Bioinformatics, 31(6), 2015, 919–925
doi: 10.1093/bioinformatics/btu732
Advance Access Publication Date: 11 November 2014
Original Paper

OXFORD

Databases and ontologies

GlycoRDF: an ontology to standardize glycomics data in RDF

Rene Ranzinger[1,*], Kiyoko F. Aoki-Kinoshita[2,3,*], Matthew P. Campbell[4], Shin Kawano[5], Thomas Lütteke[6], Shujiro Okuda[7], Daisuke Shinmachi[3], Toshihide Shikanai[2], Hiromichi Sawaki[2], Philip Toukach[8], Masaaki Matsubara[9], Issaku Yamada[9] and Hisashi Narimatsu[2]

[1]Complex Carbohydrate Research Center, University of Georgia, Athens, GA, USA, [2]Research Center for Medical Glycoscience, National Institute of Advanced Industrial Science and Technology, Tsukuba, Japan, [3]Faculty of Engineering, Soka University, Tokyo, Japan, [4]Biomolecular Frontiers Research Centre, Macquarie University, Sydney, Australia, [5]Database Center for Life Science, Research Organization of Information and Systems, Chiba, Japan, [6]Institute of Veterinary Physiology and Biochemistry, Justus-Liebig-University Giessen, Giessen, Germany, [7]Graduate School of Medical and Dental Sciences, Niigata University, Niigata, Japan, [8]N.D. Zelinsky Institute of Organic Chemistry, Russian Academy of Sciences, Moscow, Russia and [9]Laboratory of Glyco-organic Chemistry, The Noguchi Institute, Tokyo, Japan

*To whom correspondence should be addressed.

Associate Editor: Jonathan Wren

Received on August 28, 2014; revised on October 12, 2014; accepted on October 28, 2014

Abstract

Motivation: Over the last decades several glycomics-based bioinformatics resources and databases have been created and released to the public. Unfortunately, there is no common standard in the representation of the stored information or a common machine-readable interface allowing bioinformatics groups to easily extract and cross-reference the stored information.

Results: An international group of bioinformatics experts in the field of glycomics have worked together to create a standard Resource Description Framework (RDF) representation for glycomics data, focused on glycan sequences and related biological source, publications and experimental data. This RDF standard is defined by the GlycoRDF ontology and will be used by database providers to generate common machine-readable exports of the data stored in their databases.

Availability and implementation: The ontology, supporting documentation and source code used by database providers to generate standardized RDF are available online (http://www.glycoinfo.org/GlycoRDF/).

Contact: rene@ccrc.uga.edu or kkiyoko@soka.ac.jp

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Over the last decade the growth in publicly accessible data collections has steadily increased in complexity and coverage in all life science fields, including glycomics. However, compared with other fields, the content of the glycomics databases and the representations of glycan sequences used in these databases remain diverse and are not standardized (Campbell *et al.*, 2014a). The discontinuation of CarbBank (Doubet and Albersheim, 1992), the first publicly available carbohydrate structure database (CSDB), in the mid-1990s led to the development of a number of independent databases and carbohydrate-related web resources. Many have been developed by individual research groups and few are collaborative projects, all providing their own unique datasets and functionalities as reviewed by Aoki-Kinoshita *et al.* (2013a, b), Campbell *et al.* (2014a), Frank

and Schloissnig (2010) and Lütteke (2012). GLYCOSCIENCES.de (Lütteke *et al*., 2006), one of the earliest web portals for glycomics data, developed at the German Cancer Research Center was seeded from the efforts of CarbBank and is now focused on the three-dimensional conformations of carbohydrates. In the early 2000s, KEGG GLYCAN (Hashimoto *et al*., 2006) was added to the KEGG resources (Kanehisa *et al*., 2012) as a new glycan structure database that is linked to genomic and pathway information. Around the same time, the Consortium for Functional Glycomics developed a glycan structure database to support their glycan array, mass spectrometry profiling, glyco-gene knock-out mouse and glyco-gene microarray data resources (Raman *et al*., 2006). In Russia, the bacterial (Toukach, 2011) and plant and fungal (Egorova and Toukach, 2014) CSDBs were developed, containing carbohydrate structures and nuclear magnetic resonance (NMR) data from bacterial, plant and fungal species collected from the scientific literature. Meanwhile in Japan, the Japan Consortium for Glycobiology and Glycotechnology (JCGG) developed a web portal and databases to integrate all national carbohydrate resources (http://jcggdb.jp/index_en.html). One of the major databases incorporated into JCGG database (JCGGDB) is GlycoEpitope (Kawasaki *et al*., 2006) (http://www.glycoepitope.jp), which is a database of carbohydrate epitopes and antibodies, accumulated from the literature. JCGGDB also has released a glycoprotein database called GlycoProtDB (Kaji *et al*., 2012), which contains experimentally verified glycosylation site information. In 2005, the European Union funded EUROCarbDB, a design study focused on developing a framework for storing and sharing experimental data of carbohydrates (von der Lieth *et al*., 2011). An integrated module of EUROCarbDB was MonosaccharideDB (http://www.monosaccharidedb.org), which provides a controlled vocabulary of unique monosaccharide residues. The dynamic, self-extensible dictionary and translation routines of MonosaccharideDB ensure monosaccharide name consistency and allow translation and mapping of monosaccharide names used in different resources. After the end of the EUROCarbDB project, the data resources and software, which are all available as open source software, were taken on by the UniCarbKB project (Campbell *et al*., 2011, 2014b). GlycoSuiteDB (Cooper *et al*., 2001, 2003), originally developed as a commercial database in Australia, contains published glycoprotein glycan structures together with information on the glycoproteins and further knowledge of the biological context in which the glycans have been identified. GlycoSuiteDB has recently become a part of the ExPASy portal and is now integrated into UniCarbKB. In addition, several other small databases used in local laboratories have been developed in parallel providing overlapping or complementary information.

Although all these databases contain valuable data, a lack of interoperability hampers development of data mash-up applications. Despite some efforts to standardize and exchange their data (Packer *et al*., 2008; Toukach *et al*., 2007), most glycomics databases still have to be regarded as 'disconnected islands' (Lütteke, 2008). The standardization of carbohydrate primary structures is more difficult than in genomics or proteomics, partly due to the inherent structural complexity of saccharides exemplified by complex branching, glycosidic linkages, anomericity and residue modifications. Individual databases developed their own formats to cope with these problems and encode glycan primary structures in machine-readable ways (Aoki-Kinoshita *et al*., 2013a, b) creating a large collection of incompatible glycan sequence formats. GlycomeDB was designed to integrate all glycan structures and species information from most of these databases by converting them into one consistent representation and providing a single web portal allowing for searching across all these databases for particular structures (Ranzinger *et al*., 2008, 2009). Using the cross-references to the original databases provided for each structure, it became possible to access the web pages of these databases and to retrieve the provided information related to a glycan. All databases provide their information using web pages, restricting the query possibilities to the limited search options provided by the developers. Only a few provide web service interfaces allowing retrieval of the data in a machine-readable non-Hypertext Markup Language format. The few web service interfaces that have been implemented return proprietary non-standard formats making it hard to access and integrate data from several resources into a single result.

1.1 Integration and sharing of knowledge

The combination and integration of information from several databases are necessary for a better understanding of the biological processes in which glycans are involved. This requires access not just to the primary structures but also to associated data such as biological contexts where glycans have been found (including information on the proteins to which glycans are linked), specification of glycan-binding proteins and references to publications or experimental data. However, this information is still spread over the various resources, which are (e.g. in the context of proteins) not limited to only glyco-related databases, and stored in different formats and representation schemes. A model that enables the integration of these diverse data will permit users to answer more complex questions compared with individual database queries or cross-linking primary structures would allow. Connecting glycomics resources with other life science data also enables the integration of glycan information into systems biology approaches.

The premise for integrating data from multiple heterogeneous sources is not new. The review by Stein (2003), although focused on genomics, identified that biological databases have been invaluable for managing data collections but 'so far their integration has proved problematic'. A series of steps to enable data integration was introduced by Davidson *et al*. (1995) that included the following: transformation of existing datasets into a common data model; matching semantically related data objects and transformation of data into a federated database on demand.

1.2 Semantic technologies for data integration

The Semantic Web Community has successfully implemented data exchange and integration solutions based on a series of standards: the Resource Description Framework (RDF) format (http://www.w3.org/TR/rdf-concepts/) for documents and the Web Ontology Language (OWL) for ontology specification (Belleau *et al*., 2008; Jupp *et al*., 2014). RDF is a manner of describing concepts as resources and their linking relations. In particular, RDF defines data using triples, which consist of three components, subject, predicate and object, to describe relationships between pieces of information. Subjects and objects can be Uniform Resource Identifiers (URIs), such that they can be linked across the Semantic Web. Moreover, objects can serve as subjects to other objects, thus forming a graph, or network, of triples. Ontologies are further used in RDF to define classes, which describe the subjects and objects being represented. OWL is a manner of describing relations in a formal logic that allows not only specifying the classes and predicates to be used and their relationships but also for checking integrity and data inconsistencies. An OWL ontology can be defined as an explicit specification of a conceptualization, which is an abstract and simplified view of the world that needs to be represented for some purpose (Gruber, 1995). For a given knowledge base or knowledge system, it means

that a conceptual language should be used to define the objects and the relations to be represented. The conversion of data from proprietary formats or relational databases to RDF for the purpose of providing a machine-readable dataset usable for the Semantic Web is a process commonly called 'RDFization'.

Recently, the European Bioinformatics Institute (EBI) launched the EBI RDF platform to coordinate and connect the vast data collections accumulated across the institute (Jupp *et al.*, 2014). The platform provides an innovative approach to query and explore rich biological data collections, allowing researchers and developers to execute searches that were previously not possible. This effort clearly demonstrates the usefulness and power of semantics technologies.

To make use of the advantages of the Semantic Web and the possibilities provided by SPARQL (SPARQL Protocol and RDF Query Language) for querying different glycomics databases, we have commenced translating databases into RDF and demonstrated the usefulness of this approach (Aoki-Kinoshita *et al.*, 2013a, b). To effectively provide diverse glycomics data types including glycan sequences, biological source information, literature references and experimental data, a common RDF standard is required. Here, we present our efforts to develop an ontology for generating standardized RDF for glycan structures and related data (known as GlycoRDF), which has been agreed and adopted by most database designers and developers in the glycomics discipline.

2 GlycoRDF ontology

GlycoRDF is an ontology to define the RDF namespace, concepts and relationships between concepts to be used for exporting glycomics data into a standardized representation using RDF. The ontology contains classes and predicates necessary for the diverse data types used in glycomics databases, and also reuses concepts from other well-established ontologies, such as UniProt core (UniProtConsortium, 2012), the Bibliographic Ontology, Dublin Core Metadata Initiative (DCMI) Metadata Terms and Human Proteome Organization Proteomics Standards Initiative (HUPO-PSI) Mass Spectrometry Ontology (PSI-MS) (Mayer *et al.*, 2013). The objective of GlycoRDF is to minimize the development of multiple RDF dialects that complicate the querying and mash-up of the information across several resources. The development of a standard format for glycomics information needs to address the diverse information in the glycomics databases, which ranges from simple publications that are attached to a glycan structure to detailed biological source information and experimental datasets. To capture all this information, a broad ontology design is required which also has to be extensible in the future to capture further information. Figure 1 shows a simplified unified modeling language diagram of the major concepts and classes in GlycoRDF. RDF examples demonstrating the utility of the classes and properties, described in detail below, are provided in the accompanying complementary material with supporting ontology documentation.

2.1 Molecules—the Compound class

The *Compound* class represents biological molecules and contains information related to the chemical properties of the molecule (Fig. 1, bottom middle). There are several more specific subclasses for different kinds of molecules, omitted in Figure 1 for simplicity. Glycan structures are represented in most databases by the *Saccharide* class or, more specifically, they can be represented by subclasses of *Saccharide* such as *Nglycan* or *Oglycan*. Each level in the inheritance hierarchy contains additional information compared with the parent class. For example, in addition to the compound properties, each saccharide has a sequence in a glycan-specific format, a monosaccharide composition and can have links to glycan-specific image representations. The sequence representation for glycans may utilize any known sequence format, including GlycoCT (Herget *et al.*, 2008), LinearCode (Banin *et al.*, 2002), International Union of Pure and Applied Chemistry (McNaught, 1997), as well as the recently developed Web3 Unique Representation of Carbohydrate Structures (WURCS) (Tanaka *et al.*, 2014). There are also classes representing glycoconjugate structures [molecules that contain not only the glycan but also other attached molecules (aglycons) such as peptides, proteins or lipids] that are not shown in Figure 1. These glycoconjugates are defined by references to a glycan structure, a reference to the aglycon molecule, such as lipid or protein, and in some cases the attachment position of the glycan. This hierarchy allows one to represent the commonly used glycan and glycoconjugate types, but can also be extended to include molecules not yet supported by this version. Although the predicate *has_identifier* is mandatory, the identifier (ID) used is the identifier of the compound (glycan) in the database generating the RDF. Since each database provider is free to use the sequence format they prefer and there is currently no globally accepted identifier for glycan structures (Aoki-Kinoshita *et al.*, 2013a, b), mapping RDF information from a glycan entry in one database to RDF information of the same glycan entry in another database is not possible using this information. Therefore, each database provides cross-references to other database entries describing the same structure where available. In addition GlycomeDB serves as a provider for cross-references not directly stored in the source database. If the referenced database only provides web pages, the predicate *rdfs:seeAlso* is used. If there is an RDF description available for the glycan in the referenced database, the predicate *owl:sameAs* is used. With this information it is possible to map structures from different databases to each other regardless of the ID and the generated glycan sequence.

2.2 References—the Citation class

The second concept, *Citation* (Fig. 1, middle left), is used to create instances of objects representing literature references, such as articles, book chapters or thesis documents describing a glycan. There are no subclasses and only a few predicates in the GlycoRDF ontology for this class. Instead, many predicates from the DCMI (http://www.dublincore.org) and the Bibliographic Ontology (http://bibliontology.com) have been reused to describe and represent publication information, such as title or authors. It is possible to refer to other resources describing the publication, such as PubMed (Wheeler *et al.*, 2000) or Digital Object Identifier (DOI). These references are optional, since not all publications are present in DOI or PubMed; thus, it is still possible to add citations with a comprehensive list of information and no reference to other sources.

2.3 Origin—the Source class

The third major concept, *Source* (Fig. 1, top middle), describes the origin of a glycan molecule. There are more specific subclasses, two of which are shown in Fig. 1, that have different sets of predicates to represent the source information. If a glycan was found in a sample extracted from a biological organism (*SourceNatural*), the RDF allows the specification of the organism and the sample location. Information that can be provided are the species and species subclass, organ, tissue, fluid, cell type, cell line, strains, life stage and related diseases. In addition, it is possible to express that a sample was derived from an organism (e.g. viruses or bacteria) that was hosted by another organism. The list of predicates was chosen based

on the maximum level of detail that is currently available in glycan databases. Although most of the predicates are part of GlycoRDF, the actual representation of this information is derived from existing dictionaries or ontologies [e.g. UniProt Taxonomy for species information or Medical Subject Headings (Lowe and Barnett, 1994) for tissue]. The class *SourceSynthetic* can be used to represent an artificial origin of a glycan structure, such as a structure that has been chemically synthesized. This class has predicates to store the type of synthesis and a reference to a model organism in case the structure was synthesized after a biological structure. With the two presented subclasses, it is possible to represent most of the source information present in glycan databases. But it can be easily extended by creating additional subclasses if needed.

2.4 Experimental data—the Evidence class

The fourth concept storing glycomics data is *Evidence* (Fig. 1, middle left), which is used for RDFized experimental data present in glycan databases. The list of experimental techniques used to study glycan structures is long and very diverse. However, the current freely available databases only contain NMR, mass spectrometric and Liquid chromatography–mass spectrometry data. As such, we have concentrated on creating subclasses for these techniques: *EvidenceNMR*, *EvidenceLC* and *EvidenceMS* (two of which are shown in Fig. 1). Each of the subclasses has its own set of predicates and concepts defined in GlycoRDF to represent the information produced by these kinds of experiments. To represent the mass spectrometric information, several terms and classes from HUPO PSI-MS have been reused. The list of techniques can easily be extended in the future by creating new subclasses of Evidence and predicates to represent the technique-specific information.

2.5 Bringing the pieces together—the Referenced Compound class

The concepts described are used to store the primary information about glycan structures present in current databases: the glycan molecule (*Compound*), bibliographic references (*Citation*), origin (*Source*) and experimental data (*Evidence*). Since the same glycan can be found in multiple organisms, can be published in multiple papers or can be identified by multiple experiments, it is necessary to group this primary information. Such a grouping allows one to specify which experimental data provide evidence that a structure was found in a certain organism and in which of the papers this statement was published. For this purpose the concept *ReferencedCompound* is used. For each group of primary information an instance of *ReferencedCompound* is created containing data from at least two of the four primary information types. This grouping allows one to easily perform queries such as 'Which experimental data support that glycan X was found in organism Y?' or 'Which papers report the identification of structures containing sialic acid by positive mode electrospray ionization mass spectrometry?'.

In addition to these major concepts, their subclasses, classes for storing meta-information and predicates connecting the different classes, our ontology contains several dictionaries to store commonly used glycomics terms and information. Examples for these dictionaries are classes for absolute and relative monosaccharide configuration, common monosaccharide substituents, graphical glycan representation schemes and glycan sequence formats. By defining these terms in GlycoRDF, they are assigned a unique URI which will be reused by all GlycoRDF providers rather than each provider defining their own URI for these terms and requiring an URI mapping when querying this information.

The presented version of the GlycoRDF ontology is designed to capture information currently available in freely accessible glycomics databases. The structure of the ontology and also the dictionaries can be extended in the future if more information has to be represented and no other ontology can provide the necessary concepts. The documented ontology has been submitted to BioPortal (https://bioportal.bioontology.org/) and a human readable documentation can be found on the project web page (http://www.glycoinfo.org/GlycoRDF/).

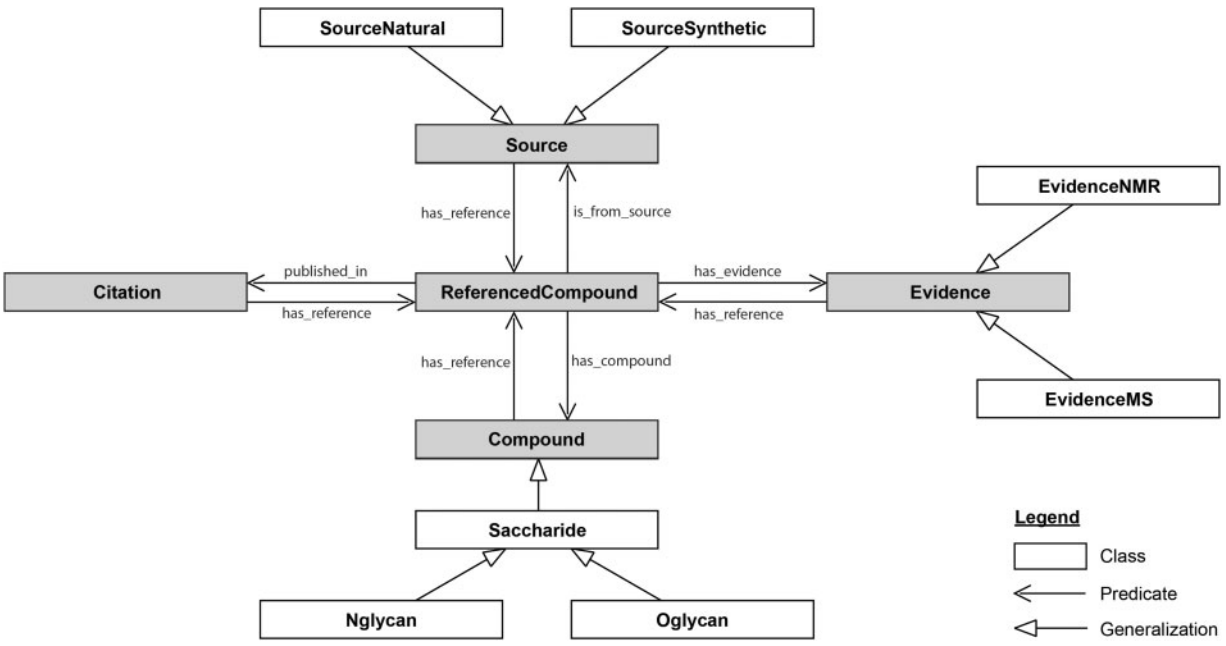

Fig. 1. UML diagram of the core classes in the GlycoRDF ontology. The five central classes (grey boxes) are shown together with their major subclasses (white boxes) and the predicates connecting those (arrows with arrowhead pointing towards the object of a RDF triple)

3 Implementation

Each of the glycan databases participating in this initiative already had existing tool chains and infrastructure for storing, managing and exporting data in place. By extending the available tools and programs, the content of the databases will be translated into RDF matching the data structure as defined by the GlycoRDF ontology to produce standard RDF for the databases. This RDFization process is unique for each resource since all databases use their own proprietary data representations, data structures and associated metadata attached to the glycan structures. At this stage, developers and database providers from several glycomics activities [including CSDB, MonosaccharideDB, GlycomeDB, UniCarbKB, GlycoEpitope, GlycoNAVI (http://www.glyconavi.org) and GlycoProtDB] have converted their database content into the RDF standard, which is publicly accessible (Table 1). An up-to-date list of the databases and links to documentation and the generated RDF can be found on the project web page (http://www.glycoinfo.org/GlycoRDF/). The access methods, which are described in detail in the specified web pages, vary from statically generated RDF files to web services that generate the RDF on the fly out of the current database content. Most resources provide the glycan structure (internal ID and sequence in one of the commonly used sequence formats), monosaccharide composition of the glycan and biological source information. Depending on the database specialization, other informations such as experimental data, literature references, images of the glycan in different notations and image formats, cross-references to other databases and information about the attached protein can be found in the provided RDF files. In Figure 2, we show an example of RDF triples representing a glycan and its sequence representation, the monosaccharide composition and the association of the glycan with a publication using the *ReferencedCompound* class. First, a glycan structure (GlycomeDB ID 773) represented in LInear Notation for Unique description of Carbohydrate Sequences (LINUCS) format may be described as shown in Figure 2A. Consequently, the three components, each consisting of a monosaccharide and its cardinality in the structure, composing this structure can be defined. An example of one of these components (the two alpha-mannoses residues) is shown in Figure 2B. To indicate the publication that references this structure, a *ReferencedCompound* is defined to refer to both this glycan and the publication (Fig. 2C). The details of the publication would then be provided as a triple with the URI http://rdf.publication/org/example/773 as the subject. At the same time, if the biological source and experimental details and data are known and available, they can all be specified with the same *ReferencedCompound* as subject. Further detailed examples of RDF data including an example for a citation and NMR data are provided in the Supplementary Materials.

Table 1. A list of GlycoRDF partners and available datasets with links to supporting documentation

Database	RDF documentation
CSDB	http://csdb.glycoscience.ru/bacterial/core/help.php?db=bacterial&topic=extras#rdf
GlycomeDB	https://github.com/ReneRanzinger/GlycoRDF/wiki/GlycomeDB-documentation
MonosaccharideDB	http://www.monosaccharidedb.org/rdf/about.action
UniCarbKB	https://github.com/ReneRanzinger/GlycoRDF/wiki/UniCarbKB—RDF
GlycoEpitope	https://github.com/ReneRanzinger/GlycoRDF/wiki/GlycoEpitope-documentation
GlycoNAVI	http://ws.glyconavi.org/WebTool/rdf.php
GlycoProtDB	https://github.com/ReneRanzinger/GlycoRDF/wiki/GlycoProtDB-documentation

Note: For further information, links to example datasets and SPARQL queries refer to the project wiki webpage (https://github.com/ReneRanzinger/GlycoRDF/wiki).

4 Conclusion

One of the greatest problems that the bioinformatics community faces in glycomics is that information about glycans is distributed across several databases with all of these databases having partially overlapping, partially complementary data content but use different representation formats for the information. In addition, most database providers only present their data in the form of web pages without providing machine-readable interfaces, such as web services. Currently, collecting comprehensive information about a single glycan structure is a manual and cumbersome procedure. By translating the available information into RDF, it becomes possible to bypass the limited web interfaces and access the database content in a format that can be easily processed by computer programs. To avoid duplicated effort and the generation of several different RDF representations of

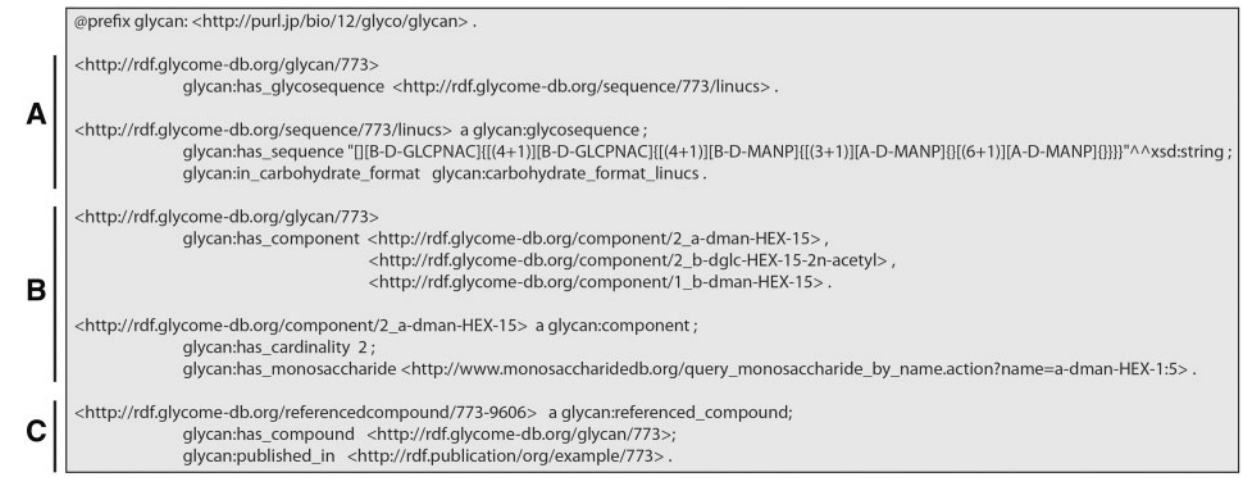

Fig. 2. RDF example for the encoding of (A) a glycan and its LINUCS representation, (B) the monosaccharide composition of the glycan, and (C) usage of a *ReferencedCompound* to link the glycan with a publication

glycans and glycan-related information by different database providers, we have defined a common RDF standard that many of the glycomics database providers have agreed upon. By supporting a generally accepted format defined in the GlycoRDF ontology, querying the data will be greatly simplified since all glycan-related information will be represented by the same predicates and terms, making it unnecessary to map between different RDF namespaces when accessing information from several resources. The GlycoRDF architecture is amenable to the inclusion of new features and relevant extensions; therefore, it is intended to be manageable for long-term use by being flexible enough to cater for future trends and meta-data descriptions.

GlycoRDF is a future-thinking collaborative effort that addresses the requirement for sophisticated data mash-ups to answer questions by combining data from different resources. The standard representation will connect glycan-related databases and resources ultimately providing a platform that enhances data discovery in glycomics. As described in the implementation section, several glycomics database providers have adopted the RDF standard and now provide their data in this format. The RDF files can be used by researchers to access database content and to use the stored information for machine learning or create data mash-up applications by interfacing with other databases, such as protein information from UniProt. At the same time, the provided web services can be used for live data requests from the databases for the purpose of data processing or enrichment of glycan data with additional information.

There are three major objectives that arise from the described work. The first is to expand the list of database providers to RDFize more glycan-related data, and to make these data freely available for data analysis. The second objective is to link and combine the glycan-related information with information from other disciplines such as proteomics and genomics. This will allow researchers to perform queries that are beyond the scope of the databases of a single domain. For example, researchers will be able to perform potential queries such as 'Find gene-related information of enzymes that took part in the generation of a glycan structure' or 'Mash-up protein-related information with the information of glycans that are attached to the protein'. Future glycomics and bioinformatics meetings will be the perfect stage to present this work and invite researchers to join this effort. The third objective is the provision of easy-to-use querying interface through a web interface rather than machine-readable interfaces designed for computer scientists. Currently, there are discussions to set up triplestores with SPARQL endpoints that will allow direct querying of the information without downloading and local handling of the RDF files. It will also allow for remote querying and mash-up of the information content of several databases using federated queries.

Acknowledgements

The GlycoRDF project was made possible by the generous contributions by the Japanese BioHackathon workshop held in Toyama (Japan) in 2012 and an international Glyco-BioHackathon workshop in Dalian (China) in 2013. The authors are grateful for organization of these meetings by the National Bioscience Database Center of Japan Science and Technology Agency, the Database Center for Life Science of Research Organization of Information and Systems, National Institute of Advanced Industrial Science and Technology and the JCGGDB project. We acknowledge Jerven Bolleman (Swiss Institute of Bioinformatics) for his valuable input and guidance in the development of GlycoRDF.

Funding

This work was supported by the National Institute of General Medical Sciences (8P41GM103490 to R.R.); National Bioscience Database Center of Japan Science and Technology Agency (Integrated Database Project to H.N.); Russian Foundation for Basic Research (N12-04-00324 to P.T.); and Australian National eResearch Collaboration Tools and Resources (RT016 to M.C.).

Conflict of Interest: none declared.

References

Aoki-Kinoshita,K.F. *et al.* (2013a) Introducing glycomics data into the Semantic Web. *J. Biomed. Semantics*, **4**, 39.

Aoki-Kinoshita,K.F. *et al.* (2013b) The fifth ACGG-DB meeting report: towards an international glycan structure repository. *Glycobiology*, **23**, 1422–1424.

Banin,E. *et al.* (2002) A novel Linear Code(R) nomenclature for complex carbohydrates. *Trends Glycosci. Glycotechnol.*, **14**, 127–137.

Belleau,F. *et al.* (2008) Bio2RDF: towards a mashup to build bioinformatics knowledge systems. *J. Biomed. Inform.*, **41**, 706–716.

Campbell,M.P. *et al.* (2011) UniCarbKB: putting the pieces together for glycomics research. *Proteomics*, **11**, 4117–4121.

Campbell,M.P. *et al.* (2014a) Toolboxes for a standardised and systematic study of glycans. *BMC Bioinformatics*, **15** (Suppl. 1), S9.

Campbell,M.P. *et al.* (2014b) UniCarbKB: building a knowledge platform for glycoproteomics. *Nucleic Acids Res.*, **42**, D215–D221.

Cooper,C.A. *et al.* (2001) GlycoSuiteDB: a new curated relational database of glycoprotein glycan structures and their biological sources. *Nucleic Acids Res.*, **29**, 332–335.

Cooper,C.A. *et al.* (2003) GlycoSuiteDB: a curated relational database of glycoprotein glycan structures and their biological sources. 2003 update. *Nucleic Acids Res.*, **31**, 511–513.

Davidson,S.B. *et al.* (1995) Challenges in integrating biological data sources. *J. Comput. Biol.*, **2**, 557–572.

Doubet,S. and Albersheim,P. (1992) CarbBank. *Glycobiology*, **2**, 505.

Egorova,K.S. and Toukach,P.V. (2014) Expansion of coverage of Carbohydrate Structure Database (CSDB). *Carbohydr. Res.*, **389**, 112–114.

Frank,M. and Schloissnig,S. (2010) Bioinformatics and molecular modeling in glycobiology. *Cell. Mol. Life Sci.*, **67**, 2749–2772.

Gruber,T.R. (1995) Toward principles for the design of ontologies used for knowledge sharing?. *Int. J. Hum. Comput. Stud.*, **43**, 907–928.

Hashimoto,K. *et al.* (2006) KEGG as a glycome informatics resource. *Glycobiology*, **16**, 63R–70R.

Herget,S. *et al.* (2008) GlycoCT—a unifying sequence format for carbohydrates. *Carbohydr. Res.*, **343**, 2162–2171.

Jupp,S. *et al.* (2014) The EBI RDF platform: linked open data for the life sciences. *Bioinformatics*, **30**, 1338–1339.

Kaji,H. *et al.* (2012) Large-scale identification of N-glycosylated proteins of mouse tissues and construction of a glycoprotein database, GlycoProtDB. *J. Proteome Res.*, **11**, 4553–4566.

Kanehisa,M. *et al.* (2012) KEGG for integration and interpretation of large-scale molecular data sets. *Nucleic Acids Res.*, **40**, D109–D114.

Kawasaki,T. *et al.* (2006) GlycoEpitope: the integrated database of carbohydrate antigens and antibodies. *Trends in Glycosci. Glycotechnol.*, **18**, 267–272.

Lowe,H.J. and Barnett,G.O. (1994) Understanding and using the medical subject headings (MeSH) vocabulary to perform literature searches. *J. Am. Med. Assoc.*, **271**, 1103–1108.

Lütteke,T. *et al.* (2006) GLYCOSCIENCES.de: an Internet portal to support glycomics and glycobiology research. *Glycobiology*, **16**, 71R–81R.

Lütteke,T. (2008) Web resources for the glycoscientist. *Chembiochem*, **9**, 2155–2160.

Lütteke,T. (2012) The use of glycoinformatics in glycochemistry. *Beilstein J. Org. Chem.*, **8**, 915–929.

Mayer,G. *et al.* (2013) The HUPO proteomics standards initiative—mass spectrometry controlled vocabulary. *Database*, **2013**, bat009.

McNaught,A.D. (1997) Nomenclature of carbohydrates (recommendations 1996). *Adv. Carbohydr. Chem. Biochem.*, **52**, 43–177.

Packer,N.H. *et al.* (2008) Frontiers in glycomics: bioinformatics and biomarkers in disease. An NIH white paper prepared from discussions by the

focus groups at a workshop on the NIH campus, Bethesda MD (September 11-13, 2006). *Proteomics*, **8**, 8–20.

Raman,R. *et al.* (2006) Advancing glycomics: implementation strategies at the consortium for functional glycomics. *Glycobiology*, **16**, 82R–90R.

Ranzinger,R. *et al.* (2008) GlycomeDB—integration of open-access carbohydrate structure databases. *BMC Bioinformatics*, **9**, 384.

Ranzinger,R. *et al.* (2009) Glycome-DB.org: a portal for querying across the digital world of carbohydrate sequences. *Glycobiology*, **19**, 1563–1567.

Stein,L.D. (2003) Integrating biological databases, *Nat. Rev. Genet.*, **4**, 337–345.

Tanaka,K. *et al.* (2014) WURCS: the Web3 unique representation of carbohydrate structures. *J. Chem. Inf. Model.*, **54**, 1558–1566.

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**, D280–D286.

Toukach,P.V. (2011) Bacterial carbohydrate structure database 3: principles and realization. *J. Chem. Inf. Model.*, **51**, 159–170.

UniProtConsortium. (2012) Reorganizing the protein space at the Universal Protein Resource (UniProt). *Nucleic Acids Research*, **40**, D71–D75.

von der Lieth,C.W. *et al.* (2011) EUROCarbDB: an open-access platform for glycoinformatics. *Glycobiology*, **21**, 493–502.

Wheeler,D.L. *et al.* (2000) Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.*, **28**, 10–14.