Semantic Web Technologies Applied to Glycoscience Data to Integrate with Life Science Databases

Kiyoko F. Aoki-Kinoshita

Department of Bioinformatics, Faculty of Engineering, Soka University,
1 – 236 Tangi-machi, Hachioji, Tokyo, 192 – 8577 Japan

E-Mail: kkiyoko@soka.ac.jp

Received: 1st November 2013 / Published: 22nd December 2014

Abstract

The World Wide Web (WWW) essentially consists of web pages containing data that are linked to other resources or pages on the Internet. Therefore, to accumulate information regarding a particular carbohydrate, for example, a user would either make searches in individual databases and/or read the scientific literature and then follow various links on the Web to get relevant information. In order to overcome such tedious tasks, the Semantic Web was born from the concept of incorporating semantics, or meanings, into each data item, which is represented as a web page, or URI (Uniform Resource Identifier). Thus, a single carbohydrate structure, for example "Man_9", would be assigned a URI, and then semantics are assigned to it by preparing a dataset containing information, or annotations, about it. Each data item is annotated in the form of triples consisting of Subject, Predicate and Object. Thus, for example, "Man_9 part_of N-linked_glycan" would be a triple where Man_9 is the Subject, part_of is the Predicate and N-linked_glycan is the Object. If there was another triple, "Mannose part_of Man_9," then it can be computationally inferred that Mannose is a part of N-linked_glycan. Moreover, this triple data can be queried as a database.

This article is part of the Proceedings of the Beilstein Glyco-Bioinformatics Symposium 2013.
www.proceedings.beilstein-symposia.org

In the life sciences, many major databases have started making their data available on the Semantic Web in the form of triples, including UniProt and EBI, and the Integrated Database Project of Japan has decided to use the Semantic Web as the integrating factor among life science databases in Japan. As a part of this project, the Japanese Consortium for Glycobiology and Glycotechnology Database (JCGGDB) has called upon glycomics database providers to use Semantic Web technologies in their databases so that they can all be integrated. Thus, developers of GlycomeDB, BCSDB, MonosaccharideDB and UniCarbKB were invited to Japan to participate in a BioHackathon to learn about the Semantic Web and develop a standard by which glycomics data can be presented as triples. As an initial proof-of-concept, we were able to successfully generate on-the-fly queries across multiple databases using Semantic Web technology. In this manuscript, an introduction to the Semantic Web and these efforts to integrate glycoscience databases will be described.

Introduction

The World Wide Web (WWW), or simply the Web, has grown tremendously since it was first developed as a file server. The first version of the Web, or Web 1.0, was simply a large number of file servers that statically served files to clients through a web browser. Other than file requests, users did not have any way to interact with the server. Eventually, some web servers started providing search engines to allow users to search for data of interest. This became a part of Web 2.0, where users started becoming more active in presenting data and interacting with the Web, with blogs, tweets, social networking and of course searching.

Now it has come to be understood that even Web 2.0 is insufficient. Especially for researchers who often use the web to search for literature and existing research results, much of their time involves collecting information and integrating them to gain a better understanding of a particular research topic, for example. Thus currently, Web 3.0 has started to enable researchers to more easily search for relevant data by providing semantics to data on the Web. This is also known as the Semantic Web [1], where links are now annotated with semantic information, which the computer can potentially use to make inferences regarding relationships between available data. Realization of the Semantic Web requires the data to be annotated in the first place. Thus, in this manuscript, we describe how this is done by presenting recent work on glycomics data to transition from Web 2.0 to Web 3.0. Figure 1 illustrates the differences between Web 1.0, 2.0 and 3.0.

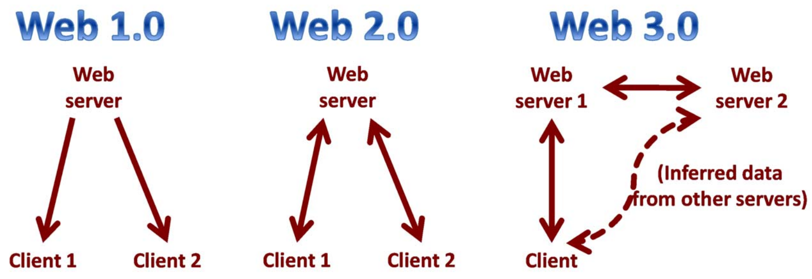

Figure 1. The transition from Web 1.0 to 2.0 to 3.0, where more interaction and eventually computational inference between data on the Web has become possible.

Background

The Semantic Web is made possible by the fact that all data is linked to one another. This network of linked data is part of what is known as the Linked Open Data (LOD), which is a network of all linked data centered on DBpedia and linked to a large variety of fields. A figure of the range of fields on the LOD can be seen at http://lod-cloud.net/. In order to add semantics to linked data, the concept of Resource Description Framework (RDF) is used, which defines data in terms of *triples* consisting of subject, predicate and object. Subjects and objects can be URLs (universal resource locators) or literals (strings). As an example, we can take the subject of the 3rd Beilstein Symposium on Glyco-Bioinformatics and define some triples about it.

Subject: 3rd Beilstein Symposium on Glyco-Bioinformatics

–Has_Theme: Discovering the Subtleties of Sugars

–Dates_Held: June 10 – 14, 2013

–At_Location: Potsdam, Germany

–Sponsor: Beilstein-Institut

Here, Has_Theme, Dates_Held, At_Location and Sponsor are predicates and "Discovering the Subtleties of Sugars," "June 10 – 14, 2013," "Potsdam, Germany" and "Beilstein-Institut" are their respective objects. We can further define more triples on the subject of Potsdam, as follows.

Subject: Potsdam, Germany

–In_State: Brandenburg

–Has_Population: 158,902

–Has_Website: http://www.potsdam.de

Note that Potsdam in the previous subject can then be linked to this subject, as illustrated in Figure 2.

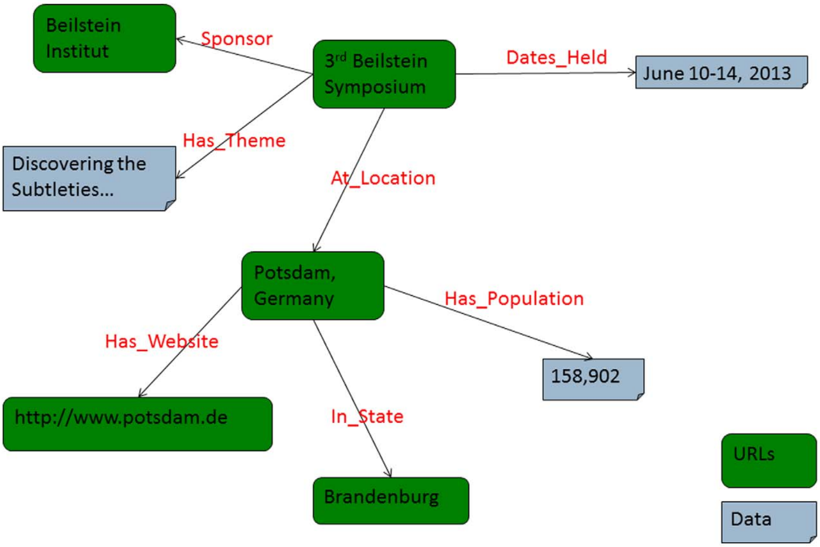

Figure 2. A network graph illustrating the linked relationships between the triples described as examples in the text.

Predicates assign meaning to data, and they are usually defined in an *ontology*. An ontology is traditionally defined as “a systematic account of existence”, and more recently has taken on the following definition: “the hierarchical structuring of knowledge about things by subcategorizing them according to their essential qualities.” The Gene Ontology (GO) is a well-known ontology of hierarchically organized terms that can be used to annotation gene information [2]. Using GO, genes can be annotated with consistent terms, making it easier for users to search for genes with similar function or cellular localization, etc. The Semantic Web relies heavily on ontologies in order to add meaning to data and their relationships in a standardized manner.

Once data is annotated with triples and stored in RDF format, the data are entered into a special RDF database called a triplestore and made available to queries at a special URI which is called a SPARQL endpoint. The data can then be queried using the SPARQL query language for RDF data. Once the data is stored in a triplestore, software can be used to make inferences on the data. For example, using the example described previously, computers may make an inference on the Beilstein Symposium as follows: "The 3rd Beilstein Symposium on Glyco-Bioinformatics will be held in the German state of Brandenburg." In this example, note that the data never directly linked the information about the 3rd Beilstein Symposium with Brandenburg. However, the computer can make the inference through the direct links with Potsdam. To extend this example in terms of glycosciences, glycan structures, for example, can be linked with information about the related glycoproteins, glycolipids, diseases, gene information, etc. Then computers would be able to make inferences and provide clues to researchers about glycan function through these links. Such work is currently done by all researchers when surveying the literature about their target glycan or protein of interest. With the aid of the Semantic Web, however, the time and effort required for such literature surveys can be greatly decreased.

Next, we focus on the technical aspects that must be addressed in order to transfer the glycomics data stored in the currently publicly available glyco-databases onto the Semantic Web. Databases such as JCGGDB, GlycomeDB, BCSDB, UniCarbKB and MonosaccharideDB were represented by the corresponding database developers, who gathered at BioHackathons held in Japan and China in 2012 and 2013, respectively. We will describe the results of these hackathons in this manuscript.

Methods

GlycoRDF

The National Bioscience Database Center (NBDC) and Database Center for Life Science (DBCLS) in Japan have held BioHackathons annually since 2008. BioHackathon stands for Biology + Hacking + Marathon, where programmers come together to develop software together for a few days (usually a week). The 5th Annual DBCLS BioHackathon was held in Toyama city, Japan, from September 2 – 7, 2012. Developers of major glycan databases worldwide gathered together in Toyama to learn about the Semantic Web and develop RDF versions of their respective databases. The databases involved are GlycomeDB [3], UniCarbKB [4], Japan Consortium for Glycobiology and Glycotechnology DataBase (JCGGDB) (http://jcggdb.jp/index_en.html), UniProt, GlycoEpitope (http://www.glyco.is.ritsumei.ac.jp/epitope2/), MonosaccharideDB (http://www.monosaccharidedb.org) and Bacterial Carbohydrate Structure Database (BCSDB) [5]. Their main focus at the BioHackathon was to ensure consistency between the predicates and the links used in each RDF data set.

SPARQL

After preliminary RDF data were generated from each participant, all of the data was stored in a local triplestore so as to test SPARQL queries on the integrated data. Several queries were tested, which are summarized in Table 1.

Table 1. List of queries tested at the 5th BioHackathon to assess useability of RDF on glycoscience data.

	Description	**Databases Involved**	**Input: Output**
Query 1	Obtain UniProt protein IDs for glycan structures in JCGGDB.	JCGGDB, GlycomeDB, UniCarbKB, UniProt	JCGGDB ID: UniProt ID
Query 2	Retrieve the glycan structures involved with lectins across all databases.	Lectin Frontier Database (LfDB) of JCGGDB, GlycomeDB	Lectin data: Glycan structure data
Query 3	Search for the carrier proteins of glycan epitopes.	GlycoEpitope, UniProt, GlycoProtDB (Kaji, et al., 2012) of JCGGDB	GlycoEpitope data: NCBI RefSeq IDs referenced from GlycoProtDB

Results

Each of the queries tested during the 5th BioHackathon could be successfully carried out on our preliminary RDF data set. Although each query involved multiple databases, a single SPARQL query could be written to obtain the requested data in one run. That is, each query could be made with one SPARQL statement, whereas users would be required to access several different databases to obtain the same information over the traditional Web. Thus, the power of the Semantic Web could be demonstrated. As a result, the glyco-database developers could also get a better understanding of what is involved in developing data for the Semantic Web. Detailed results have been published in [6].

The results of the 5th BioHackathon enabled glyco-database developers to get an idea for the potential of the Semantic Web. However, the development process of RDF data generation from each database and generating SPARQL queries also showed us the importance of a standardizing ontology of glyco-data. In particular, the Semantic Web enables computers to make inferences on the RDF data, which requires a consistent ontology to be defined in the first place. Thus, the developers at the 5th Hackathon decided to gather again at a glyco-hackathon, which was held during the GLYCO22 meeting in Dalian, China, June 23 – 28, 2013 [7]. This meeting was sponsored by the Advanced Institute for Science and Technology (AIST) and was focused on the development of a standardized ontology for glycan structures and related data such as publications, experimental procedures and biological samples. The results of this work will allow database developers to generate RDF using consistent predicates on the appropriate subject classes and for the relevant object classes. Moreover, recently the EBI has announced the availability of their life science data on the Semantic Web [8]. Thus, in addition to UniProt, more biological data are now available to be linked.

Conclusion and Future Perspectives

The next steps will include the development of more intuitive user interfaces to query and make inferences on the glycomics and related life science data that are slowly become a part of the Semantic Web. As one of the informatics aims proposed in the NAS report on glycosciences [9], the next five and ten years will see an increased availability of glycomics data on the web along with better computational tools. It can be expected that the Semantics Web technologies will serve as key contributors to this progress.

Acknowledgements

This manuscript presents work that was made possible by the glyco-database developers who participated in the BioHackathons in Toyama and Dalian: Jerven Bolleman (UniProt), Matthew P. Campbell (UniCarbKB), Shin Kawano and Jin-Dong Kim (DBCLS), Thomas Luetteke (MonosaccharideDB), Shujiro Okuda (GlycoEpitope), Rene Ranzinger (GlycomeDB), Hiro Sawaki, Daisuke Shinmachi and Hisashi Narimatsu (JCGGDB at AIST), and Philip Toukach (BCSDB). This work has also been supported by National Bioscience Database Center (NBDC) of Japan Science and Technology Agency (JST), National Institute of Advanced Industrial Science and Technology (AIST) in Japan, and the Database Center for Life Science (DBCLS) in Japan. The developers also recognize the invaluable contributions from the community and those efforts to curate and share structural and experimental data collections.

References

[1] Berners-Lee, T., Hendler, J. and Lassila, O. (2001) The Semantic Web. *Scientific American*. pp. 29 – 37.

[2] Blake, J.A.*, et al.* (2013) Gene Ontology annotations and resources. *Nucleic Acids Res.* **41**:D530 – 535.
doi: http://dx.doi.org/10.1093/nar/gks1050.

[3] Ranzinger, R.*, et al.* (2011) GlycomeDB – a unified database for carbohydrate structures. *Nucleic Acids Res.* **39**:D373 – 376.
doi: http://dx.doi.org/10.1093/nar/gkq1014.

[4] Campbell, M.P.*, et al.* (2011) UniCarbKB: putting the pieces together for glycomics research. *Proteomics* **11**:4117 – 4121.
doi: http://dx.doi.org/10.1002/pmic.201100302.

[5] Toukach, P., *et al.* (2007) Sharing of worldwide distributed carbohydrate-related digital resources: online connection of the Bacterial Carbohydrate Structure DataBase and GLYCOSCIENCES.de. *Nucleic Acids Res.* **35**:D 280 – 286.
doi: http://dx.doi.org/10.1093/nar/gkl883.

[6] Aoki-Kinoshita, K.F., *et al.* (2013) Introducing glycomics data into the Semantic Web. *Journal of Biomedical Semantics* 2013, **4**:39.
doi: http://dx.doi.org/10.1186/2041-1480-4-39.

[7] Aoki-Kinoshita, K.F., *et al.* (2013) The Fifth ACGG-DB Meeting Report: Towards an international glycan structure repository. *Glycobiology*, **23** (12): 1422 – 1424.
doi: http://dx.doi.org/10.1093/glycob/cwt084.

[8] EMBL-EBI (2013) Bioinformatics embraces Semantic Web technologies.

[9] Committee on Assessing the Importance and Impact of Glycomics and Glycosciences, N.R.C.U. (2012) *Transforming Glycoscience: A Roadmap for the Future.* National Academies Press Washington, DC, USA.