授 理学研究科, 物理学専攻, 兼任・併任, 助教授	時田 恵一郎	Tokita KEI	男		6842	6859	.osaka-u.ac.jp	http://www.cp.cmc.osaka-u.ac.jp/~tokita/	早稲田大学 理工学部 応用物理学科 大学 卒業, 工学士, 1989.3 早稲田大学 理工学研究科 物理学及応用物理学専攻 修士 修了, 理学修士, 1991.3 東京大学 理学研究科 相関理化学専攻 博士 修了, 博士(理学), 1994.3	理学部量子物理学第二講座助手, 1994.4 ハーバード大学客員研究員, 1995.11, 1996.1 理学研究科物理学専攻量子物理学講座助手, 1996.4 ハーバード大学客員研究員, 1996.8, 1998.8 大阪大学大学院理学研究科物理学専攻協力講座併任, 2000.4 ...more »	複雑系の統計力学的研究	個体群生態学会 日本進化学会 日本数理生物学会 日本生態学会 日本物理学会 ...more »		専門書, ネットワーク科学への招待, 時田恵一郎, サ 専門書, 進化経済学ハンドブック, 時田恵一郎, 共立 専門書, ブラントミメティックス, 時田恵一郎, エヌティ 専門書, 複雑系の構造と予測, 時田恵一郎, 共立出 専門書, ゲーム理論のフロンティア, 時田恵一郎, サ »	社, 2008.7 2006.9 2006.8 6.6 社, 2005.12
外池俊幸	エコトピア科学研究所情報・通信科学研究部門	外池 俊幸	Tonoike Toshiyuki	男		4346	4346		http://www.lang.nagoya-u.ac.jp/~tonoike/	1982年, 東京都立大学人文学部英語学 1985年, 東京都立大学人文科学研究科英語学修士課程		言語学	日本英語学会、日本言語学会、認知科学会、人工知能学会、情報処理学会、認知科学会、計量国語学会、言語処理学会			
有	大学院情報科学	有田	ARITA	男	19			nagoya-u.jp	http://www.alife.cs.is.nagoya-	1983年, 東		人工生命	日本数理生物学		"Evolutionary Simulation of Pheromone Communica	Ant Foragin

Figure 61. Uniform table view of integrated data from the university directories.

5.6.2 A musical community website



Odaiba Campur - A Musical Community

Tokyo International Exchange Center

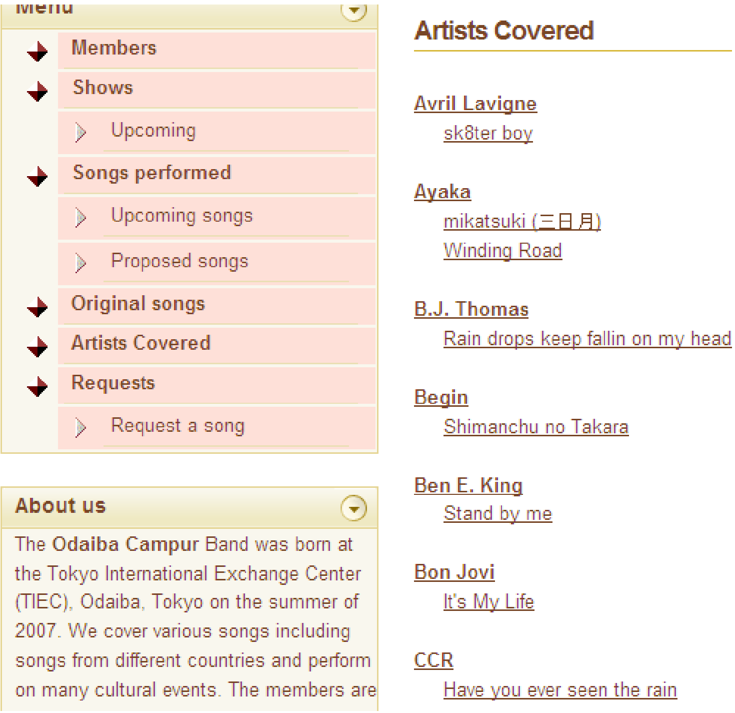
The Odaiba Campur Band was born at

Figure 62. The TIEC musical community website.

StYLiD has also been used to create a dynamic website for a musical community[89] in the Tokyo International Exchange Center, Odaiba, Tokyo. The community consists of talented residents of the exchange center coming from different countries. It is an ad hoc community because new residents join regularly and old ones leave, although some are constant members. The community performs in musical shows during various cultural events.

A screenshot of the site is shown in Figure 62. Various types of information are published and shared through this community website. Various concepts have been defined, for e.g., members, songs, shows, etc and data is posted in structured schematic form using these concepts. The data is highly interlinked. For e.g., shows are linked to the songs performed and songs are linked to the members appearing in the song. The system automatically produces the backlinks in reverse direction. Only members of the community are authorized to post data. However, anyone can post comments or vote contents.

The site acts as an informal information management system for the community as well as its homepage for web presence and publicity. A number of useful views have also been created implemented through SPARQL queries. For example, the list of artists, whose songs the group has covered, is automatically generated as shown in Figure 63. The list of covered songs of each artist is also shown. Similarly, other simple views like the upcoming songs, songs performed, upcoming shows, etc are also provided in the site menu.



The **Odaiba Campur** Band was born at the Tokyo International Exchange Center (TIEC), Odaiba, Tokyo on the summer of 2007. We cover various songs including songs from different countries and perform on many cultural events. The members are

Artists Covered

Avril Lavigne
sk8ter boy

Ayaka
mikatsuki (三日月)
Winding Road

B.J. Thomas
Rain drops keep fallin on my head

Begin
Shimanchu no Takara

Ben E. King
Stand by me

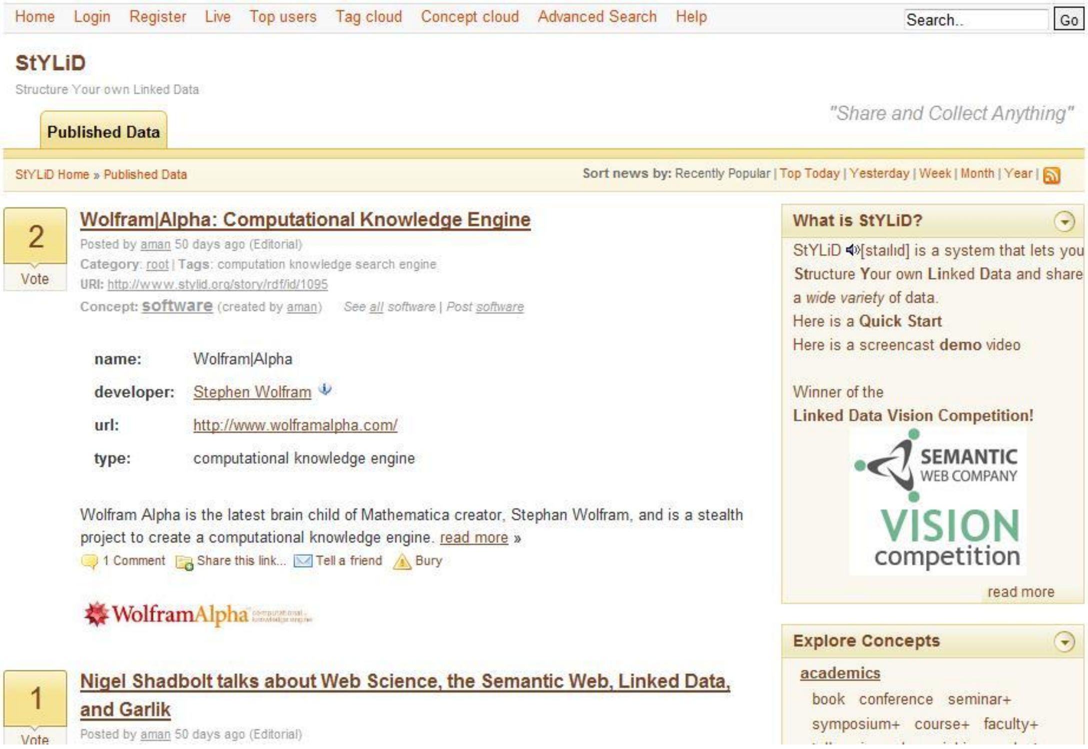
Bon Jovi
It's My Life

CCR
Have you ever seen the rain

Figure 63. View showing list of artists covered.

[89] http://www.odaibacampur.com/

5.6.3 Social data bookmarking site StYLiD.org



StYLiD

Structure Your own Linked Data

"Share and Collect Anything"

Wolfram|Alpha: Computational Knowledge Engine

Posted by aman 50 days ago (Editorial)

Category: root | Tags: computation knowledge search engine

URI: http://www.stylid.org/story/rdf/id/1095

Concept: software (created by aman) See all software | Post software

name:	Wolfram|Alpha
developer:	Stephen Wolfram
url:	http://www.wolframalpha.com/
type:	computational knowledge engine

Wolfram Alpha is the latest brain child of Mathematica creator, Stephan Wolfram, and is a stealth project to create a computational knowledge engine. read more »

Nigel Shadbolt talks about Web Science, the Semantic Web, Linked Data, and Garlik

Posted by aman 50 days ago (Editorial)

What is StYLiD?

StYLiD [staılıd] is a system that lets you Structure Your own Linked Data and share a *wide variety* of data.

Here is a **Quick Start**

Here is a screencast **demo** video

Winner of the

Linked Data Vision Competition!

Explore Concepts

academics

book conference seminar+ symposium+ course+ faculty+

Figure 64. Screenshot of www.stylid.org

An installation of StYLiD has been online as the social data bookmarking site *www.stylid.org* (Figure 64). It is an open social site where anyone can bookmark and share various types of data by defining their own structured concepts. Initially, to bootstrap the site, it was populated with some sample data in the academic domain with different versions of concepts like faculty, courses, seminars, etc. Heterogeneity is common in such data because academic institutes have different systems and formats. Some wrappers were created using the Dapper online service to scrap some data from the website of NII, Japan and others. However, the site is open for general purpose.

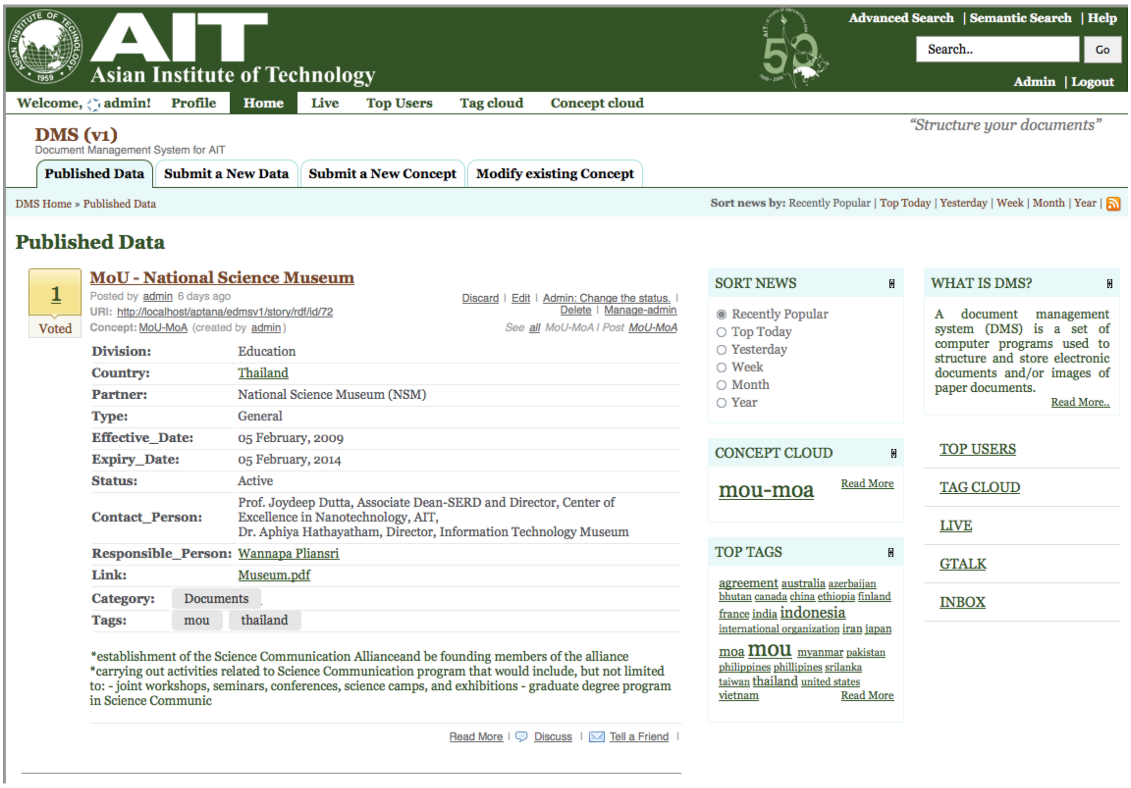
As of May 4, 2009, the following statistics has been noted for the website, using Google Analytics[90].

- 120 registered users
- Total 75 concepts, 613 instances
- 9 concepts with multiple versions
- 2512 absolute unique visitors from 96 countries

5.6.4 A document management system at AIT

StYLiD is also being used to develop an ad hoc document management system (DMS) in the Asian Institute of Technology (AIT), Thailand. A screenshot of the DMS system is shown in Figure 65.

[90] http://www.google.com/analytics/



AIT
Asian Institute of Technology
Advanced Search | Semantic Search | Help
Search..
Go
Admin | Logout
Welcome, admin!
Profile
Home
Live
Top Users
Tag cloud
Concept cloud
"Structure your documents"
DMS (v1)
Document Management System for AIT
Published Data
Submit a New Data
Submit a New Concept
Modify existing Concept
DMS Home » Published Data
Sort news by: Recently Popular | Top Today | Yesterday | Week | Month | Year |

Published Data

MoU - National Science Museum

1
Voted

Posted by admin 6 days ago
URI: http://localhost/aptana/edmsv1/story/rdf/id/72
Concept: MoU-MoA (created by admin)

Discard | Edit | Admin: Change the status. | Delete | Manage-admin
See all MoU-MoA / Post MoU-MoA

Division:	Education
Country:	Thailand
Partner:	National Science Museum (NSM)
Type:	General
Effective_Date:	05 February, 2009
Expiry_Date:	05 February, 2014
Status:	Active
Contact_Person:	Prof. Joydeep Dutta, Associate Dean-SERD and Director, Center of Excellence in Nanotechnology, AIT, Dr. Aphiya Hathayatham, Director, Information Technology Museum
Responsible_Person:	Wannapa Pliansri
Link:	Museum.pdf
Category:	Documents
Tags:	mou thailand

*establishment of the Science Communication Allianceand be founding members of the alliance
*carrying out activities related to Science Communication program that would include, but not limited to: - joint workshops, seminars, conferences, science camps, and exhibitions - graduate degree program in Science Communic

Read More | Discuss | Tell a Friend |

SORT NEWS

● Recently Popular
○ Top Today
○ Yesterday
○ Week
○ Month
○ Year

WHAT IS DMS?

A document management system (DMS) is a set of computer programs used to structure and store electronic documents and/or images of paper documents.
Read More..

CONCEPT CLOUD

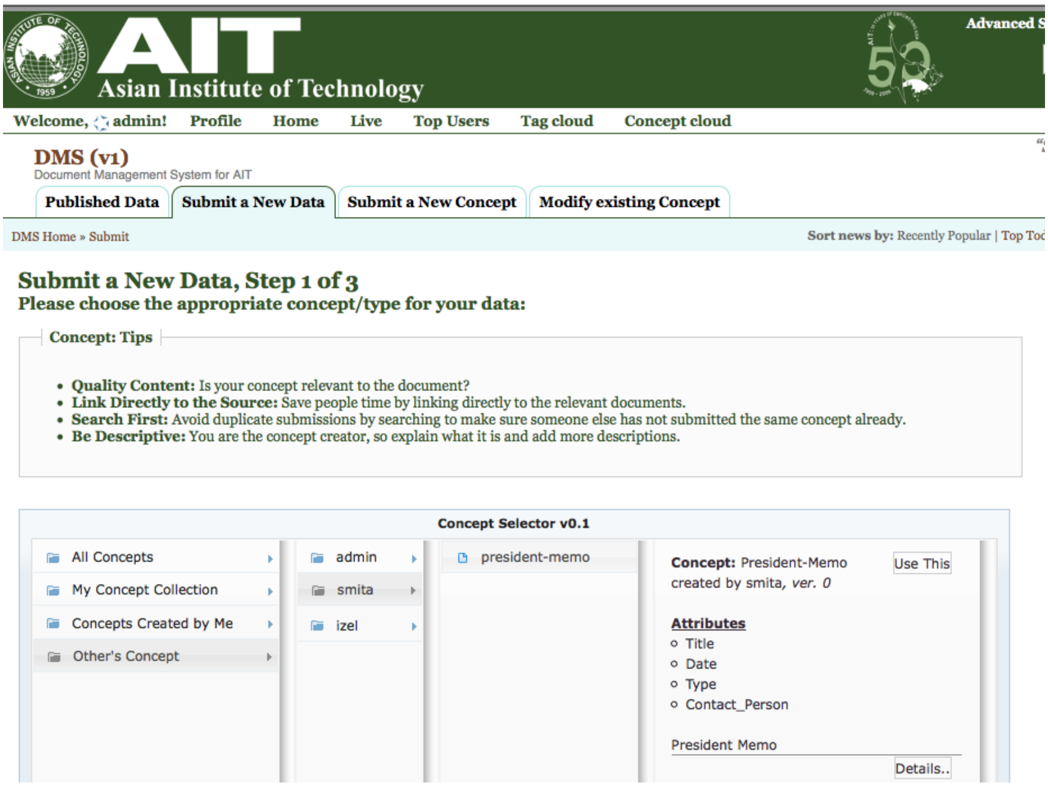
mou-moa
Read More

TOP USERS
TAG CLOUD
LIVE
GTALK
INBOX

TOP TAGS

agreement australia azerbaijan bhutan canada china ethiopia finland france india indonesia international organization iran japan moa mou myanmar pakistan philippines phillipines srilanka taiwan thailand united states vietnam
Read More

Figure 65. A screenshot of the DMS system at AIT.



Submit a New Data, Step 1 of 3
Please choose the appropriate concept/type for your data:

Concept: Tips

- **Quality Content:** Is your concept relevant to the document?
- **Link Directly to the Source:** Save people time by linking directly to the relevant documents.
- **Search First:** Avoid duplicate submissions by searching to make sure someone else has not submitted the same concept already.
- **Be Descriptive:** You are the concept creator, so explain what it is and add more descriptions.

Concept Selector v0.1

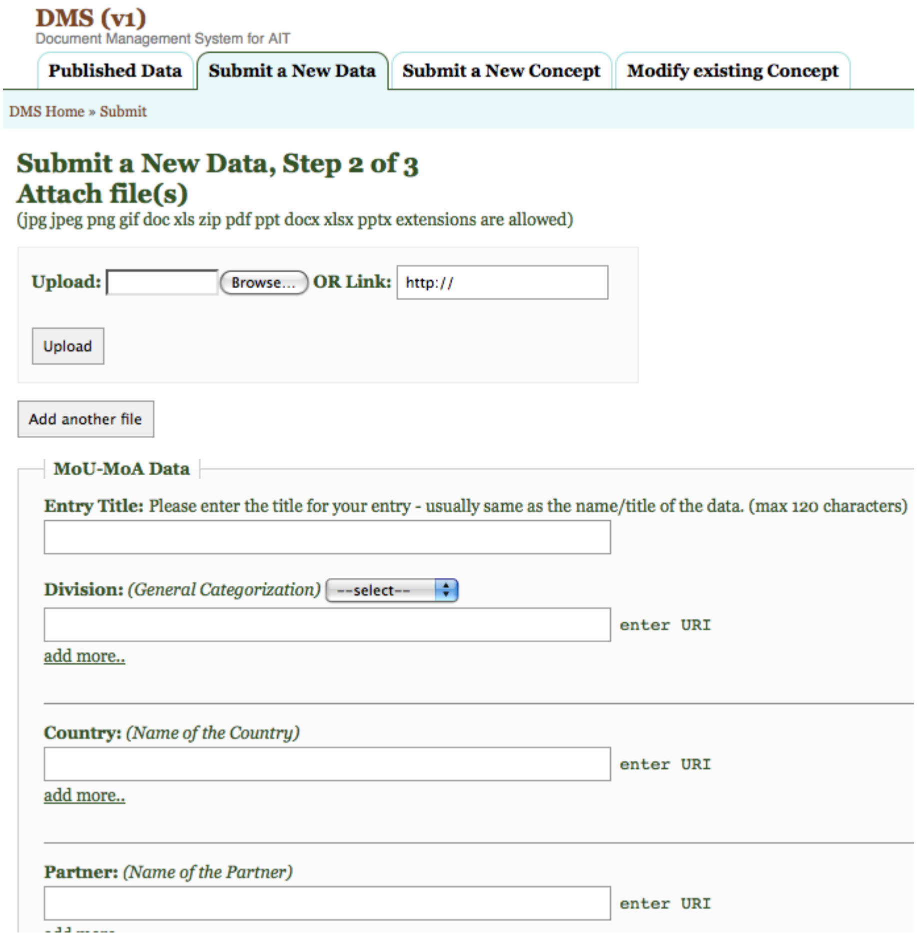
Concept: President-Memo
created by smita, *ver. 0*

Attributes
- Title
- Date
- Type
- Contact_Person

President Memo

Figure 66. The concept explorer/selector interface.

A new interface for selecting and exploring concepts, as shown in Figure 66, was created which was suitable for the document management system. The structured data input interface is also slightly improved, as shown in Figure 67, for better supporting document uploads. An access control mechanism has also been implemented in the system so that access of a document can be restricted to selected users.


DMS (v1)
Document Management System for AIT

Submit a New Data, Step 2 of 3
Attach file(s)
(jpg jpeg png gif doc xls zip pdf ppt docx xlsx pptx extensions are allowed)

Upload: [ ] Browse... OR Link: http://

MoU-MoA Data

Entry Title: Please enter the title for your entry - usually same as the name/title of the data. (max 120 characters)

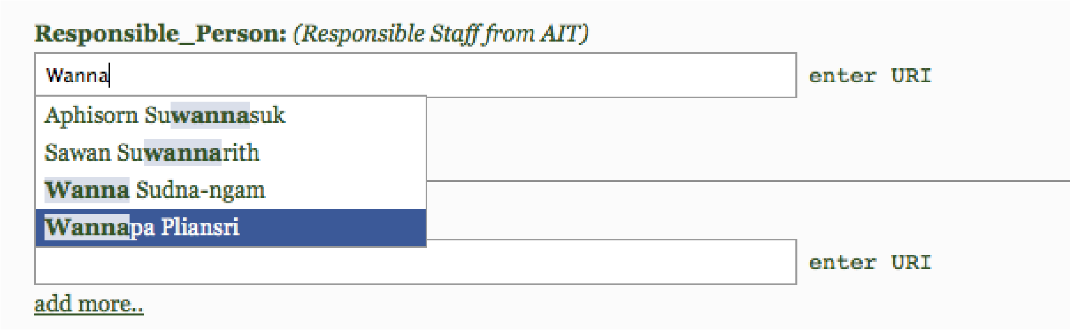
Division: *(General Categorization)* --select--
enter URI
add more..

Country: *(Name of the Country)*
enter URI
add more..

Partner: *(Name of the Partner)*
enter URI

Figure 67. Structured data input interface for the DMS.

An auto-complete interface has also been implemented to link the document entry to the staff of AIT who is responsible for the document (see Figure 68). Convenient widgets for selecting country names and dates have also been implemented as shown in Figure 69 and Figure 70. Besides these, an auto-complete mechanism for the advanced search is also being implemented.


Responsible_Person: *(Responsible Staff from AIT)*

Wanna

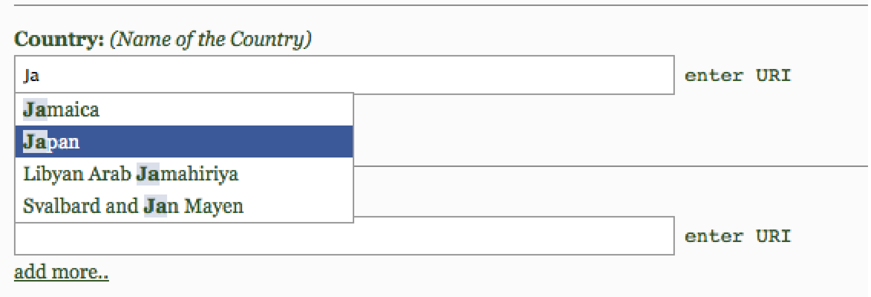
enter URI

Aphisorn Suwannasuk
Sawan Suwannarith
Wanna Sudna-ngam
Wannapa Pliansri

enter URI

add more..

Figure 68. Auto-complete to select the staff.



Country: *(Name of the Country)*

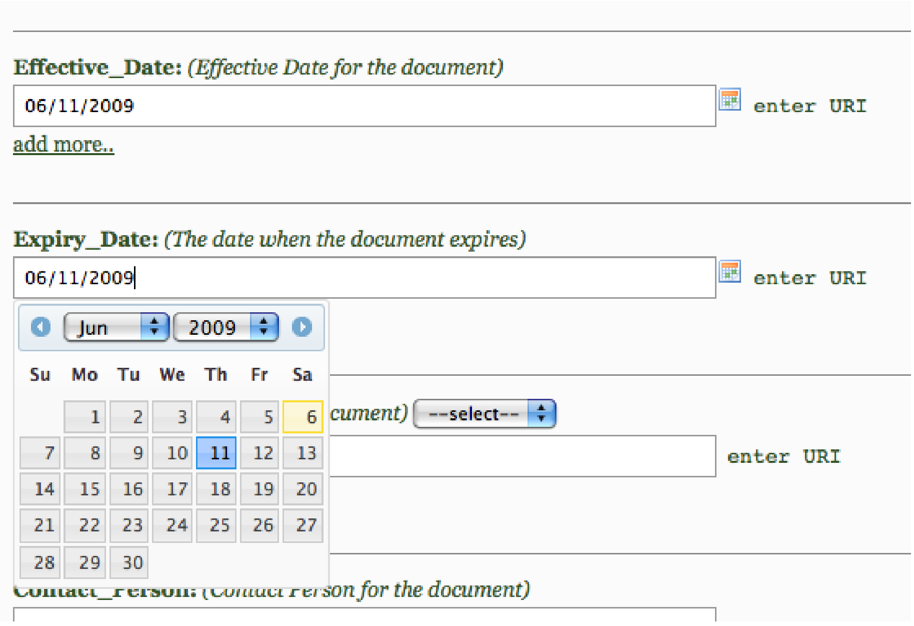
Ja — enter URI

- Jamaica
- Japan
- Libyan Arab Jamahiriya
- Svalbard and Jan Mayen

enter URI

add more..

Figure 69. Country selector widget.



Effective_Date: *(Effective Date for the document)*

06/11/2009 — enter URI

add more..

Expiry_Date: *(The date when the document expires)*

06/11/2009 — enter URI

Jun 2009

Su	Mo	Tu	We	Th	Fr	Sa
	1	2	3	4	5	6
7	8	9	10	11	12	13
14	15	16	17	18	19	20
21	22	23	24	25	26	27
28	29	30				

cument) --select--

enter URI

Contact_Person: (Contact Person for the document)

Figure 70. Date selector widget.

5.6.5 OntoBlog

The OntoBlog[91] semantic blogging prototype demonstrates a possible application of ontology and structured data for organizing and utilizing unstructured social data. Traditionally, blog entries are scattered snippets of text and it is difficult to navigate, organize and retrieve the contents. OntoBlog links up unstructured blog entries to well structured ontology instances through semi-automatic semantic annotation. This enables effective navigation, organization and retrieval of unstructured contents. In organizations which maintain knowledge bases, a lot of high quality information may remain locked because direct use of such knowledge bases is difficult for non-technical users. Coupling a knowledge base with informal techniques like blogging can expose such valuable data for useful applications. Thus, semantic contents can be used in social platforms like blogs for informal knowledge management.

[91] An online demo can be found at http://dutar.ex.nii.ac.jp/ontoblog/blog/default/

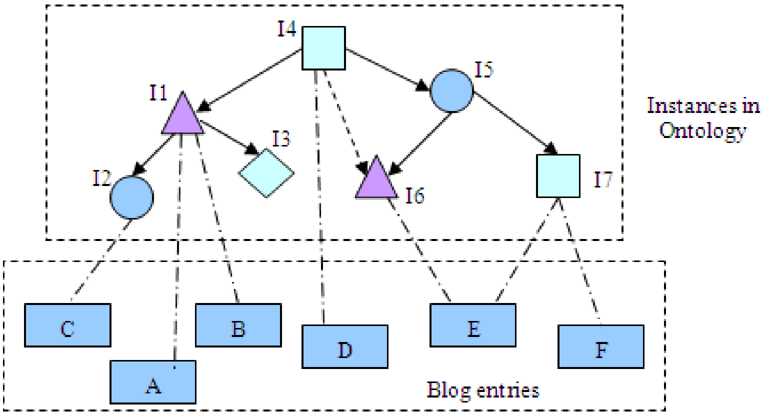


Figure 71. Example semantic annotation of blog entries.

Semantic annotation of blog entries allows us to relate different blog entries using the structure of the ontology as illustrated in Figure 71. In the figure, instances (I1-I7) in the ontology are represented by different shapes, each shape representing a concept. Instances are connected to each other by different relations (indicated by the solid arrows). Linking blog entries to ontology helps in linking related blog entries implicitly. Blog entries (A to F) are annotated by the ontology by linking them to the instances, as shown by the dash-dotted lines. Blog entries ‘A’ and ‘B’ are related to each other because they are both mapped to the same instance ‘I1’. Instance ‘I1’ is related to ‘I2’. Hence, blog entry ‘A’ is indirectly related to ‘C’, which has been mapped to ‘I2’. Instances may also be linked by implicit relations (shown by dashed arrow) that can be discovered by inference. Instance ‘I4’ is related to ‘I6’ by an inferred link. Thus, blog entry ‘D’ (mapped to ‘I4’) is related to ‘E’ (mapped to ‘I6’).

An example application

As an example domain, we can consider the case of a computer department of a university. The department maintains an ontology with concepts like course, topic, teacher, research, etc. The knowledge base including the ontology and instances is maintained by the ontology engineer or administrator. The department also maintains a community blog as illustrated in Figure 72. When the users publish or update a blog entry, the system automatically suggests instances related to the blog entry. If a related instance or concept is not shown by the system, the user may enter appropriate suggestion for a new instance and/or concept. The suggestions can be viewed by the administrator who can make appropriate additions or improvements to the knowledge base. Then, the users can access the blog entries effectively with the help of various semantic capabilities provided.

A prototype implementation of this scenario has been demonstrated considering the Computer Science and Information Management department of the Asian Institute of Technology, Thailand.

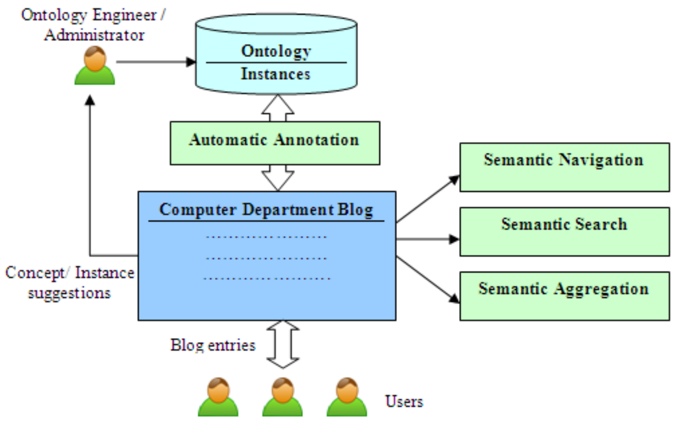


Figure 72. Example scenario for OntoBlog.

Example Ontology. A simple example ontology of the computer science department is shown in Figure 73. It is adapted from the SHOE Computer Department Ontology[92]. However, only few concepts and relations of the ontology have been used. Some inference is also possible with axioms like - "*for_course* and *has_topic* are inverse relations", "*is_broader_than* and *is_narrower_than* are inverse relations", "*teaches* and *taught_by* are inverse relations", "*has_prerequisite* and *is_broader_than* are transitive", etc.

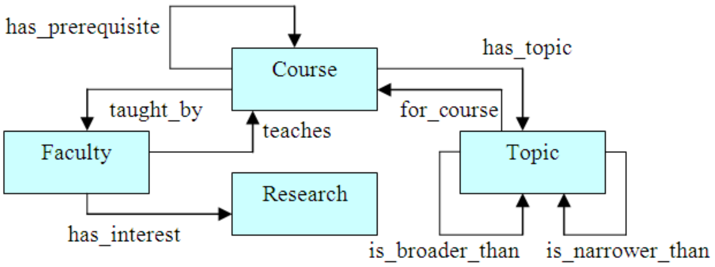


Figure 73. A part of a computer department ontology.

Useful services

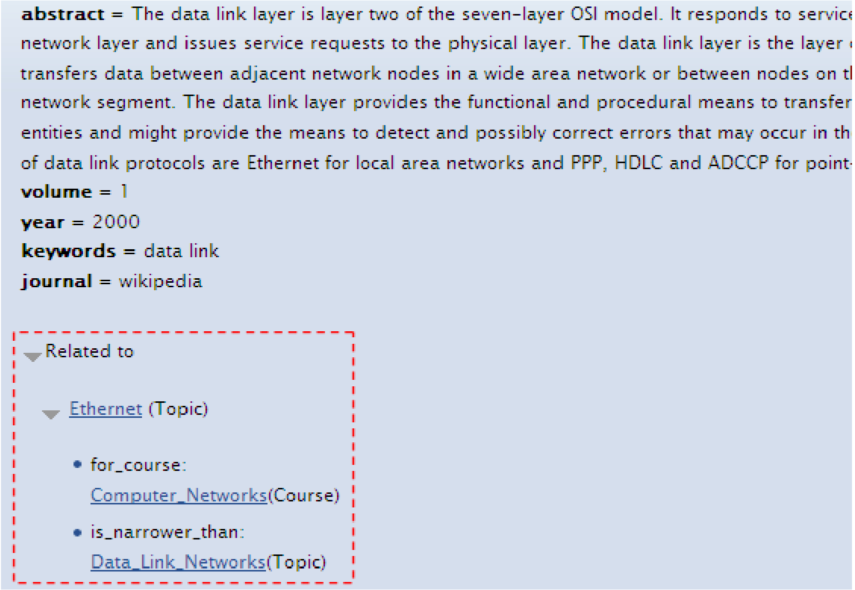
Some useful services demonstrated by the system are as follows. Such services motivate the user by providing instant gratification in return to their contribution as demonstrated in Mangrove (McDowell et al., 2003).

Semantic navigation. Semantic navigation helps the user to browse through related blog entries. For example, suppose we view a blog entry B about "Database Programming". The blog entry may be connected to {"computer programming", "databases", "software development", "Prof. Takeda"....}. "Computer programming" may be involved in the relations {"is taught by", "has prerequisite",..... }. Thus, there may be links like

[computer programming]
– is taught by – [Prof. Takeda]
– has prerequisite – [databases],etc.

[92] http://www.cs.umd.edu/projects/plus/SHOE/onts/cs1.1.html

Clicking on [databases] will lead to the blog entries related to databases. When a blog entry is opened, the semantic navigation links are shown in a collapsible "Related to" block (shown in Figure 74).



abstract = The data link layer is layer two of the seven-layer OSI model. It responds to servic
network layer and issues service requests to the physical layer. The data link layer is the layer
transfers data between adjacent network nodes in a wide area network or between nodes on t
network segment. The data link layer provides the functional and procedural means to transfer
entities and might provide the means to detect and possibly correct errors that may occur in th
of data link protocols are Ethernet for local area networks and PPP, HDLC and ADCCP for point
volume = 1
year = 2000
keywords = data link
journal = wikipedia

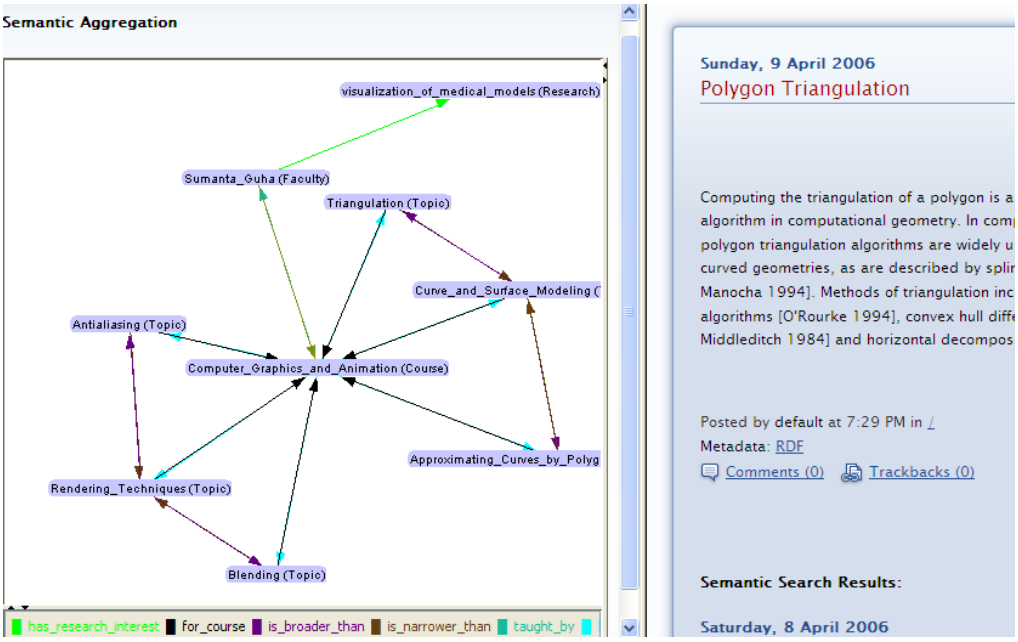
Related to

Ethernet (Topic)

- for_course: Computer_Networks(Course)
- is_narrower_than: Data_Link_Networks(Topic)

Figure 74. Semantic navigation.

Semantic search. The text search in the system is augmented by semantic search. It traces semantic links of the ontology to retrieve related contents. Guha et al. (2003) have presented extensive research on semantic search along with sophisticated implementation. OntoBlog just provides a simple demonstration of its applicability. The depth of semantic links followed by the semantic search may also be configured.


Semantic Aggregation
visualization_of_medical_models (Research)
Sumanta_Guha (Faculty)
Triangulation (Topic)
Curve_and_Surface_Modeling (
Antialiasing (Topic)
Computer_Graphics_and_Animation (Course)
Approximating_Curves_by_Polyg
Rendering_Techniques (Topic)
Blending (Topic)
🟩 has_research_interest ■ for_course ■ is_broader_than ■ is_narrower_than ■ taught_by

Sunday, 9 April 2006
Polygon Triangulation

Computing the triangulation of a polygon is a
algorithm in computational geometry. In com
polygon triangulation algorithms are widely u
curved geometries, as are described by splir
Manocha 1994]. Methods of triangulation inc
algorithms [O'Rourke 1994], convex hull diffe
Middleditch 1984] and horizontal decompos

Posted by default at 7:29 PM in /
Metadata: RDF
Comments (0) Trackbacks (0)

Semantic Search Results:

Saturday, 8 April 2006

Figure 75. Semantic aggregation.

Semantic aggregation. Semantic aggregation can be introduced in the system to collect and organize search results relevant to a topic of interest. Semantic aggregation is depicted in Figure 75. The user runs semantic aggregation by searching on a topic of interest. The search results are listed on the right-hand side frame. Related instances from the ontology are aggregated and visualized on the left-hand side frame as directed graphs. The nodes represent concept instances and the links represent the relations between them. The relation type is identified by the color of the link and shown in an index. When a node is clicked, blog entries related to that node are displayed on the right-hand side.

OntoBlog has been built upon the Blojsom blogging platform using Java, just as SocioBiblog. Similarly, the same Jena Semantic Web framework backed by a MySQL database has been used to deal with RDF data. Protégé[93] was used for creating the example ontology. GraphML[94] along with the Prefuse package[95] has been used for the graphical visualization of interlinked data.

[93] http://protege.stanford.edu/
[94] http://graphml.graphdrawing.org/
[95] http://prefuse.sourceforge.net/

5.7 Comparison with Existing Works

Looking back at the state-of-art of structured data creation in the social Semantic Web, as discussed in Section 2.4, we can say that the proposed approach advances it by overcoming some limitations of the existing approaches. StYLiD provides an easy platform for structured data sharing just like the existing approaches for direct structured instance data creation, for e.g., semantic blogging, semantic bookmarking, etc. However, these systems can produce only limited types of instance data and the concepts and ontologies do not evolve. StYLiD advances such approaches towards producing evolving concepts and ontologies too. If we ignore the capability of StYLiD to create new concepts, it functions quite like a semantic blogging or semantic bookmarking platform.

StYLiD addresses some specific issues in existing works for collaborative creation of structured resources and ontologies. The following Table 15 shows the comparison among some related works in collaborative knowledge base creation, revisiting the analysis summarized in Table 1. The prominent representative works are the semantic wikis, Freebase, myOntology and the ontology maturing approach (Braun et al., 2007). Freebase is the closest system to StYLiD considering the functionalities. The systems can be compared along several dimensions.

1. Ease of use. First considering the usability, the detailed experiments indicate that StYLiD is more usable than Freebase for the features considered in the experiments. StYLiD was found to be easier to use for posting structured data and creating concept schemas. It should be reiterated that this does not mean at all that StYLiD is better than Freebase overall. This cannot be taken as an evaluation of Freebase. Freebase has many attractive capabilities that are not present in our approach. In Freebase, the same topic instance may have multiple concept types. Data may be imported in bulk. Data is organized into spaces called bases. A lot of data is already populated in Freebase and users can simply build collections by creating views. The ways of exploring and searching data are also different in the systems. The experiments indicate that searching is still difficult in both the systems. Freebase has more functionalities and much elaborate interface than StYLiD. On the other hand, this may rather overwhelm or confuse the user unlike the focused interface of StYLiD.

More powerful and expressive systems tend to be more complex. The Semantic MediaWiki requires some training to get started because users need to learn the extended wiki syntax. Even with interface enhancements (Pfisterer et al., 2008) it still seems to have considerable learning curve for ordinary people and lower SUS ratings. For myOntology users need to have some understanding of ontologies. Some orientation and training would be required to explain the notions to the users. Moreover, it may be difficult to motivate people to directly help in the ontology construction process as there does not seem to be any obvious benefit for them.

2. Expressiveness. Next, if we consider the expressiveness, StYLiD is moderately expressive. The concepts serve as class/property frame definitions. The concepts act as the common vocabulary to share structured data and grouping of semantically similar concepts serves as a thesaurus. However, the lightweight ontologies that emerge are informal. Semantic Wikis provide more rigorous semantics. However, they are mostly meant for creating semantically structured instances. Semantic wikis, like the Semantic MediaWiki(SMW), are mainly used for annotating wiki pages with metadata. It is also possible to create schemas in SMW in the form of templates.

However, it is generally limited to the administrator and templates have to be created using a custom markup language. Both StYLiD and Freebase basically express concept schemas and structured instances. Freebase is somewhat more expressive because range restrictions and other constraints are used.

3. Constraints. Most of the systems including the semantic wikis, Freebase and myOntology impose strict constraints on the data. StYLiD does not impose many constraints and keeps the range specifications only suggestive. This makes it easy to input any type of data, even incomplete and unforeseen. The flexible and relaxed interface of StYLiD offers more freedom to the users. But at the same time the chances of erroneous input also increases.

5. Multiplicity. Allowing multiple conceptualizations and consolidating them at the same time is a unique aspect of our approach.

6. Consensus. Having consensus over the conceptualizations is also optional in StYLiD unlike other approaches. In Freebase, the schemas in individual user spaces do not require consensus. However, when the schemas are in the common shared space, general consensus is assumed.

StYLiD also has some other features that are not present in other collaborative systems.

- In Freebase, it is difficult to link to external resources. This can be done easily in StYLiD and any arbitrary resources can be linked using the URI.
- StYLiD directly produces data in linked data format at the time of authoring. However, systems like Freebase and SMW convert the data later into linked data format.
- Also it is not possible to specify multiple concepts as property value range in Freebase.
- The listed collaborative systems do not provide concept grouping by similarity.
- The systems also do not embed structured data as in StYLiD which uses RDFa.

The Table 16 shows a rough comparison of some features among StYLiD and the two most prominent related works, Freebase and the Semantic MediaWiki (SMW). Concept and instance creation is UI supported with form-based interface in both StYLiD and Freebase. The SMW uses a template markup language for concept schema creation. It uses extended wiki syntax for instance data, along with a forms extension developed later. In StYLiD, data authoring is more like blogging or social data bookmarking. Freebase is a structured wiki for collaborative maintenance of data. In the SMW, it is basically done as semantic annotation of the wiki text. External wrappers have to be used in StYLiD to import data. A bulk data import facility is provided by Freebase. Bulk imports are not directly possible in SMW. Constraints are flexible in StYLiD unlike the other systems and multiplicity of concept definitions is allowed. Consolidation is done at schema-level in StYLiD. Freebase provides some instance level consolidation but not at the schema-level. Concepts can be organized by semi-automatic grouping in StYLiD. Bases are created in freebase to put together related items. The SMW simply uses conventional categories for organizing the resource pages.

Table 15. Comparision with existing works.

	Ease of use	**Expressiveness**	**Constraints**	**Multiplicity**	**Consensus**
Semantic Wikis	*Complex* - extended wiki syntax - some training needed	*Moderate* - Mainly instances, concept schemas possible	*strict type constraints*	*No*	*Needed* - Wiki way
Freebase	*Moderate* - Interactive but elaborate interface	*Moderate* - Concept schemas, instances	*strict type constraints*	Allowed but concepts not related	*Mostly needed* - Wiki way - selected by admin
myOntology	*Complex* - understanding of ontology needed	*Moderate* - Concepts, relations, instances	*Strict logical constraints*	*No*	*Needed* - Wiki way
Ontology maturing approach	*Fairly easy* - need to build taxonomy	*Low* - Concept hierarchy	*free tagging*	*No*	*Needed* - By interaction
StYLiD	*Easy* - easier than Freebase	*Moderate*	*Minimum*	*Yes*	*Optional*

Table 16. Comparison of some features with Freebase and SMW.

	StYLiD	**Freebase**	**Semantic MediaWiki**
Concept creation	UI supported	UI supported	Template markup
Instance creation	Form-based	Form-based	Extended wiki syntax+ forms
Data authoring	Blogging / social bookmarking	Structured wiki	Wiki text annotation
Data import	Wrappers	Bulk import facility	Not possible
Constraints	Flexible	Strict type constraints	Strict type constraints
Multiplicity	Allowed	Partly	No
Consolidation	Schema-level	Some instances	No
Organization	Concept grouping	Bases	Categories

Other approaches for multiple conceptualizations. There are a number of works based on the idea of multiple conceptualizations and combining them, introduced in Section 3.2.1. But these are currently outside the scope of the social Semantic Web and mainly in the area of formal ontology engineering. The thesis attempts to introduce these ideas into the area of collaborative creation of ontology and structured resources in the social Semantic Web. As such, the thesis is not intended towards advancing the state-of-art of these works and hence, there is no point in direct comparison. Nevertheless, the proposed approach in the thesis can be related to these works as discussed below.

Approaches like DDL (Borgida & Serafini, 2003), C-OWL (Bouquet et l., 2004), ε-connections (Kutz et al., 2004; Grau et al., 2004) provide formalisms to represent ontologies in multiple contexts and to relate or connect them. Such representations can in fact be adopted in our approach. These basically express the alignments in formal logic. The schema alignments in our approach can also be exported in C-OWL format. However, these approaches do not deal with the issue of building up conceptualizations. Takeda et al. (1995) also provided a logical representation mechanism to represent heterogeneous ontologies from multiple aspects. Such multiple aspects are then mapped to enable multi-agent communication. Similar information exchange and interoperation among different parties maintaining multiple conceptualizations is also possible in our approach through query translations.

The works based on DOGMA formal ontology engineering approach (Meersman, 1999; Jarrar & Meersman, 2002; Jarrar & Meersman, 2008; De Leenheer & Debruyne, 2008; De Leenheer et al., 2009) distinguish multiple application-specific axiomatizations while maintaining a common generalized domain-specific ontology base, along with context representation. Ontological elements are modeled elegantly keeping them reusable across multiple application perspectives. Reconciling multiple perspectives is mainly based on generalization of the different cases and not on creating mappings between the different sources. Such formal ontology engineering process usually requires considerable effort by ontology engineers, although some tool support is provided. Moreover, it would be difficult to involve the community because these things are difficult to understand for non-technical people.

The eCOIN approach (Firat et al., 2007; Firat et al., 2005) assumes the existence of a generic ontology. The generic ontological terms are then specialized to multiple perspectives and mappings among them are enabled through a conversion function network. In fact, it would be useful to incorporate such a conversion function network into our approach too. However, the approach does not offer a mechanism to build up such generalized ontologies.

5.8 Discussion

The primary goal of the thesis is to enable people to share various types of information effectively. Social web applications are effective platforms to share information in communities. They are easy to use for ordinary people and facilitate free contribution. However, semantic structure is also important for effective sharing of information. Therefore, the thesis adopts the path of combining social web framework and Semantic Web technologies. The thesis proposes a new approach to address some specific issues in this area. Working systems have been implemented based on the approach and some experimental evaluations have also been done. Overall, the thesis can address some current needs of people as discussed in the beginning in Section 1.2 while facilitating effective information sharing. This demonstrates the practical significance of using structure and semantics in data.

1. Effective processing and retrieval. The proposed approach produces structured data which can be sorted and filtered by different fields and even exported for further processing. Flexible and powerful retrieval is possible through structured queries. Furthermore, categorization of data under concepts and grouping by similarity also helps in easy retrieval and browsing.

2. Automation and useful applications. The implemented system based on the proposed approach is already being used for some real practical applications in various scenarios. Other useful applications are also possible in future.

3. Interoperability. The ontologies formed by the consolidation process provide the basis for interoperability among different sources. The schema mappings make the interoperation among heterogeneous sources possible, which enable instance and query translation among the sources. As the emerging ontology gradually stabilizes, it can serve as the standard for interoperability among external systems too. The thesis also proposed how structured information can be disseminated across multiple interoperating systems and compatibility is maintained with existing systems.

4. Integration. This has been partly demonstrated by the application on integration of heterogeneous staff directories of the Japanese universities. As a result, the multiple sources can be searched together like a single repository. Integration of data from external resources like DBpedia has also been demonstrated.

The proposed approach can provide some technical support to facilitate the process of knowledge emergence in organizations or communities as proposed by Nonaka and Takeuchi (1995). They have modeled knowledge emergence in four continuous modes of conversion as follows.

1. Socialization. Socialization helps in sharing tacit knowledge in the community. The proposed social framework can support such a process by enabling people to share data of their interest and interact online.

2. Externalization. Externalization is the process of articulating tacit knowledge into explicit knowledge. The capability of conceptualizing various types of information and publishing online can effectively support this process.

3. Combination. Explicit knowledge from various sources can be combined and reconciled to form more complete knowledge. The proposed concept consolidation mechanism can help in consolidating multiple conceptualizations by different people and relate different perspectives to form a unified view.

4. Internalization. Internalization is the process of absorbing the explicit knowledge to increase tacit knowledge. The proposed approach facilitates this by providing ways to utilize the information effectively. The community gains knowledge from the collection of conceptually organized and shared information.

Knowledge in the organization or community grows in a continuous spiral through these modes while knowledge transforms from tacit to explicit and vice-versa.

5.8.1 Strengths

The proposed approach has already been compared with existing approaches in the previous section. The main strengths and salient aspects of the approach are as follows.

Social information sharing

- It provides an easy social platform enabling people to share information. Experiments showed that the implementation is quite usable even for ordinary people and needs minimum training.
- Unlike many social semantic applications which are meant for specific types of data, the proposed approach facilitates sharing of a wide variety of data. Different concepts of interest can be defined by the people themselves. Also the concepts obtained from ordinary people are likely to be the basic level of human conceptualization which has special advantages over concepts defined by experts.
- The flexible interface poses minimal constraints encouraging freedom in contribution, easing the process of concept definition and keeping it open for evolving needs.

Consolidating multiple conceptualizations

- Multiple conceptualizations of the same thing are allowed depending upon the perspective or context. Individual requirements and perspectives can be maintained in a democratic way. At the same time, a unified global model is also formed by consolidating these conceptualizations.
- As individuals in the community can contribute partial conceptualizations from their own perspective, the system facilitates a loose collaboration among the contributors requiring minimal interaction. Global consensus is not required over any conceptualization and it is not even necessary for people to fully understand each others' definitions for contributing. The semantics gradually becomes apparent with data instances.
- Unlike many ontology based approaches, no shared ontology is needed beforehand. The ontology emerges as a by-product of the information sharing activity. In this approach, community participation is not directly for the purpose of ontology creation. People basically participate to share information without being bothered about any ontology creation process.

Others

- As individual conceptualizations can be maintained and mapped, it is easy to connect legacy systems without changing the existing technologies.
- Major processes in the system, including schema alignment and concept grouping, are partly automatic. This eases the burden of the user. Also instead of implementing from scratch, the alignment task has been delegated to an existing ontology alignment module which may also be replaced by more sophisticated implementations in the future.
- Structured query unfolding is easy in the approach because it is based on GaV.
- The structured data is exposed on the Semantic Web following linked data principles. This makes the data instantly available to other applications which can reuse or link to the data and provide useful services. The linked data network effectively adds value to the exposed data. The system facilitates direct authoring of linked data, unlike most existing works which export the data from a system as linked data only in a later stage.
- The system also provides embedded RDFa data.
- The approach can serve a wide range of application scenarios from social information sharing to information integration from heterogeneous online sources, as demonstrated.

5.8.2 Limitations

The proposed approach definitely is not meant to solve all issues. It also has a number of limitations and weaknesses. Overcoming some of these limitations is considered for future work in the next chapter.

- The concept definitions by ordinary people may not be elegant theoretically. There may be some duplication of concepts and redundancies in definitions. Also while using the concepts defined by others, there may be some misunderstandings and different use from what had been intended.
- Another limitation is that currently an instance can only be of one concept type. However, there may be cases when the same thing may be of multiple concept types. Allowing the use of multiple concept types, like in Freebase, is a solution. But, on the other hand, this makes the system somewhat complex.
- Allowing some noisy data is inherent in the approach. Data entered by people may not be perfect and no restrictions are in place. So we have to rely on people to input sensible data and tolerate some rough data. Some cleaning and preprocessing may be necessary for applications requiring good quality data.
- Although the approach shows several ways for motivating users, it does not guarantee user motivation as such. It largely depends upon the particular application and so the responsibility is delegated to the application.
- The approach assumes that multiple conceptualizations can generally be consolidated and the schemas can be aligned with simple relations. However, the alignment may be quite complex in some cases due to different modeling.

Currently, such parts, for which alignment is not so easy, are left in the source conceptualizations without aligning.

- The system only performs the alignment at a generic level. However, specific nuances of the attributes may vary by local contexts. Currently, handling such details is delegated to the application. A conversion function network as proposed by Firat et al. (2007) may be needed to handle complex conversions.
- The concept grouping simply indicates the semantic proximity of the grouped concepts. It cannot determine the actual relation among the concepts.
- The ontologies that emerge are quite informal and lightweight. As such, logical inferences cannot be made directly. Knowledge engineers would be required to formalize these ontologies if needed.
- Currently, the approach does not explicitly support the reuse of existing external ontologies. It is rather focused on creating new vocabulary for the long tail of information domains.

6. Conclusions and Future Directions

6.1 Conclusions

The basic motivation of the thesis is effective information sharing on the web. This includes information publishing, information semantics and information dissemination as the key aspects. The significance of structured data and Semantic Web for representing the semantics of information to be shared was discussed. The role of social web applications for realizing information sharing in communities by enabling people to contribute, participate, collaborate and disseminate information was also discussed. Structured information creation and sharing by the combination of these two areas into a social Semantic Web is the subject of this thesis.

As mentioned in the introduction, the thesis mainly covered the following aspects in this field; identified some specific problems and proposed some new solutions. The proposed approaches were implemented into working systems serving as proof of concept. Several experiments and observations provide support for the approaches and also offer lessons to be learnt. The practical applicability of the proposed approach has also been demonstrated by some real world applications.

1. Obtaining structured data from people. The thesis proposes enabling ordinary people to author structured data for the Semantic Web by providing easy to use social web application interfaces. As extensions to existing social platforms, semantic blogging systems like SocioBiblog and OntoBlog which were implemented can facilitate easy publishing of particular types of data, for e.g., bibliographic data in case of SocioBiblog. However, it was soon realized that it is difficult to extend such systems for new types of data and even the existing types cannot evolve to accommodate requirements of people. Hence, a more flexible and generalized system, StYLiD, was implemented which enables people to share a wide variety of data of their interest by defining their own conceptual schemas. Keeping the input interface flexible and relaxed enables the users to contribute freely and easily. Freedom, ease of use and benefits are important factors for gaining social participation. Experiments showed that StYLiD is quite usable and almost requires no training to start contributing structured data. The lessons learned from the experiments can help in further refining the implementation to make it easier for people.

2. Collaborative ontology creation. To model the wide variety of data to be shared, new ontologies are required. Ontologies should be formed collaboratively to cover the requirements of different people. Some specific problems were identified in this area. Creating perfect concept definitions and building ontologies is a difficult process. It is difficult to achieve consensus on conceptualizations through direct collaboration. Therefore, following solutions were proposed.

- *Defining concepts freely*. People should be allowed to define their own concepts to meet their needs and concept definitions should not be rigid and constrained. Experimental evidences were also presented supporting that people can and do express conceptual schemas and that constrained concept definitions can create problems for data contribution.
- *Allowing multiple conceptualizations*. Multiple conceptualizations should be allowed because people have different perspectives over the same thing or

different contexts to be considered. Experimental evidences were also provided that different people have multiple conceptualizations

- *Consolidation of multiple conceptualizations.* Such conceptualizations can be consolidated to form a unified model. This is possible with data integration principles and semi-automatic schema alignment methods. Consolidation serves as a new collaborative approach for creation of conceptualizations from the community. It is a loose collaboration requiring minimal interaction and consensus and facilitates collaborative knowledge formation while satisfying individual requirements. It was experimentally observed that conceptualizations of the same thing by different people overlap significantly and can be consolidated. It was also verified that satisfactory precision and coverage can be achieved even with simple methods of schema alignment.
- *Emergence of informal lightweight ontologies.* Consolidation of concepts produces a unified common vocabulary for sharing different types of structured data. Concepts can further be grouped and organized semi-automatically. It was experimentally demonstrated that concept schemas can be grouped by similarity calculations with satisfactory precision and coverage. Concepts can evolve and emerge out of the cloud of concepts in the same manner as popular tags from a tag cloud.

3. Structured information dissemination. It is also important to have mechanisms for dissemination of the structured information in communities. Social web applications serve as excellent platforms for this by connecting people but are usually centralized and confine information within themselves. So a decentralized approach was proposed for dissemination of information across system boundaries. RSS feeds can easily be extended to transport structured data too. This was demonstrated by implementing SocioBiblog for sharing bibliographic information through social network links. It combines the capabilities of publishing and aggregating information into a single unit that can aggregate, filter and redistribute information. An evolving distributed network of such units can help in delivering relevant streams of information to people. The Semantic Web provides semantic structure and interoperability essential in such a decentralized environment. The proposed approach is applicable for any other system supporting RSS aggregation, including StYLiD.

The thesis demonstrates new ways by which various aspects of social software and Semantic Web technologies can be combined into a synergetic whole for information sharing. Mass contributions can be obtained from the community through social platforms providing abundant structured data for the social Semantic Web. However, having people contribute structured data is challenging and easy social interfaces need to be provided. Ontologies, needed for providing semantic structure to information, can emerge as a by-product of information sharing activities of the community and integration of heterogeneous information sources. It should be noted that effective information sharing is the main goal of the community rather than attempting to build ontologies directly. The emerging ontologies can form the basis for meaningful information sharing among disparate systems in the distributed web. Combination of social and semantic technologies can facilitate effective dissemination of information in the community using such systems.

Hence, the objectives set at the beginning of the thesis have mostly been addressed. However, there are several untouched areas, lessons learnt and more problems uncovered for future research. Some are discussed in the following section.

6.2 Future Directions

The presented work can be enhanced in various ways and several directions are still open for future research. Some are mentioned below.

1. *Computing concept relations*. Further work can be done on computing hierarchical and non-hierarchical relations between structured concepts besides just similarity relations. Subsumption relations between concepts can be determined to form concept hierarchies. Ideas from works on deriving ontologies from folksonomies may be adapted for the purpose.
2. *Better schema alignments*. More sophisticated alignment techniques may be employed to enable more complex alignments. Features for maintaining the alignments collaboratively may also be improved, following the wiki paradigm, rather than solely relying on sophisticated computations.
3. *Consolidation of data instances*. The presented work only considers consolidation of concepts. Consolidation of data instances is still an important open problem though some dataset-specific automated linking algorithms have been demonstrated (Bizer et al., 2007).
4. *Use of existing vocabularies and ontologies*. This work focused on the creation of new concept definitions by the community. However, ways to reuse existing vocabularies and ontologies or to map concept definitions to them should also be devised. The posts containing the structured data objects may also be weaved into the social linked data web using SIOC (Bojārs et al, 2008b). However, existing ontologies should be introduced carefully because ontologies are too complex for people to understand and use.
5. *Utilizing the structured data*. Plugins and mash-ups can be introduced, which may be contributed by the community itself, to make use of different types of structured data produced. These would provide instant benefits to people.
6. *Scraping web pages*. Scrapers may be provided to the users for collecting data from existing websites easily. Entering data forms manually still seems to be tedious. Visual interactive scraper creation tools may also be provided so that users can easily create and share such scrapers.
7. *Using existing Semantic Web data*. Besides scraping unstructured web pages, we may also directly use existing structured Semantic Web data. Data embedded in pages using RDFa or HTML5 may easily be picked up by the browser and fed into the system. If such formats become widespread in the future, this may be an easy and accurate way of picking up data from the web.
8. *Extended Evaluation*. The possible evaluation at this stage has been limited. The system is still in its initial stages of deployment and use. More evaluation can be done in the future after stable use of the system in real applications with real users for some period. The system is already being used in some applications. These can be monitored and lessons can be learnt through the experiences. The degree of user satisfaction or how well their requirements are met can be used as a measure for evaluation. Better empirical research and evaluation can be done with more users, from different backgrounds, for extended usability experiments. It may be interesting to discover statistically significant differences between users with or without IT background. Moreover, we should also study the nature of concepts contributed by the

users, how well these can be consolidated and how useful these consolidated concepts really are. We may also compare the emerging ontologies of consolidated and grouped concepts with some existing ontologies built traditionally.

9. Better searching and browsing interfaces could be developed to access and utilize the structured data and concepts. Traditional ranked keyword search mechanisms may be combined with structured semantic search. Faceted browsing also seems to be useful and popular.
10. Besides providing linked open data, embedded data and SPARQL interface, structured data may also be exposed through an API or extended RSS.
11. Recommendation systems may be used for disseminating useful information to targeted people. The recommendations may be based on the semantics of contents and social configuration of the person.
12. Issues of privacy, ownership, copyright and authenticity of shared data are also areas of high practical importance that are out of scope of this thesis.

References

Aberer, K., Cudré-Mauroux, P., Ouksel, A.M., Catarci, T., Hacid, M.S., Illarramendi, A., Kashyap, V., Mecella, M., Mena, E., Neuhold, E.J., Troyer, O.D., Risse, T., Scannapieco, M., Saltor, F., de Santis, L., Spaccapietra, S., Staab, S. & Studer, R (2004). Emergent Semantics Principles and Issues. In *Database Systems for Advanced Applications 9th International Conference, LNCS, 2973*, 25-38, Springer.

Ankolekar, A., Krötzsch, M., Tran, T. & Vrandečić, D. (2007). The two cultures: Mashing up web 2.0 and the Semantic Web. In *Proceedings of the 16th International World Wide Web Conference (WWW2007)*, Banff, Alberta, Canada, ACM Press, New York.

Auer, S., Bizer, C., Kobilarov, G., Lehmann, J., Cyganiak, R. & Ives, Z. (2007). Dbpedia: A nucleus for a web of open data. In K. Aberer, K.S. Choi, N. Noy, D. Allemang, K.I. Lee, L.J.B. Nixon, J. Golbeck, P. Mika, D. Maynard, G. Schreiber & P. Cudré-Mauroux, eds., *Proceedings of the 6th International Semantic Web Conference and 2nd Asian Semantic Web Conference (ISWC/ASWC2007), LNCS, 4825*, 715–728. Springer, Heidelberg.

Auer, S., Dietzold, S., Lehmann, J., Hellmann, S. & Aumueller, D. (2009). Triplify – Light-Weight Linked Data Publication from Relational Databases. In *Proceedings of WWW 2009*, Madrid, Spain.

Barabási, A.L. (2003). *Linked: How everything is connected to everything else and what it means for business, science, and everyday life*. Plume Books.

Baumgartner, R., Flesca, S. & Gottlob, G. (2001). Visual web information extraction with Lixto. In *Proceedings of the International Conference on Very Large Data Bases*, 119-128.

Bergman, M.K. (2007). What is the structured web? AI3 Blog. Retrieved August 10, 2007, from http://www.mkbergman.com/?p=390

Berners-Lee, T. (2000). Semantic Web - XML2000, slide 10. Retrieved April 22, 2009, from http://www.w3.org/2000/Talks/1206-xml2k-tbl/slide10-0.html

Berner-Lee, T., Hendler, J. & Lassila, O. (2001). The Semantic Web. *Scientific American*, *284*, 34-43.

Berners-Lee, T. (2006). Linked data. *World Wide Web design issues*. Retrieved December 24, 2008, from http://www.w3.org/DesignIssues/LinkedData.html

Bizer, C. & Cyganiak, R. (2006). D2R Server-Publishing Relational Databases on the Semantic Web. In *5th International Semantic Web Conference*.

Bizer, C., Cyganiak, R. & Heath, T. (2007a). How to publish linked data on the web. Retrieved December 24, 2008, from http://www4.wiwiss.fu-berlin.de/bizer/pub/LinkedDataTutorial/

Bizer, C., Cyganiak, R. & Gauss, T. (2007b). The RDF Book Mashup: from Web APIs to a web of data. In *Proceedings of the 3rd Workshop on Scripting for the Semantic Web, at 4th European Semantic Web Conference (ESWC 2007)*.

Blumauer, A. & Pellegrini, T. (2008). *Social Semantic Web: Web 2.0 - was nun?* Springer, Berlin.

Bojārs, U., Breslin, J. & Passant, A. (2008a). Data Portability with SIOC and FOAF. In *Proceedings of XTech 2008: The Web on the Move,* Dublin, Ireland.

Bojārs, U., Passant, A., Cyganiak, R., & Breslin, J. (2008b). Weaving SIOC into the web of linked data. In *Proceedings of the Workshop on Linked Data on the Web (LDOW2008).*

Bollacker, K., Tufts, P., Pierce, T. & Cook, R. (2007). A Platform for Scalable, Collaborative, Structured Information Integration. In *Sixth International Workshop on Information Integration on the Web.* Association for the Advancement of Artificial Intelligence.

Borgida, A. & Serafini, L. (2003). Distributed description logics: Assimilating information from peer sources. *Journal of Data Semantics*, *1,* 153-184.

Bouquet, P., Giunchiglia, F., van Harmelen, F., Serafini, L. & Stuckenschmidt, H. (2004). Contextualizing ontologies. *Web Semantics: Science, Services and Agents on the World Wide Web, 1*(4), 325-343. Elsevier.

Brank, J., Grobelnik, M., & Mladenić, D. (2005). A survey of ontology evaluation techniques. In *Proceedings of the Conference on Data Mining and Data Warehouses (SiKDD 2005).*

Bratt, S. (2007). Semantic Web, and Other Technologies to Watch, World Wide Web Consortium: January 2007, slide 24. Retrieved April 22, 2009, from http://www.w3.org/2007/Talks/0130-sb-W3CTechSemWeb/#(24)

Braun, S., Schmidt, A., Walter, A., Nagypal, G. & Zacharias, V. (2007). Ontology maturing: a collaborative web 2.0 approach to ontology engineering. In *Proceedings of the Workshop on Social and Collaborative Construction of Structured Knowledge at the 16th International World Wide Web Conference (WWW 07),* Banff, Canada, 8-12.

Breslin, J.G., Harth, A., Bojars, U. & Decker, S. (2005). Towards semantically-interlinked online communities. In *The Semantic Web: Research and Applications, vol. 3532 of Lecture Notes in Computer Science*, 500-514, Springer Berlin / Heidelberg.

Breslin, J. G., Passant, A., Decker, S. (in press). *The Social Semantic Web*. Springer.

Brooke, J. (1996). SUS: a "quick and dirty" usability scale. *Usability evaluation in industry*, 189-194.

Buffa, M., Gandon, F., Ereteo, G., Sander, P. & Faron, C. (2008). SweetWiki: A semantic wiki. *Journal of Web Semantics, 6*(1), 84-97.

Castano, S., De Antonellis, V., Fugini, M. & Pernici, B. (1998). Conceptual schema analysis: techniques and applications. *ACM Transactions on Database Systems (TODS), 23*(3), 286-333, ACM, New York.

Cayzer, S. (2004a). Semantic blogging and decentralized knowledge management. *Communications of the ACM, 47*(12), 48-52.

Cayzer, S. (2004b). Semantic blogging: Spreading the semantic web meme. In *Proceedings of XML Europe 2004*, Amsterdam, Netherlands, 18–21.

Cayzer, S. (2005). SWAD-Europe deliverable 12.1.4: Semantic blogging – lessons learnt. http://www.w3.org/2001/sw/Europe/reports/demo_1_report/.

Cayzer, S. (2006). What next for semantic blogging. In *Proceedings of the SEMANTICS 2006 conference*, 71-81.

Corcho, O., Fernández-López, M. & Gómez-Pérez, A. (2003). Methodologies, tools and languages for building ontologies. Where is their meeting point? *Data & Knowledge Engineering, 46*(1), 41-64.

Dataserver/reconciliation. (2008). In *Freebase*. Retrieved December 25, 2008, from http://www.freebase.com/view/guid/9202a8c04000641f8000000007beed56

De Leenheer, P. & Debruyne, C. (2008). DOGMA-MESS: A Tool for Fact-Oriented Collaborative Ontology Evolution. In R. Meersman, Z. Tari, & P. Herrero , eds., *OTM 2008 Workshops, LNCS: Vol. 5333*, 797–806, Springer-Verlag, Berlin Heidelberg.

De Leenheer, P., Christiaens, S., & Meersman, R. (2009). Business Semantics Management: a Case Study for Competency-centric HRM. *Journal of Computers in Industry: Special Issue about Semantic Web Computing in Industry*. Elsevier.

Decker, S. & Frank, M. (2004). The social semantic desktop. In *WWW2004 Workshop Application Design, Development and Implementation Issues in the Semantic Web, 9*.

Di Maio, P. (2008). Towards global user models for semantic technologies: emergent perspectives. In Ronchetti, M., ed., *Proceedings of the 1st Workshop on Human Factors and the Semantic Web, Bangkok, Thailand.*

Ding, Y., Embley, D.W. & Liddle, S.W. (2006). Automatic creation and simplified querying of semantic Web content: An approach based on information-extraction ontologies. In *Proceedings of the First Asian Semantic Web Conference (ASWC 2006), vol. 4185 of Lecture Notes in Computer Science*, 400-414, Springer.

Dzbor, M., Domingue, J. & Motta, E. (2003). Magpie - towards a semantic web browser. In *The Semantic Web - ISWC 2003 , vol. 2870 of Lecture Notes in Computer Science*, 690-705, Springer-Verlag Berlin Heidelberg.

Dzbor, M., Motta, E. & Domingue, J. (2004). Opening up magpie via semantic services. In *The Semantic Web - ISWC 2004 , vol. 3298 of Lecture Notes in Computer Science*, 635-649, Springer Berlin / Heidelberg.

Euzenat, J. (2004). An API for ontology alignment. In S.A. McIlraith, D. Plexousakis, & F. Van Harmelen, eds., *International Semantic Web Conference. LNCS, Vol. 3298,* 698–712, Springer.

Euzenat, J., Le Bach, T., Barasa, J. et al. (2004). State of the art on ontology alignment. *Knowledge Web Deliverable D2.2.3*.

Firat, A., Madnick, S. & Grosof, B. (2007). Contextual alignment of ontologies in the eCOIN semantic interoperability framework. *Information Technology and Management, 8*(1), 47-63, Springer.

Firat, A., Madnick, S. & Manola, F. (2005). Multi-dimensional ontology views via contexts in the ECOIN semantic interoperability framework. *Contexts and Ontologies: Theory, Practice and Applications*, 1-8, AAAI Press.

Goldberg, D. Nichols D., Oki, B. M. & Terry, D. (1992). Using collaborative filtering to weave an information tapestry. *Communications of the ACM , 35* (12), 61–70.

Gómez-Pérez, A. (2003). Ontology Evaluation. In S. Staab and R. Studer, eds., *Handbook on Ontologies,* 251-274, Springer.

Guarino, N. & Welty, C. A. (2000). A formal ontology of properties. *Knowledge Acquisition, Modeling and Management*, 97–112, Springer.

Guha, R., McCool, R. & Miller, E. (2003). Semantic search. In *Proceedings of the twelfth international conference on World Wide Web*, Budapest, Hungary, 700–709, ACM Press New York, USA.

Grau, B., Parsia, B. & Sirin, E. (2004). Working with multiple ontologies on the semantic web. In *The Semantic Web – ISWC 2004, LNCS: Vol. 3298*, 620-634, Springer, Berlin / Heidelberg.

Groza, T., Handschuh, S., Moeller, K., Grimnes, G., Sauermann, L., Minack, E., Mesnage, C., Jazayeri, M., Reif, G. & Gudjonsdottir, R. (2007). The NEPOMUK project - on the way to the social semantic desktop. In *Proceedings of I-Semantics, 7,* 201-211.

Gruber, T. (2007). Ontology of Folksonomy: A Mash-up of Apples and Oranges. *International Journal of Semantic Web and Information Systems, 3,* 1-11.

Gruber, T. (2008). Collective knowledge systems: Where the social web meets the semantic web. *Journal of Web Semantics, 6*(1), 4–13.

Gruber, T.R. (1993). A translation approach to portable ontology specifications. *Knowledge Acquisition, 5*, 199-220.

Guarino, N. (1998). Formal ontologies and information systems. In *Proceedings of FOIS, 98*, 3-15, IOS Press.

Haase, P., Schnizler, B., Broekstra, J., Ehrig, M., van Harmelen, F., Menken, M., Mika, P., Plechawski, M., Pyszlak, P., Siebes, R., Staab, S. & Tempich, C. (2004). Bibster - a semantics-based bibliographic peer-to-peer system. In *Proceedings of the International Semantic Web Conference (ISWC2004), vol. 3298 of Lecture Notes in Computer Science,* Hiroshima, Japan.

Hahn, U. & Romacker, M. (2001). The SynDiKATe text knowledge base generator. In *Proceedings of the first International Conference on Human Language Technology Research*, Morristown, NJ, USA, 1–6, Association for Computational Linguistics.

Halevy, A., Etzioni, O., Doan, A., Ives, Z., Madhavan, J., McDowell, L. & Tatarinov, I. (2003). Crossing the structure chasm. In *Proceedings of the First Biennial Conference on Innovative Data Systems Research (CIDR)*.

Halpin, H. (2009). A query-driven characterization of linked data. In *Proceedings of WWW 2009 Workshop: Linked Data on the Web (LDOW2009), Madrid, Spain*.

Handschuh, S., Staab, S. & Ciravegna, F. (2002). S-CREAM – semiautomatic creation of metadata. In *13th International Conference on Knowledge Engineering and Knowledge Management (EKAW02),* Madrid, Spain.

Hasan, T. & Jameson, A. (2008). Bridging the Motivation Gap for Individual Annotators: What Can We Learn From Photo Annotation Systems? In *Proceedings of the 1st Workshop on Incentives for the Semantic Web (INSEMTIVE 2008),* Karlsruhe, Germany.

Heath, T. & Motta, E. (2007). Revyu.com: A reviewing and rating site for the web of data. In K. Aberer, K.S. Choi, N.F. Noy, D. Allemang, K.I. Lee, L.J.B. Nixon, J. Golbeck, P. Mika, D. Maynard, R. Mizoguchi, G. Schreiber & P. Cudré-Mauroux, eds., *ISWC/ASWC, vol. 4825 of Lecture Notes in Computer Science*, 895-902, Springer.

Hendler, J. (2008). Web 3.0: Chicken farms on the semantic web. *Computer-IEEE Computer Society, 41*.

Hepp, M. (2007). Possible ontologies: How reality constrains the development of relevant ontologies. *IEEE Internet Computing, 11*(1), 90-96.

Horrocks, I., Parsia, B., Patel-Schneider, P. & Hendler, J. (2005). Semantic web architecture: Stack or two towers? *Lecture notes in computer science, 3703*.

Hotho, A., Jäschke, R., Schmitz, C. & Stumme, G. (2006). BibSonomy: A social bookmark and publication sharing system. In *Proceedings of the Conceptual Structures Tool Interoperability Workshop at the 14th International Conference on Conceptual Structures,* 87-102, Aalborg University Press.

Hughes, T.C. & Ashpole, B.C. (2004). The Semantics of Ontology Alignment. In *Proceedings of Information Interpretation and Integration Conference (I3CON), Performance Metrics for Intelligent Systems, PerMIS '04.*

Huynh, D., Karger, D. & Miller, R. (2007a). Exhibit: lightweight structured data publishing. In *Proceedings of the 16th international conference on World Wide Web*, 737-746, ACM Press New York, USA.

Huynh, D., Mazzocchi, S. & Karger, D. (2007b). Piggy bank: Experience the semantic web inside your web browser. *Web Semantics: Science, Services and Agents on the World Wide Web, 5*, 16-27.

Huynh, D.F., Miller, R.C. & Karger, D.R. (2007c). Potluck: Data mashup tool for casual users. In K. Aberer, K.S. Choi, N.F. Noy, D. Allemang, K.I. Lee, L.J.B. Nixon, J. Golbeck, P. Mika, D. Maynard, R. Mizoguchi, G. Schreiber & P. Cudré-Mauroux, eds., *ISWC/ASWC, vol. 4825 of Lecture Notes in Computer Science*, 239-252, Springer.

Iskold, A. (2007). The structured web - a primer. Read Write Web. Retrieved August 10, 2007, from http://www.readwriteweb.com/archives/structured_web_primer.php

Jarrar, M. & Meersman, R. (2002). Formal ontology engineering in the DOGMA approach. In R. Meersman, & Z. Tari , eds., *Proceedings of DOA/CoopIS/ODBASE 2002, LNCS: Vol. 2519*, 1238-1254, Springer-Verlag, Berlin Heidelberg.

Jarrar, M. & Meersman, R. (2008). Ontology Engineering-The DOGMA Approach. *Advances in Web Semantic 1,LNCS: Vol. 4891*, Springer.

Jhingran, A. (2008, June). Web 2.0, Enterprise 2.0 and Information Management. *Keynote speech at the Linked Data Planet conference & expo, spring 2008,* New York City.

Kahan, J., Koivunen, M.R., Prud'Hommeaux, E. & Swick, R.R. (2001). Annotea: an open rdf infrastructure for shared web annotations. In *Proceedings of the 10th International World Wide Web Conference*, 623-632, Hong Kong.

Karger, D.R. & Quan, D. (2005). What would it mean to blog on the semantic web? *Journal of Web Semantics, 3*(2), 147-157.

Kim, H., Breslin, J., Yang, S. & Kim, H. (2008). int. ere. st: building a tag sharing service with the SCOT ontology. In *Proceedings of the AAAI 2008 Spring Symposium on Social Information Processing,* Stanford University, California.

Kiryakov, A., Popov, B., Ognyanoff, D., Manov, D., Kirilov, A. & Goranov, M. (2003). Semantic annotation, indexing, and retrieval. In *2nd International Semantic Web Conference (ISWC 2003), vol. 2870 of Lecture Notes in Computer Science*, 484-499, Springer-Verlag Berlin Heidelberg.

Koivunen, M.R. (2005). Annotea and semantic web supported collaboration. In *ESWC 2005, UserSWeb workshop*.

Kruk, S. R., Decker, S. & Zieborak, L. (2005). JeromeDL - Adding Semantic Web technologies to digital libraries. In *Proceedings of Database and Expert Systems Applications, 16th International Conference, DEXA 2005,* Copenhagen, Denmark, 716-725.

Krötzsch, M., Vrandečić, D. & Völkel, M. (2006). Semantic MediaWiki. In *Proceedings of the 5th International Semantic Web Conference (ISWC06)*, 935-942, Springer.

Kuhn, H. W. (1955). The Hungarian Method for the assignment problem. *Naval Research Logistics Quarterly, 2*, 83–97.

Kurakawa, K., Shakya, A. & Takeda, H. (2009). A trial on extracting and integrating concepts of university staff directories for federated search. In *Proceedings of the 23rd Annual Conference of the Japanese Society for Artificial Intelligence*.

Kurematsu, M., Iwade, T. & Yamaguchi, T. (2004). DODDLE II: A domain ontology development environment using a MRD and text corpus. *IEICE transactions on information and systems, 87*(4), 908–916.

Kutz, O., Lutz, C., Wolter, F., & Zakharyaschev, M. (2004). ε-connections of abstract description systems. *Artificial Intelligence, 156*(1), 1-73.

Lakoff, G. (1987). *Women, fire, and dangerous things: What categories reveal about the mind*. University of Chicago Press Chicago.

Lenzerini, M. (2002). Data integration: A theoretical perspective. In *Proceedings of the twenty-first ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*, 233-246, ACM New York, NY, USA.

Lin, D. (1998). An information-theoretic definition of similarity. In *Proceedings of the 15th International Conference on Machine Learning*.

Liu, H., & Singh, P. (2004). ConceptNet – a practical commonsense reasoning tool-kit. *BT Technology Journal, 22*(4), 211-226. Springer.

McDowell, L., Etzioni, O., Gribble, S., Halevy, A., Levy, H., Pentney, W., Verma, D. & Vlasseva, S. (2003). Mangrove: Enticing ordinary people onto the semantic web via instant gratification. *Lecture notes in computer science*, 754-770.

Maedche, A. & Staab, S. (2001). Ontology learning for the Semantic Web. *IEEE Intelligent Systems*, *16*(2), 72–79.

Meersman, R. (1999). The use of lexicons and other computer-linguistic tools in semantics, design and cooperation of database systems. In *Proceedings of the Second International Symposium on Cooperative Database Systems for Advanced Applications (CODAS'99), Wollongong, Australia,* 1-14, Springer Verlag.

Mika, P. (2007). Ontologies are us: A unified model of social networks and semantics. *Web Semantics: Science, Services and Agents on the World Wide Web, 5*(1), 5-15.

Mika, P., Klein, M. & Serban, R. (2005). Semantics-based publication management using RSS and FOAF. In *Proceedings of the Semantic Desktop Workshop at the ISWC, vol. 175 of CEUR Workshop Proceedings, Galway, Ireland*.

Mika, P., 2005. Bibliography Management using RSS Technology (BuRST). Retrieved August 10, 2007, from http://www.cs.vu.nl/~pmika/research/burst/BuRST.html

Möller, K. & Decker, S. (2005). Harvesting desktop data for semantic blogging. In *Proceedings of the 1st Workshop on The Semantic Desktop – Next Generation Personal Information Management and Collaboration Infrastructure at ISWC2005,* Galway, Ireland, 79-91.

Möller, K., Breslin, J.G. & Decker, S. (2005). semiblog - semantic publishing of desktop data. In *14th Conference on Information Systems Development (ISD2005),* Karlstad, Sweden.

Möller, K., Bojārs, U.U. & Breslin, J.G. (2006). Using semantics to enhance the blogging experience. In *The Semantic Web: Research and Applications, vol. 4011 of Lecture Notes in Computer Science,* 679-696, Springer Berlin / Heidelberg.

Möller, K., Reif, G. & Handschuh, S. (2007). Moving Stuff – Linking Desktops with semiBlog, the Semantic Clipboard and RDFa. In *Proceedings of WWW2007, Developers Track*, Banff, Canada.

Morbidoni, C., Le Phuoc, D., Polleres, A., Samwald, M. & Tummarello, G. (2008). Previewing Semantic Web Pipes. In *Proceedings of the 5th European Semantic Web Conference (ESWC2008),* Tenerife, Spain, 843-848.

Murphy, G. (2004). *The big book of concepts*. Bradford Book.

Newman, R. (2005). Tag ontology design. Retrieved December 25, 2008, from http://www.holygoat.co.uk/projects/tags/

Nielsen, J. (2006). Participation inequality: encouraging more users to contribute. *Alertbox: Current Issues in Web Usability*. Retrieved April 16, 2009, from http://www.useit.com/alertbox/participation_inequality.html

Nielsen, J. (2008). Web users 'getting more ruthless'. In *BBC News*. Retrieved 24 May 2008 from, http://news.bbc.co.uk/2/hi/technology/7417496.stm

Nonaka, I. & Takeuchi, H. (1995). *The knowledge-creating company: How Japanese companies create the dynamics of innovation*. Oxford University press, US.

Ohmukai, I. & Takeda, H. (2004). Semblog: Personal knowledge publishing suite. In *Proceedings of WWW 2004 Workshop on the Weblogging Ecosystem: Aggregation, Analysis and Dynamics,* New York, USA.

Oren, E., Delbru, R., Möller, K., Völkel, M. & Handschuh, S. (2006). Annotation and navigation in semantic wikis. In *Proceedings of SemWiki in ESWC, 2006*.

O’Reilly, T. (2005). What is Web 2.0: Design patterns and business models for the next generation of software. Retrieved April 16, 2009, from http://www.oreillynet.com/pub/a/oreilly/tim/news/2005/09/30/what-is-web-20.html

Passant, A. (2008). LODr-A Linking Open Data Tagging System. In *Proceekings of the Social Data on the Web (SDoW2008) workshop,* Karlsruhe, Germany.

Passant, A. & Laublet, P. (2008). Meaning Of A Tag: A collaborative approach to bridge the gap between tagging and Linked Data. In *Proceedings of the WWW 2008 Workshop Linked Data on the Web (LDOW2008),* Beijing, China.

Passant, A., Hastrup, T., Bojārs, U. & Breslin, J. (2008). Microblogging: a semantic and distributed approach. In *Proceedings of the 4th Workshop on Scripting for the Semantic Web, Tenerife, Spain*.

Pell, I. B. (2007). Powerset - Natural language and the Semantic Web. Invited talk in *The 6th International Semantic Web Conference and the 2nd Asian Semantic Web Conference,* Busan, Korea. Retrieved May 10, 2009 from http://videolectures.net/iswc07_pell_nlpsw/

Peroni, S., Motta, E. & D'Aquin, M. (2008). Identifying Key Concepts in an Ontology, through the Integration of Cognitive Principles with Statistical and Topological Measures. In *Proceedings of the 3rd Asian Semantic Web Conference on The Semantic Web*, 242-256, Springer-Verlag Berlin, Heidelberg.

Pfisterer, F., Nitsche, M., Jameson, A. & Barbu, C. (2008). User-Centered Design and Evaluation of Interface Enhancements to the Semantic MediaWiki. In *Workshop on Semantic Web User Interaction at CHI 2008.*

Popov, B., Kiryakov, A., Kirilov, A., Manov, D., Ognyanoff, D. & Goranov, M. (2003). KIM-Semantic Annotation Platform. In *2nd International Semantic Web Conference (ISWC2003), vol. 2870* of Lecture Notes in Computer Science, 834-849, Springer.

Provost, D. (2008). On the cusp: A global review of the semantic web industry. Technical report.

Quan, D., Huynh, D. & Karger, D. R. (2003). Haystack: a platform for authoring end user Semantic Web applications. In *Proceedings of ISWC 2003*, 738-753.

Rahm, E. & Bernstein, P. (2001). A survey of approaches to automatic schema matching. *The VLDB Journal – The International Journal on Very Large Data Bases, 10*(4), 334-350.

Reif, G., Morger, M. & Gall, H.C. (2006). Semantic clipboard – semantically enriched data exchange between desktop applications. In *Semantic Desktop and Social Semantic Collaboration Workshopat the 5th International Semantic Web Conference ISWC06,* Athens, Geogria, *USA*.

Rowe, M. & Ciravegna, F. (2008). Getting to me: Exporting semantic social network from facebook.

Sauermann, L. (2003). *The Gnowsis-Using Semantic Web Technologies to build a Semantic Desktop*. Diploma thesis, Technical University of Vienna.

Sauermann, L., Bernardi, A. & Dengel, A. (2005). Overview and outlook on the semantic desktop. In *Proceedings of the 1st Workshop on The Semantic Desktop at the ISWC 2005 Conference*.

Sauermann, L., Van Elst, L. & Dengel, A. (2007). PIMO – a framework for representing personal information models. In *Proceedings of I-Semantics*, 270-277.

Schaffert, S. (2006a). IkeWiki: A Semantic Wiki for Collaborative Knowledge Management. In *Proceedings of the 15th IEEE International Workshops on Enabling Technologies: Infrastructure for Collaborative Enterprises*, 388-396, IEEE Computer Society Washington, DC, USA.

Schaffert, S. (2006b). Semantic social software: Semantically enabled social software or socially enabled semantic web? In *Proceedings of the SEMANTICS 2006 conference*, 99-112, OCG, Vienna, Austria.

Schaffert, S., Gruber, A. & Westenthaler, R. (2005). A semantic wiki for collaborative knowledge formation. In *Proceedings of SEMANTICS 2005 Conference,* Vienna, Austria.

Shakya, A., Takeda, H., Wuwongse, V. & Ohmukai, I. (2007a). SocioBiblog: A decentralized platform for sharing bibliographic information. In Isaías, P., Nunes, M. B., Barroso, J., eds., *Proceedings of the IADIS International Conference WWW/Internet 2007, 1,* 371-380, IADIS Press, Vila Real, Portugal.

Shakya, A., Wuwongse, V., Takeda, H. & Ohmukai, I. (2007b). OntoBlog: Linking ontology and blogs. In S. Handschuh, N. Collier, T. Groza, R. Dieng-Kuntz, A. de Waard & M. Sintek, eds., *Proceedings of the Semantic Authoring, Annotation and Knowledge Markup*, 47-54, Whistler, British Columbia, Canada, located at the 4th International Conference on Knowledge Capture (KCap 2007).

Shakya, A., Wuwongse, V., Takeda, H. & Ohmukai, I. (2008a). Ontoblog: Informal knowledge management by semantic blogging. In Y. Ouzrout & A. Hossain, eds., *Proceedings of the 2nd International Conference on Software, Knowledge, Information Management and Applications (SKIMA 2008),* 197-202, Kathmandu, Nepal.

Shakya, A., Takeda, H., Wuwongse, V. & Ohmukai, I. (2008b). SocioBiblog: Decentralized bibliographic information sharing through social links. *IADIS International Journal on WWW/Internet, 6*(2), 31-46.

Shamsfard, M. & Barforoush, A. A. (2003). The state of the art in ontology learning: a framework for comparison. *The Knowledge Engineering Review, 18*(4), 293–316.

Simpson, T., & Dao, T. (2005). *WordNet-based semantic similarity measurement.* Retrieved December 25, 2008, from http://www.codeproject.com/KB/string/semanticsimilaritywordnet.aspx

Siorpaes, K. & Hepp, M. (2007a). myOntology: The marriage of ontology engineering and collective intelligence. In *Bridging the Gap between Semantic Web and Web 2.0 (SemNet 2007)*, 127-138.

Siorpaes, K. & Hepp, M. (2007b). OntoGame: Towards Overcoming the Incentive Bottleneck in Ontology Building. In *Proceedings of the 3rd International IFIP Workshop on Semantic Web and Web Semantics (SWWS'07), Vilamoura, Portugal. On Semantic Web and Web Semantics (SWWS07), co-located with OTM Federated Conferences, Vilamoura, Portugal*, 1222-1232.

Specia, L. & Motta, E. (2007). Integrating folksonomies with the semantic web. In E. Franconi, M. Kifer & W. May, eds., *Proceedings of the European Semantic Web Conference (ESWC2007), vol. 4519 of LNCS,* 624-639, Springer-Verlag, Berlin Heidelberg, Germany.

Stuckenschmidt, H. & Van Harmelen, F. (2005). *Information sharing on the semantic web*. Springer-Verlag, Berlin Heidelberg, Germany..

Sure, Y., Gómez-Pérez, A., Daelemans, W., Reinberger, M.L., Guarino, N. & Noy, N.F. (2004). Why evaluate ontology technologies? because it works! *IEEE Intelligent Systems, 19*(4), 74-81, IEEE Educational Activities Department Piscataway, NJ, USA.

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

Takeda, H. (2008). Ontology for People-Vagueness and multiplicity. In *Proceedings of InterOntology08,* Tokyo, Japan.

Takeda, H., Iino, K. & Nishida, T. (1995). Agent organization and communication with multiple ontologies. *International Journal of Cooperative Information Systems, 4*(4), 321-337.

Tijerino, Y., Embley, D., Lonsdale, D., Ding, Y. & Nagy, G. (2005). *World Wide Web: Internet and Web Information Systems*, *8*(3), 261–285. Springer, Netherlands.

United States Intelligence Community (2008, February 22). *Information Sharing Strategy*. Retrieved May 10, 2009 from http://www.dni.gov/reports/IC_Information_Sharing_Strategy.pdf

Uren, V., Cimiano, P., Iria, J., Handschuh, S., Vargas-Vera, M., Motta, E. & Ciravegna, F. (2006). Semantic annotation for knowledge management: Requirements and a survey of the state of the art. *Journal of Web Semantics: Science, Services and Agents on the World Wide Web*, *4,* 14-28.

Van Damme, C., Hepp, M. & Siorpaes, K. (2007). Folksontology: An integrated approach for turning folksonomies into ontologies. In *Bridging the Gap between Semantic Web and Web 2.0 (SemNet 2007)*, 57-70.

Watts, D.J. (1999). *Small worlds: The dynamics of networks between order and randomness*. Princeton University Press.

Watts, D.J. (2003). *Six degrees: The science of a connected age*. WW Norton & Company, New York.

Zhdanova, A. & Shvaiko, P. (2006). Community-driven ontology matching. In *The Semantic Web: Research and Applications, LNCS: Vol. 4011*, 34-49, Springer, Berlin / Heidelberg.

Publications

Journal Publications

Shakya, A., Takeda, H. & Wuwongse, V. (2009). Community-driven linked data authoring and production of consolidated linked data. *International Journal on Semantic Web and Information Systems (Special Issue on Linked Data), 5*(3). IGI Global.

Shakya, A., Takeda, H., Wuwongse, V. & Ohmukai, I. (2008). SocioBiblog: Decentralized bibliographic information sharing through social links. *IADIS International Journal on WWW/Internet, 6* (2), 31-46.

International Conference Proceedings

Shakya, A., Takeda, H. & Wuwongse, V. (2008). Consolidating user-defined concepts with StYLiD. In J. Domingue & C. Anutariya, eds., *Proceedings of the 3rd Asian Semantic Web Conference (ASWC 2008), Bangkok, Thailand, LNCS, vol. 5367*, 287-301, Springer.

Shakya, A., Wuwongse, V., Takeda, H. & Ohmukai, I. (2008). OntoBlog: Informal knowledge management by semantic blogging. In A. Hossain & Y. Ouzrout, eds., *Proceedings of the 2nd International Conference on Software, Knowledge, Information Management and Applications, (SKIMA 2008), Kathmandu, Nepal*, 197-202.

Shakya, A., Takeda, H., Wuwongse, V. & Ohmukai, I. (2007). SocioBiblog: A decentralized platform for sharing bibliographic information. In P. Isaías, M. B. Nunes & J. Barroso, eds., *Proceedings of the IADIS International Conference WWW/Internet 2007, Vila Real, Portugal, 1*, 371-380, IADIS Press.

- *Awarded "Best Paper in the area of Web 2.0"*

Muljadi, H., Takeda, H., Shakya, A., Kawamoto, S., Kobayashi, S., Fujiyama, A. & Ando, K. (2006). Semantic wiki as a lightweight knowledge management system. In R. Mizoguchi, Z. Shi & F. Giunchiglia, eds., *The Semantic Web - ASWC 2006, Beijing, China, LNCS, vol. 4185*, 65-71, Springer.

International Workshop Proceedings

Shakya, A., Takeda, H. & Wuwongse, V. (2008). StYLiD: Social information sharing with free creation of structured linked data. In *Proceedings of the Social Web and Knowledge Management Workshop, (SWKM 2008) located at the 17th World Wide Web Conference (WWW2008), Beijing, China*, 33-40.

Shakya, A., Wuwongse, V., Takeda, H. & Ohmukai, I. (2007). OntoBlog: Linking ontology and blogs. In S. Handschuh, N. Collier, T. Groza, R. Dieng-Kuntz, A.D. Waard & M. Sintek, eds., *Proceedings of the Semantic Authoring, Annotation and Knowledge Markup Workshop (SAAKM 2007) located at the 4th International Conference on Knowledge Capture (KCap 2007), Whistler, British Columbia, Canada*, 47-54.

Shakya, A., Takeda, H., Ohmukai, I. & Wuwongse, V. (2008). A publication aggregation system using semantic blogging. In G. Li, Y. Liang & M. Ronchetti, eds., *The Semantic Web - ASWC 2006 Workshops Proceedings, Beijing, China*, 55-62, Jilin University Press.

International Poster and Demonstration Proceedings

Shakya, A., Takeda, H. & Wuwongse, V. (2008). Consolidating multiple concept definitions with StYLiD. In C. Bizer & A. Joshi, eds., *Proceedings of the Poster and Demonstration Session of the 7th International Semantic Web Conference, (ISWC 2008), Karlsruhe, Germany*.

Shakya, A., Takeda, H. & Wuwongse, V. (2008). StYLiD: Structure your own linked data. In *Poster Proceedings of the 2nd International Conference on Weblogs and Social Media, (ICWSM 2008), Seattle, Washington, USA*, 220-221.

Shakya, A., Takeda, H., Wuwongse, V. & Ohmukai, I. (2007). SocioBiblog: A decentralized platform for sharing bibliographic information. In *Poster + Demo Proceedings of the 6th International Semantic Web Conference and the 2nd Asian Semantic Web Conference, (ISWC 2007+ASWC 2007), Busan, Korea*, 93-94.

Shakya, A., Takeda, H., Wuwongse, V. & Ohmukai, I. (2007). SocioBiblog: Enabling communication on bibliography with semantic blogging. In *Poster proceedings of the International Conference on Weblogs and Social Media - ICWSM 2007, Boulder, Colorado, USA*, 297-298.

Local Conferences and Workshop Proceedings

Kurakawa, K., Shakya, A. & Takeda, H. (2009). A trial on extracting and integrating concepts of university staff directories for federated search. In *Proceedings of the 23rd Annual Conference of the Japanese Society for Artificial Intelligence*.

Shakya, A., Takeda, H. & Wuwongse, V. (2008). StYLiD: Structured information sharing with user-defined concepts. In *Proceedings of the annual conference of Japanese Business Model Society (Biz-model 2008), web technology and business model session, Tokyo, Japan*, 111-115.

Shakya, A. & Takeda, H. (2008). A report on linked data. *Technical report presented at SIG-SWO, セマンティックWeb とオントロジー研究会, A801-07, Tokyo, Japan*.

Shakya, A. & Takeda, H. (2008). Information sharing on the social semantic web. In *Proceedings of the second NEA-JC (Nepal Engineers' Association - Japan Chapter) Workshop on Current and Future Technologies, Tokyo, Japan*.

Appendix

Appendix A: Tasks for the Experiment on Usability

1. Task 1

Task 1 for StYLiD

Input the following data instance about the **band** into the system.

name: The Beatles

genre: rock and roll

origin: Liverpool, England (*Link this to the Wikipedia page of Liverpool)

members: (*Pick these from the singers already in the system)

John Lennon,

Paul McCartney,

George Harrison,

Ringo Starr

films:

A Hard Day's Night,

Help!

website: www.beatles.com

manager: Brian Epstein (URI: http://dbpedia.org/resource/Brian_Epstein)

past_members: Pete Best, Stuart Sutcliffe

description:

The Beatles are one of the most commercially successful and critically acclaimed bands in the history of popular music.

Task 1 for Freebase

Input the following data instance about the **band** into the system.

name: The Beatles

genre: rock and roll

origin: Liverpool, England

members: (*Pick these from the singers already in the system)

John Lennon,

Paul McCartney,

George Harrison,

Ringo Starr

films:

A Hard Day's Night,

Help!

website: www.beatles.com

manager: Brian Epstein

past_members: Pete Best, Stuart Sutcliffe

description:

The Beatles are one of the most commercially successful and critically acclaimed bands in the history of popular music.

2. Task 2

Task 2 for StYLiD

Input the “**Concert**” concept with the following attributes.

- **title** [description: title of the concert]
- **performer**

  (suggest that the performer may be a "**band**")
- **date**
- **venue** [description: location where the concert takes place]
- **type** [description: type of the concert]

  (enumerate *rock*, *classical*, *jazz*, *pop* as some possible values)
- **organizer**

  (suggest that the organizer may an “organization” or a “band”)

Description of the concept:
A concert is a live performance, usually of music, before an audience.

Task 2 for Freebase

Input the "**Concert**" concept with the following attributes.

- **title** [description: title of the concert]
- **performer**

  (suggest that the performer may be a "**band**")
- **date**
- **venue** [description: location where the concert takes place]
- **type** [description: type of the concert]
- **organizer**

Description of the concept:

A concert is a live performance, usually of music, before an audience.

3. Task 3

Input the following “**singer**” concept.

- **name** [description: name of the singer]
- **nationality** [description: country born]
- **genre** (List *rock,pop,classical,jazz,country* as some possible values}
- **member-of** (Singer may be member of a “band” or an “organization”)
- **years-active** [description: years when the singer is performing]
- **live-performances** (this may be “concerts”)

Description:

A person who is singing is called a singer or vocalist.

4. Task 4

Modify the "singer" concept that you just created to add the following attributes

- instrument [description: instrument played by the singer]
- website

5. Task 5

Task 5 for StYLiD

Post data about the following “**album**”

- **title:** A Hard Day's Night
- **artist:** The Beatles
- **released:** July 1964
- **genre:** Rock and roll, beat music, Rock
- **producer:** George Martin

Task 5 for Freebase

Post data about the following “**album**”

- **title:** A Hard Day's Night
- **artist:** The Beatles
- **released:** July 1964
- **genre:** Rock and roll, beat music, Rock
- **producer:** George Martin

6. Task 6

Task 6 For StYLiD

Find all the movies directed by “Martin Scorsese” which has “Leonardo DiCaprio” in the starcast**.**

Task 6 for Freebase

Find all the films directed by “Martin Scorsese” which has “Leonardo DiCaprio” as an actor**.**

Appendix B: Questionnaires

1. Participant Details

Participant Details

Name:

Nationality:

Gender:

Age:

Email:

Postal address *(your award will be sent by post)*:

Affiliation:

Qualification:

Current status:

Field of study/research area:

Interests/hobbies:

Note: Your personal information will not be disclosed anywhere.

2. Task-specific Questionnaire

Task number:

System: □ StYLiD □ Freebase

1. How confident did you feel?

□ very low □ low □ medium □ high □ very high

2. How easy was it?

□ very easy □ easy □ moderate □ difficult □ very difficult

3. Please mention if something was difficult:

4. Any comments/suggestions:

3. Task-specific Comparative Questionnaire

Task number:

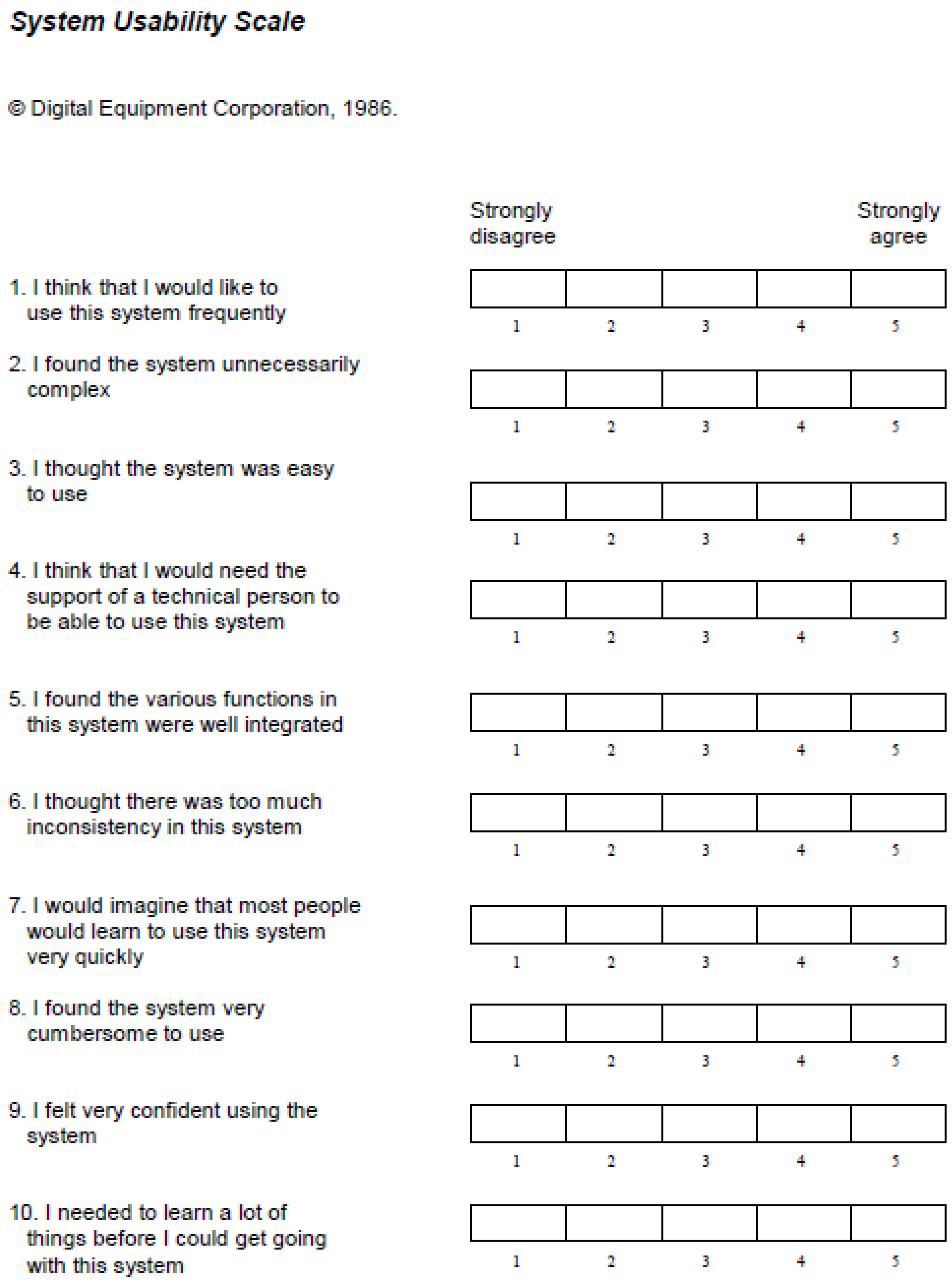
Which system did you feel more confident with for this task?

□ StYLiD □ Freebase □ almost same

Which system was easier for this task?

□ StYLiD □ Freebase □ almost same

4. System Usability Scale



System Usability Scale

	Strongly disagree				Strongly agree
1. I think that I would like to use this system frequently	□ 1	□ 2	□ 3	□ 4	□ 5
2. I found the system unnecessarily complex	□ 1	□ 2	□ 3	□ 4	□ 5
3. I thought the system was easy to use	□ 1	□ 2	□ 3	□ 4	□ 5
4. I think that I would need the support of a technical person to be able to use this system	□ 1	□ 2	□ 3	□ 4	□ 5
5. I found the various functions in this system were well integrated	□ 1	□ 2	□ 3	□ 4	□ 5
6. I thought there was too much inconsistency in this system	□ 1	□ 2	□ 3	□ 4	□ 5
7. I would imagine that most people would learn to use this system very quickly	□ 1	□ 2	□ 3	□ 4	□ 5
8. I found the system very cumbersome to use	□ 1	□ 2	□ 3	□ 4	□ 5
9. I felt very confident using the system	□ 1	□ 2	□ 3	□ 4	□ 5
10. I needed to learn a lot of things before I could get going with this system	□ 1	□ 2	□ 3	□ 4	□ 5

5. Final Questionnaire

Final Questionnaire

1. Do you know what the Semantic Web is? Yes No

2. Have you ever heard of the Semantic Web? Yes No

3. Have you ever done any database design? Yes No

4. Did you know about Freebase? Yes No

5. Did you know about Wikipedia? Yes No

6. Did you know about StYLiD? Yes No

7. Do you usually read the manual/help when using a new online system? Yes No

8. Please write if you have any more comments/suggestions overall.

Appendix C: Experiment on Conceptualization

1. Conceptualization Task

Please read the provided texts and list facts about the thing the text is about. (The facts should be in the given text and not from outside or your general knowledge).

List the facts as attributes and values as in the following examples.

Concept: *Band*

Attribute	**Value**
name	*The Beatles*
origin	*England*
members	*John Lennon, Paul McCartney, George Harrison, Ringo Starr*
years active	*1960's*
genre	*Rock, pop*

Concept: *Concert*

Attribute	**Value**
Name	*Eric Clapton & Jeff Beck live in Japan*
artist	*Eric Clapton,* *Jeff Beck*
date	*February 2009*
venue	*Saitama, Japan*

2. Texts Provided to Participants

The Edo-Tokyo Open Air Architectural Museum

The Edo-Tokyo Open Air Architectural Museum exhibits a range of historic buildings from the Tokyo area. The buildings were relocated or reconstructed there in order to preserve a chapter of architectural history, which has been almost completely lost in fires, earthquakes, wars and city redevelopment.

Most of the buildings exhibited are from the Meiji Period (1868-1912) or more recent times, and include among others, a politician's elegant former residence, a farm house, a public bathhouse, various shops and a police box.

The Edo-Tokyo Open Air Architectural Museum is a branch museum of the superb Edo-Tokyo Museum.

The open air museum is located in the western part of Koganei Park, Koganei City, 25 minutes west of Tokyo's Shinjuku Station by train. From Shinjuku, you can either access it by the Seibu Shinjuku Line (260 yen to Hana-Koganei Station) or JR Chuo Line (290 yen to Musashi-Koganei Station).

From either station, the park is a 5 to 10 minute bus ride or 15 to 30 minute walk.

The Historic Village of Hokkaido

The Historic Village of Hokkaido (kaitaku no mura) is an open air museum in the suburbs of Sapporo. It exhibits about 60 typical buildings from all over Hokkaido, dating from the Meiji and Taisho Periods (1868 to 1926), the era when Hokkaido's development was carried out on a large scale.

The open air museum is divided into a town, fishing village, farm village and mountain village section. The Historical Museum of Hokkaido (kaitaku kinenkan), which documents the history of Hokkaido's development, can be found nearby.

The Historic Village of Hokkaido is located in the Nopporo Forest Park (Shinrin Koen) outside of Sapporo. From Sapporo Station, take a local train on the JR Hakodate Line to Shinrin Koen Station (about 15 minutes) from where the museum is a 5 minute bus ride or 15-20 minute walk.

Hasedera (Hase Temple)

Hase Temple is a temple of the Jodo sect, that is most famous for its statue of Kannon, the goddess of mercy. The statue shows Kannon with eleven heads, each representing a characteristic of the goddess. The 9.18 meter tall, gilded wooden statue is regarded as the largest wooden sculpture in Japan, and can be viewed in the temple's main building.

Visitors to Hase Temple can enjoy a great view of the coastal city of Kamakura from the terrace next to the temple's main buildings. There is also a small restaurant where Japanese sweets such as mitarashi dango, small rice flour dumplings covered with a sticky sauce made of sugar and soya sauce, other small meals and beverages are served.

Next to the temple garden and the pond stands the Bentendo, a small hall that contains a figure of Benten (or Benzaiten), a goddess of feminine beauty and wealth. Sculptures of Benten and other minor gods can be found in a small cave (Bentenkutsu) next to the Bentendo.

Hase Temple is located a 5 minute walk from the Enoden Railway Hase Station, the third station from Kamakura main station. The Enoden is a streetcar-like train that connects Kamakura with Enoshima and Fujisawa. Its terminal station in Kamakura is located just west of JR Kamakura Station.

Kiyomizudera

Kiyomizudera ("Pure Water Temple") is one of the most celebrated temples of Japan. It was founded in 780 and remains associated with the Hosso sect, one of the oldest sects within Japanese Buddhism. In 1994, the temple was added to the list of UNESCO world heritage sites. Kiyomizudera stands in the wooded hills of eastern Kyoto and offers visitors a nice view over the city from its famous wooden terrace. Below the terrace, you can taste the spring water, which gives the temple its name and which is said to have healing power.

Behind Kyomizudera's main hall stands Jishu Shrine, a shrine dedicated to the deity of love. In front of the shrine are two rocks, placed several meters apart from each other. Successfully walking from one to the other rock with your eyes closed is said to bring luck in your love live.

Part of the fun of visiting Kiyomizudera is the approach to the temple along the steep and busy lanes of the atmospheric Higashiyama district. Except early in the morning, do not expect a tranquil, spiritual atmosphere.

The many shops, restaurants and ryokan in the area have been catering to tourists and pilgrims for centuries. Products on sale range from local specialties such as Kiyomizu-yaki pottery, sweets and pickles to the standard set of souvenirs.

Kiyomizudera can be reached from Kyoto Station in about 15 minutes by bus. Take bus number 100 or 206 and get off at Kiyomizu-michi or Gojo-zaka, from where it is a 10-15 minute uphill walk to the temple.

Shinagawa Prince Hotel

Shinagawa prince hotel is located in tokyo, japan. The hotel is three kilometers from Tokyo tower. Roppongi and Tsukiji fish market are four kilometers from the hotel. Ginza, a popular shopping and entertainment district, is five kilometers away. Transportation to and from Tokyo's main attractions is available at Shinagawa station, located 200 meters from the hotel.

The four-tower hotel includes 16 restaurants and bars. The yahoo café serves light Japanese snacks and features internet access. Nanakamado Japanese restaurant serves breakfast in the mornings and pub-style food in the afternoons and evenings. Recreational activities available at the hotel include indoor and outdoor pools, a bowling alley, an indoor golf center and a game room. The hotel also features an aquarium, IMAX theater and 10-screen multiplex cinema complex.

The guestrooms at Shinagawa prince hotel are located throughout the hotel's four towers and include internet access and cable television.

For guests wishing to be environmentally friendly, the hotel participates in active environmental friendly practices. In lieu of a "no room cleaning" request, a JPY500 per room, per night credit for hotel facility-use will be issued.

The Prince Park Tower Tokyo Hotel

Rising above the green grounds of Shiba park, along the same visual parallel as nearby Tokyo tower, the prince park tower Tokyo makes a welcome addition to Tokyo's park-hotel offerings. The hotel was opened in 2005 and the majority of the property physically lies inside Shiba park. Soaring 33 stories high allows for arresting aerial views of the adjacent Zojo-ji temple, which is also situated inside Shiba park. Located within the business districts of Toranomon and Kasumigaseki, as well as the fashionable shopping district of Roppongi, the prince park tower Tokyo offers a first-class location to business and leisure travelers alike. The nearest subway is Onarimon while the Toei Asakusa, Ooedo, Hibiya and JR lines are all approximately two kilometers away. International travelers arriving at Narita international airport (NRT) can take the Narita express to connect to one of these local train lines.

The business traveler has major facilities to suit virtually any scale of event, including a 3,600-occupancy ballroom and convention hall. Dining options are plentiful which includes the Brise Verte that features fine French dining set against the backdrop of Tokyo's cityscape from 33 floors in the air. Sushi, tempura, and yakitori are also available in a restaurant dedicated to each of these distinctive Japanese cooking styles. The melody line jazz bar features live jazz entertainment, which given Tokyo's obsession with the cool music style, Tokyo and top hotel bars such as this one have talented jazz performers from around the globe.

The 673 guest rooms and suites include 397 rooms that have balconies. Each guestroom features a Jacuzzi bathtub and separate shower stall. In addition, every room is equipped with complimentary high-speed internet access and LCD flat-screen televisions. After a day of absorbing Japan's leading cultural center or a demanding day of business, guests can avail themselves of the hotel's natural hot spring spa and steam away the relentless energy of Japan's imperial city.

3. Table for Representing Conceptualization

Title:
Concept: ________________________________

Attribute	Value

Appendix D: Results of the Experiment on Usability

1. Evaluation of Task 1

Participant	Confidence		Ease		Time (in mins)		Errors		Assistance		Comparison	
	S	**F**	**S**	**F**	**S**	**F**	**S**	**F**	**S**	**F**	**confidence**	**ease**
1	2	2	1	1	13	9	4	1	3	3	same	S
2	3	3	4	2	11	14	0	0	2	6	F	S
3	4	4	3	4	9	7	2	0	2	1	F	F
4	4	2	4	2	8	10	1	0	1	3	S	S
5	4	2	3	1	7	14	0	0	0	6	S	S
6	3	2	3	2	8.5	8	1	0	1	1	S	S
7	3	1	4	1	7.5	12	2	0	1	3	S	S
8	3	1	3	1	9	15	0	0	0	4	S	S
9	3	1	3	0	8	12	0	0	0	5	S	S
10	3	2	3	1	14.5	18	0	1	2	3	S	S
11	0	0	2	1	16	12	1	2	3	6	S	S
12	3	2	3	2	9	13	1	1	2	5	S	S
13	2	2	3	2	12	13	1	1	1	4	F	S
14	3	2	4	3	10	13	0	0	2	5	S	S
15	2	2	3	3	11	11	1	1	3	4	F	F
Average	2.8	1.87	3.07	1.73	10.23	12.07	0.93	0.47	1.53	3.93		

Note: S stands for StYLiD and F stands for Freebase

2. Evaluation of Task 2

Participant	**Confidence**		**Ease**		**Time (in mins)**		**Errors**		**Assistance**		**comparison**	
	S	**F**	**S**	**F**	**S**	**F**	**S**	**F**	**S**	**F**	**confidence**	**ease**
1	3	2	3	1	8	10.5	2	1	2	1	S	S
2	4	2	4	3	7.5	13	0	0	2	1	S	S
3	2	2	2	2	6	11	2	1	1	2	F	F
4	3	2	4	3	5.5	9	0	0	0	1	S	S
5	4	3	3	2	3.5	11.5	3	0	1	4	S	S
6	3	3	4	3	2.5	5	1	0	0	0	S	S
7	3	1	4	2	6	14	0	1	0	2	S	S
8	4	3	4	3	10	12	2	0	0	1	same	S
9	3	1	3	2	8	8	1	0	1	0	S	S
10	3	3	3	2	8.5	16	2	1	0	1	S	S
11	2	2	2	2	15	13	1	0	3	4	S	S
12	4	3	4	2	4	13	0	0	1	4	S	S
13	1	2	2	2	13	12	1	1	2	5	F	F
14	3	3	3	3	6	14	1	0	1	5	S	S
15	2	2	2	3	8	9	1	1	3	4	F	F
Average	2.93	2.27	3.13	2.33	7.43	11.4	1.13	0.4	1.13	2.33		

3. Evaluation of Task 3

Participant	Confidence	Ease	Time (in mins)	Errors	Assistance
1	3	3	5.5	1	1
2	4	4	6	0	0
3	2	3	3	1	0
4	4	4	4	0	0
5	2	3	7	4	2
6	4	4	2.5	0	0
7	4	4	5	0	0
8	3	4	6.5	1	1
9	3	3	4	1	0
10	3	3	5.5	2	0
11	3	3	7	1	0
12	3	3	6	1	0
13	3	2	8	2	0
14	4	4	6	0	1
15	3	3	4	1	0
Average	3.2	3.33	5.33	1	0.33

4. Evaluation of Task 4

Participant	Confidence	Ease	Time (in mins)	Errors	Assistance
1	4	4	2	0	0
2	4	4	1	0	0
3	2	3	0.5	0	0
4	4	4	1.5	0	0
5	4	4	3.5	0	0
6	4	4	1	0	0
7	4	4	2	0	0
8	4	4	1.5	0	0
9	4	4	2	0	0
10	3	3	4	0	0
11	3	4	2	0	0
12	3	4	2	0	1
13	3	3	3	0	0
14	4	4	2	0	1
15	3	3	2	0	0
Average	3.53	3.73	2	0	0.13

5. Evaluation of Task 5

Participant	confidence		Ease		Time (in mins)		Errors		Assistance		Comparison	
	S	F	S	F	S	F	S	F	S	F	confidence	ease
1	3	3	3	2	7.5	14	1	1	1	1	S	S
2	4	2	4	2	9	15	2	1	1	1	S	S
3	2	2	2	2	2.5	7.5	1	2	0	2	F	F
4	3	3	4	3	6.5	8	0	0	0	0	same	S
5	4	2	3	1	4.5	10.5	1	1	0	2	S	S
6	4	3	4	2	4	6.5	0	0	0	0	S	S
7	3	2	4	1	5	13.5	0	2	0	3	S	S
8	2	2	4	2	14	10.5	0	1	1	1	S	S
9	3	3	3	3	7	8	1	0	0	0	S	S
10	3	1	3	1	6	17.5	1	0	0	3	S	S
11	3	2	3	2	6	11	1	2	2	4	S	S
12	4	2	4	3	3	7	1	1	0	2	S	S
13	3	3	3	2	6	8	1	1	2	2	S	same
14	4	3	4	3	4	10	1	1	0	4	S	S
15	3	2	3	2	3	9	2	2	1	5	S	S
Average	3.2	2.33	3.4	2.07	5.87	10.4	0.87	1	0.53	2		

6. Evaluation of Task 6

Participant	Confidence		Ease		Time (in mins)		Errors		Assistance		Comparison	
	S	**F**	**S**	**F**	**S**	**F**	**S**	**F**	**S**	**F**	**confidence**	**ease**
1	1	1	2	1	8	7.5	0	0	3	3	S	S
2	3	3	3	2	5.5	11	0	0	1	3	S	same
3	4	1	0	1	3	2.5	0	0	1	1	F	F
4	4	0	4	0	2.5	9	0	0	0	3	S	S
5	3	1	1	1	5.5	5	0	0	4	2	S	S
6	4	2	3	2	2	4.5	0	0	1	1	S	S
7	3	0	4	2	2	4.5	0	0	0	1	S	S
8	3	2	3	2	7.5	9	0	0	2	2	same	S
9	3	3	2	2	5	4	0	0	1	0	same	same
10	0	0	0	0	11	11	0	0	5	5	same	same
11	1	0	1	1	7	8	0	0	3	3	S	S
12	3	3	3	2	3	5	0	0	2	1	S	S
13	3	3	3	2	5	7	0	0	3	3	S	same
14	4	3	4	3	3	9	0	0	2	3	S	S
15	3	1	3	1	4	9	0	0	2	5	S	S
Average	2.8	1.53	2.4	1.47	4.93	7.07	0	0	2	2.4		