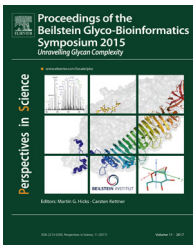
Perspectives in Science (2017) 11, 18—23

Available online at www.sciencedirect.com

ScienceDirect

journal homepage: www.elsevier.com/pisc



Latest developments in Semantic Web technologies applied to the glycosciences☆

Kiyoko F. Aoki-Kinoshita[a,b,*], Nobuyuki P. Aoki[a], Akihiro Fujita[a], Noriaki Fujita[b], Toshisuke Kawasaki[c], Masaaki Matsubara[d], Shujiro Okuda[e], Toshihide Shikanai[b], Daisuke Shinmachi[a], Elena Solovieva[b], Yoshinori Suzuki[b], Shinichiro Tsuchiya[a], Issaku Yamada[d], Hisashi Narimatsu[b]

[a] *Faculty of Science and Engineering, Soka University, Tokyo 192-8577, Japan*
[b] *Glycoscience and Glycotechnology Research Group, AIST, Ibaraki 305-8568, Japan*
[c] *Research Center for Glycobiotechnology, Ritsumeikan University, Kusatsu, Shiga 525-8577, Japan*
[d] *The Noguchi Institute, Tokyo 173-0003, Japan*
[e] *Niigata University Graduate School of Medical and Dental Sciences, Niigata 951-8510, Japan*

Received 4 January 2016; accepted 6 May 2016
Available online 20 October 2016

KEYWORDS
Semantic Web;
Glycan repository;
Glycan text representation

Summary The Integrated Life Science Database Project of Japan funded a group of glycoscientists to carry out a project to integrate glycoscience databases using Semantic Web technologies. As a continuation of the previous project period, the Japan Consortium for Glycobiology and Glycotechnology Database (JCGGDB) developed several glycoscience-related databases. The GlycoProtDB database is among those being integrated, providing an important resource to understand protein glycosylation. Another database being integrated is GlycoEpitope, a comprehensive database of carbohydrate epitopes and antibodies. In the current project period, we started the development of GlyTouCan, the international glycan structure repository providing unique accession numbers to all glycan structures. Although such databases are sufficiently important in and of themselves, their integration with other—omics data such as the protein information in UniProt will be crucial to bring glycosciences to the forefront of life sciences. However, to integrate such disparate sets of data among different fields in a way such that future maintenance costs are minimal, standardized ontologies and formats must be established.

* Corresponding author at: 1-236 Tangi-machi, Hachioji, Tokyo, Japan 192-8577, Faculty of Science and Engineering, Soka University, Tokyo 192-8577, Japan.
E-mail address: kkiyoko@soka.ac.jp (K.F. Aoki-Kinoshita).

http://dx.doi.org/10.1016/j.pisc.2016.05.012

Our latest project has attempted to define the minimal standards that are necessary to enable this integration. The technical challenges to integrate all these databases and the technologies to overcome these challenges will be described.

Introduction

Integrated Life Science Database Project

The purpose of the Integrated Life Science Database Project of Japan (http://biosciencedbc.jp/en/), sponsored by the Japan Science and Technology Agency (JST), is to integrate all life science databases both within and outside Japan. It is apparent that a plethora of data is currently available on the Internet, and it can be bewildering for students and newcomers who come from different life science fields. One of the problems with the development of life science databases is that the data are often left around or taken down after publication or the developer/student has left the organization/university. Therefore, JST had specifically included a requirement in this latest project; all funded database projects must be sustainable even after funding ends. As a result, the currently funded projects are now being developed in a way that the data is available and all developed source code is open source.

In terms of technical sustainability, considering that the previous project focused initially on web services as the main technology, it was found that web services were still limited due to the following:

- Each web service can implement only one query, having a specific input, predefined parameters and a specific output.
- Changing the data often forces changes in the web service implementation.
- Documentation of each and every web service was cumbersome, but required for others to use them.
- Different web service application programming interfaces (APIs) needed to be developed for different programming languages (although Representational State Transfer (REST;) was becoming mainstream).
- Integration with other databases required knowledge of other databases' APIs.

Semantic Web technology

Consequently, participants started looking into the Semantic Web as a means to integrate data among different databases. In contrast to web services, it was found to be simpler to implement using the Resource Description Framework (RDF), which consisted of triples of data (**subject**, **predicate** and **object**) where the **subject** and **objects** could be either a literal (i.e., a text string or numeric, for example) or Uniform Resource Identifiers (URIs) that pointed to data on the Internet. The **predicate** performs an important function as it adds the semantics to the data. **Predicates** are defined by an *ontology* which specifies the relationship between different classes of data. For example, if a particular protein X is glycosylated at position P by an *N*-linked glycan, then, the ontology would need to specify classes of data called *proteins* and *glycans*. Moreover, it would need to define the concept of *glycosylation* as a **predicate** that takes *proteins* as **subjects** and *glycans* as **objects**. Furthermore, the glycosylation position would need to be encapsulated by another **predicate** *glycosylated_at* whose **objects** are classes of data called *aminoacids*, for example. The *aminoacid* class could also serve as a **subject** of other **predicates**, such as *positionnumber* and *aminoacidtype*.

Note that such rules to specify the semantics of data are all stored in an ontology that must be consistently used by data providers for the data to be accurately linked to each other. If one database provider uses *glycosylatedAt* and another uses *glycosylated_at*, then these would be considered different predicates, and their meanings would be considered different from one another.

As a result, as long as a standardized ontology could be developed, data no longer needed to be maintained in relational databases and could be instead provided as URIs, which meant that data could be modified as needed as long as it could be accessed from the same URI consistently (Aoki-Kinoshita et al., 2015b). Thus, as a part of the Integrated Life Science Database Project, the Glycoscience Team also started looking into the Semantic Web and RDFizing existing databases, and we were able to show a proof-of-concept that it was possible to do so with minimal effort (Aoki-Kinoshita et al., 2013a; Katayama et al., 2014). From this, a new glycan ontology was developed, called GlycoRDF (Ranzinger et al., 2015), which is now used by the Carbohydrate Structure Database (Toukach and Egorova, 2015), MonosaccharideDB, UniCarbKB (Campbell et al., 2014), Glycoepitope, GlycoProtDB (Kaji et al., 2012) and GlycomeDB (Ranzinger et al., 2009). By using this ontology, it is now possible to integrate these different databases simply by referencing the URIs of other database entries using RDF.

To implement such integration, data must be stored in a specialized database for RDF data, called a triplestore. In the current work, Virtuoso is used as the triplestore data management system (Erling and Mikhailov, 2009).

Glycoscience Team project

The Glycoscience Team of the Integrated Database Project was granted funds to develop an International Glycan Structure Repository which we call GlyTouCan based on WURCS (Web 3.0 Unique Representation of Carbohydrate Structures), a new text representation of glycan structures. These two subprojects started in 2014 in addition to the