2008年 ビジネスモデル学会秋季大会
Biz-model 2008, Tokyo-u, Tokyo, JAPAN
Oct.11, 2008

StYLiD: ユーザ定義概念による構造化された情報共有

国立情報学研究所　Shakya　Aman
国立情報学研究所　武田　英明、アジア工科大学院　Vilas Wuwongse

StYLiD: Structured Information Sharing with User-defined Concepts

National Institute of Informatics　Shakya Aman,
National Institute of Informatics　Hideki Takeda ,
Asian Institute of Technology　Vilas Wuwongse

Information sharing can be effective with structured data. However, there are several challenges for having structured data. Creating structured definitions is difficult and multiple conceptualizations may exist due to different user requirements and preferences. We propose a social software called StYLiD for sharing a wide variety of structured data. It provides an easy platform for users to freely define their own structured concepts. We propose consolidating multiple concept definitions by different users into a unified virtual concept. Attributes of the multiple concept versions are aligned semi-automatically to provide a unified view. Popular concepts gradually emerge from the cloud of concepts while concepts evolve incrementally. StYLiD also helps in interlinking data including external resources like Wikipedia. It can be used as an information sharing platform, lightweight content management system and for information integration and collaborative design of data models.

information sharing, semantic web, social software

1. Introduction

People want to share a wide variety of information on the web. Different types of data can be modeled by structuring them. The semantics of structured data facilitates automated processing. The Semantic Web [Berners-Lee 01] is aimed towards a web of data with semantics that machines can understand, process automatically and provide intelligent applications. Information sharing, searching and browsing become more effective with structured data. Interoperability and integration of data from various sources becomes possible with standard or compatible formats. An ontology, in the particular area of interest, models the structure of data with well-defined semantics. An ontology is an explicit specification of a conceptualization, i.e., the modeling of the objects, concepts, and entities that exist in the area of interest and the relationships that hold among them [Gruber 93].

However, there are several challenges for sharing structured data on the web. There is a long tail of domains[Huynh 07] for which people have information to share but not many information systems or ontologies to cover the wide variety of information[Siorpaes 07]. Developing individual solutions every time is infeasible. It is not easy to create adequate definitions. Moreover, people conceptualize the same thing in different ways based on different contexts, requirements or preferences. Hence, multiple heterogeneous conceptualizations always exist.

Creating ontologies should be a widely collaborative and incremental process[Siorpaes 07]. However, to have mass participation, systems should be easy to understand and use. Unlike ontologies, social software has proven to be very successful in that. Thus, the combination of social software and Semantic Web technologies can be promising [Ankolekar 07, Gruber 08].

We propose a social software to enable general users to freely define their own conceptualizations and share structured data based on that. We propose allowing users to have multiple conceptualizations and, at the same time, consolidating them into a single virtual conceptualization. The consolidation of multiple user-defined concepts serves as a new way of building up richer concept definitions collaboratively. We have implemented a system called StYLiD[1] (an acronym for Structure Your own Linked Data) to realize our approach. Multiple definitions are consolidated by semi-automatic alignment of concept attributes. Popular concepts can gradually emerge out by usage in the same way as folksonomies or tagging. StYLiD produces linked data[Berners-Lee 06, Bizer 07] for the Semantic Web by interlinking internal and external resources. The system integrates various aspects of social software and Semantic Web technologies into a synergetic whole.

2. The StYLiD Platform

2.1 Sharing Structured Data with User-defined Concepts

Suppose a user wants to share some structured information, let's say details of a seminar. He may register an account on StYLiD and easily define his own "seminar" concept on the fly with a list of attributes like topic, speaker, date, time, venue, etc. The main interface of StYLiD is shown in Fig. 1. The users may freely define their own concepts by specifying the concept name, some description and a set of attributes. Each attribute is defined by the attribute name, description and the suggested value range as shown in Fig. 2. Once a concept has been defined, any user may post data instances using system generated online forms as shown in Fig. 3. The posts may be shared in the community using social software features. By posting data instances and inviting others to post to the system we can maintain a useful collection of information in a structured way.

The user does not need to define concepts from scratch. He may modify an existing concept to make his own version. The system creates a copy of the concept and makes modifications on it. Attributes from an existing concept may also be imported to define a different concept. Users can update their own concept definitions to add new attributes whenever needed. Thus, concept definitions can evolve incrementally along with different versions.

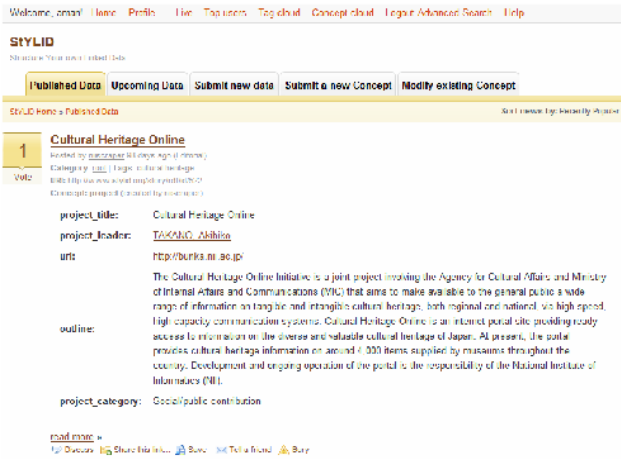

Fig. 1: StYLiD interface.

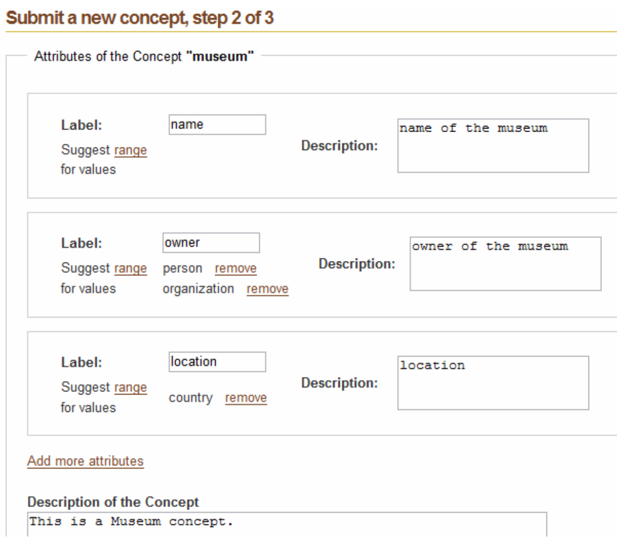

Fig. 2: Interface to create a new concept.

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

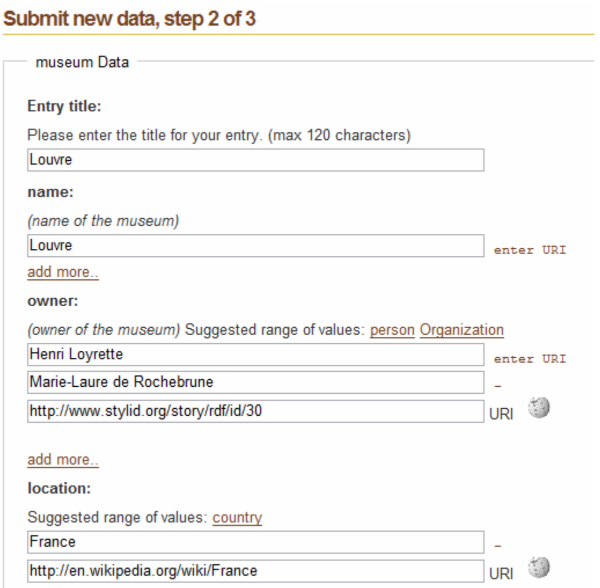

Fig. 3: Interface to enter instance data.

It is difficult to think of all attributes and all possible value ranges. Moreover, we may not always have perfect data at the time of data entry. To avoid these difficulties we allow flexible and relaxed definitions. The range of values defined for attributes, as seen in Fig. 2 and 3, is only suggestive and does not impose strict constraints. Rather the system assists the user to pick instances from the suggested range. The system accepts both literals and URI identifiers as values for any attribute. Users may input single or multiple values for any attribute.

The system also offers a personal structured data space. It provides a *Concept Collection* for each user, as seen in Fig. 5. Concepts created or adopted by the user are automatically added to this collection. Users can also add any other useful concepts to their collection. The concepts actually created by the user are shown in a separate tab.

2.2 Consolidation of Concepts

Different users may have multiple definitions for the same concept. As illustrated in Fig. 4, the same "Hotel" concept may be defined by 3 users in different ways. Even the same user may have multiple versions for the concept. The system groups them and consolidates them into a single virtual "Hotel" concept combining all the features of the individual definitions. Our 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 02].

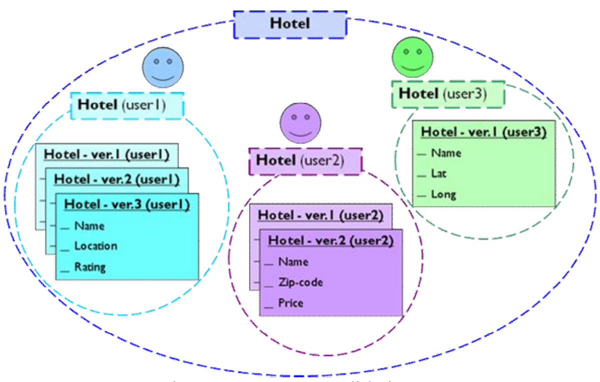

Fig. 4: Concept consolidation.

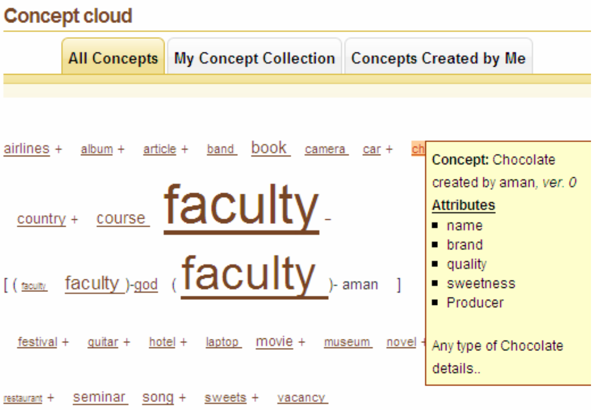

Fig. 5: Concept cloud.

Consolidated Concept Cloud. All the concepts are visualized in a Concept Cloud as shown in Fig. 5. Clicking on a concept shows all its instances. Hovering on a concept shows its details. Popular concepts with more data instances appear bigger in the cloud. Thus, stable definitions gradually emerge out from the cloud as more data instances are contributed.

Concepts defined by different users with the same name are grouped together by the system forming a single virtual concept. We have also done some work on consolidating concepts with different names and grouping similar concepts. 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. In the sub-cloud, multiple versions defined by the same user are subgrouped together. In Fig. 5, 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 shows all instances of all its versions. This is useful because the user may want to browse all instances of a concept regardless of the concept version. We can also list instances of the multiple versions of a concept defined by a single user by clicking on the user name.

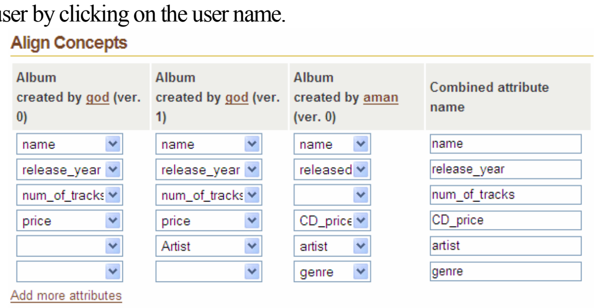

Fig. 6: Aligning the attributes of multiple concepts.

Semi-Automatic Concept Alignment. Different concepts in a consolidated group can be aligned to produce a uniform and integrated view. When the instances of a consolidated group of concepts are viewed as a single table, the system automatically suggests alignments between the attributes, as shown in Fig. 6. Matching attributes are automatically selected in the form-based interface. The Alignment API[2] with its WordNet extension has been used for the purpose. It utilizes a

[2] http://alignapi.gforge.inria.fr/

WordNet based similarity measure between attribute labels to find alignments. However, more sophisticated alignment methods may be used in the future. There is a large body of research about schema matching and ontology alignment [Euzenat 04].

It is not possible to make the alignment fully automatic and accurate. So it is necessary to have the user in loop to complete the process by adding or modifying mappings not correctly suggested by the system. The alignments are saved by the system so that once a user completes the alignment others need not do it again. Thus, both machine intelligence and human intelligence are used in getting the concepts aligned. The alignment can be incrementally updated as more concepts may be added to the consolidated group of concepts.

A Unified View. Each set of aligned attributes is mapped to a single consolidated attribute. The system automatically fills a name for each consolidated attribute, as shown in Fig. 6, 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. The table can be sorted by any field. The table of data may also be exported to spreadsheet applications like Microsoft Excel.

2.3 Creating Linked Data

The system facilitates creation of Semantic Web linked data by the use of URIs. Data instances can be linked to each other directly by entering resource URIs as attribute values (see Fig. 3). The system provides support for this by suggesting range of values for the attributes. The user may easily pick up instances from this range. The data appears as simple hyperlinked entries for the user (see Fig. 1). However, the linked data can be crawled by machines to enable powerful applications.

The system provides support to link to Wikipedia contents. The familiar Wikipedia icon is seen next to the URI field (see Fig. 3). When the user clicks on the icon it searches for the Wikipedia page about the text attribute value typed by the user and displays it as a pop-up. The user may copy the Wikipedia page URL as the URI. Transparent to the user, the system converts it into the corresponding DBpedia URI. DBpedia [Auer 07] exposes the structured data in Wikipedia on the Semantic Web like a database. Unlike DBpedia, Wikipedia is well understood by general people and user-friendly. So the users would be motivated to link to Wikipedia pages to make their data more informative, interesting and useful. The user may directly enter any external URI to link to other external resources too.

2.4 Querying and Using Structured Data

The user would be able to browse different types of data using the concepts defined by him and others. He can also navigate through linked data entries. The system provides a structured search interface, as shown in Fig. 7, to retrieve instances of a concept by specifying attribute, value pairs as criteria. The search can be done over a consolidated concept as well. In that case, the query terms are unfolded to aligned attributes of all versions of the concept. For instance, if he searches using the "venue" attribute of the seminar, the "location" attribute aligned in a different version would also be searched. The system also provides a SPARQL[3] query interface for more complex queries.

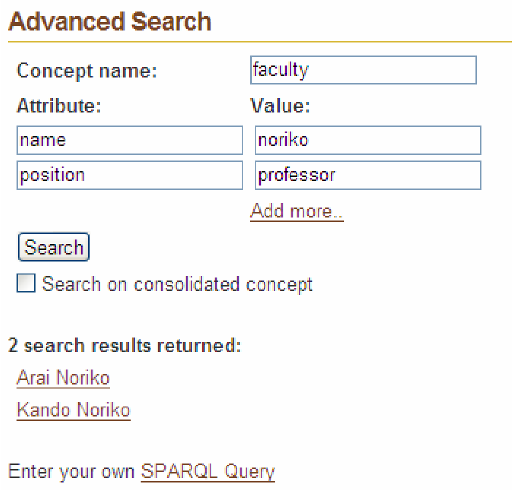

Fig. 7: Structured search interface.

The system embeds machine understandable data in the HTML posts using RDFa[4]. Many useful RDFa tools and plug-ins are available[5] and we may expect more powerful tools to be available in the future. Users with some programming knowledge may code small scripts with the Operator[6] browser extension to create useful operations for different types of data. For example, a "conference" concept may allow an operation to help in booking a hotel at the conference "venue" from "start date" to "end date". Moreover, developers may import the structured data into various useful applications.

3. Usage

StYLiD may at least be used in the following ways.

1. *Information sharing/gathering platform*: StYLiD may be used as a social platform for sharing and gathering different types of data, both in public and enterprise settings.
2. *Lightweight content management system*: It can also be used as a simple content management system or a data backend to drive dynamic online applications and websites.
3. *Information integration*: It can be used to map and integrate separately defined databases and serve inter-departmental or inter-organizational information exchange.
4. *Collaborative designing*: It can be used as a collaborative platform to create emergent ontologies or schemas. It can serve as an inexpensive tool for rapid prototyping when requirements are not well-defined initially.

4. Related Work

Freebase[7] allows users to define their own structured types. However, these are kept separate and not consolidated or related in any way. It is also difficult to contribute types and instances because of strict

---

3 http://www.w3.org/TR/rdf-sparql-query/

4 http://www.w3.org/TR/xhtml-rdfa-primer/

5 http://esw.w3.org/topic/RDFa

6 http://www.kaply.com/weblog/operator/

7 http://www.freebase.com/

constraints. Exhibit [Huynh 07] enables ordinary users to create web pages with structured data. However, authoring such pages would be cumbersome. myOntology [Siorpaes 07] uses wikis for collaborative lightweight ontology building by enabling general users. Semantic wikis [Buffa 08] make the collaborative knowledge contributed by users more explicit and formal. Freebase, myOntology and semantic wikis are all based on wikis which assume that each resource can only have a single prominent version. However, in practice, multiple conceptualizations may exist. There had been a lot of works on semantic blogging [Cayzer 04] which enhance blog items with semantic structure. However, they deal with limited types of metadata and the schemas do not evolve.

5. Conclusion and Future Work

We proposed StYLiD as a social software for sharing different types of structured data. It facilitates collaborative creation of knowledge while meeting individual requirements of people. We also proposed an approach for consolidating multiple user-defined conceptualizations. Concepts can evolve incrementally and emerge by popularity. The presented work combines various aspects of social semantic software into a more effective whole.

In the future, better alignment techniques may be employed. Existing vocabularies should be imported and reused. Other useful mash-ups may be introduced to exploit the structured data. We can facilitate users to contribute plugins for handling different types of data. Scrapers may be associated to concepts for collecting data from web pages easily.

References

[Auer 07] Auer, S., Bizer, C., Lehmann, J., Kobilarov, G., Cyganiak, R., Ives, Z. DBpedia: A nucleus for a web of open data. In *Proceedings of the 6th International Semantic Web Conference and 2nd Asian Semantic Web Conference*, Busan, South Korea. LNCS, vol. 4825, pp. 715–728. Springer, Heidelberg (2007)

[Ankolekar 07] Ankolekar, A., Krötzsch, M., Tran, T., Vrandečić, D. 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, pp. 825–834. ACM Press, New York (2007).

[Berners-Lee 01] Berners-Lee, T., Hendler, J., Lassila, O. The Semantic Web. Scientific American (2001).

[Berners-Lee 06] Berners-Lee, T. Linked data. World Wide Web design issues. (July 2006) http://www.w3.org/DesignIssues/LinkedData.html

[Bizer 07] Bizer, C., Cyganiak, R., Heath, T. How to Publish Linked Data on the Web (2007) http://www4.wiwiss.fu-berlin.de/bizer/pub/LinkedDataTutorial/

[Buffa 08] Buffa, M., Gandon, F., Ereteo, G., Sander, P., Faron, C. SweetWiki: A Semantic Wiki. Journal of Web Semantics 6(1), 84–97 (2008)

[Cayzer 04] Cayzer, S. Semantic blogging and decentralized knowledge management. Communications of the ACM 47(12), 48–52 (2004)

[Euzenat 04] Euzenat, J., Le Bach, T., Barasa, J., et al.: State of the art on ontology alignment. Knowledge Web Deliverable D2.2.3 (2004)

[Gruber 93] Gruber, T, R. A translation approach to portable ontology specifications. Knowledge Acquisition, 5(2), pp. 199–220 (1993).

[Gruber 08] Gruber, T. Collective knowledge systems:Where the social web meets the Semantic Web. Journal of Web Semantics 6(1), pp. 4–13 (2008).

[Huynh 07] Huynh, D., Karger, D., Miller, R. Exhibit: lightweight structured data publishing. In *Proceedings of the 16th international conference on World Wide Web*, pp. 737–746. ACM Press, New York (2007).

[Lenzerini 02] Lenzerini, M. Data integration: A theoretical perspective. In *Proceedings of the twenty-first ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*, pp. 233–246 (2002)

[Siorpaes 07] Siorpaes, K., Hepp, M. myOntology: The marriage of ontology engineering and collective intelligence. In *Bridging the Gap between Semantic Web and Web 2.0*