Molecular Systems Biology 6; Article number 403; doi:10.1038/msb.2010.45
Citation: ***Molecular Systems Biology*** **6:403**

www.molecularsystemsbiology.com

molecular systems biology

PERSPECTIVE

Structured digital tables on the Semantic Web: toward a structured digital literature

Kei-Hoi Cheung[1,2,3,4,*], Matthias Samwald[5,6], Raymond K Auerbach[1] and Mark B Gerstein[1,4,7,*]

[1] Program in Computational Biology and Bioinformatics, Yale University, New Haven, CT, USA,
[2] Center for Medical Informatics, School of Medicine, Yale University, New Haven, CT, USA,
[3] Department of Genetics, School of Medicine, Yale University, New Haven, CT, USA,
[4] Department of Computer Science, Yale University, New Haven, CT, USA,
[5] Digital Enterprise Research Institute, National University of Ireland Galway, IDA Business Park, Lower Dangan, Galway, Ireland,
[6] Konrad Lorenz Institute for Evolution and Cognition Research, Altenberg, Austria and
[7] Department of Molecular Biophysics and Biochemistry, Yale University, New Haven, CT, USA
* Corresponding authors. K-H Cheung, Yale Center for Medical Informatics, Yale University, 300 George Street, Suite 501, New Haven, CT 06520-8009, USA. Tel.: +1 203 737 5783; Fax: +1 203 737 5708; E-mail: kei.cheung@yale.edu or MB Gerstein, Department of Molecular Biophysics and Biochemistry, Yale University, New Haven, CT 06520, USA. Tel.: +1 203 432 6105; Fax: +1 203 432 6949; E-mail: mark.gerstein@yale.edu

Received 20.7.09; accepted 15.3.10

In parallel to the growth in bioscience databases, biomedical publications have increased exponentially in the past decade. However, the extraction of high-quality information from the corpus of scientific literature has been hampered by the lack of machine-interpretable content, despite text-mining advances. To address this, we propose creating a structured digital table as part of an overall effort in developing machine-readable, structured digital literature. In particular, we envision transforming publication tables into standardized triples using Semantic Web approaches. We identify three canonical types of tables (conveying information about properties, networks, and concept hierarchies) and show how more complex tables can be built from these basic types. We envision that authors would create tables initially using the structured triples for canonical types and then have them visually rendered for publication, and we present examples for converting representative tables into triples. Finally, we discuss how 'stub' versions of structured digital tables could be a useful bridge for connecting together the literature with databases, allowing the former to more precisely document the later.

Molecular Systems Biology **6**: 403 published online 24 August 2010; doi:10.1038/msb.2010.45
Subject Categories: bioinformatics; computational methods
Keywords: bioinformatics; data integration; semantic publishing; Semantic Web; triplification

Introduction

With the advent of high-throughput experimentation, a deluge of biomedical data of diverse types (e.g. sequences, genes, proteins, and pathways) has been generated in different formats (e.g. eXtensible Markup Language (XML), tabular, and proprietary formats) and stored in numerous public and private databases. In addition, there is a large bulk of unstructured data that is deeply buried in the literature (e.g. journal papers) and as a result, is not readily accessible for inclusion in additional analyses. Given the tremendous growth of data in the form of databases and literature, data and text mining presents promise as well as challenges in bioinformatics and computational biology. For example, as described in Jensen *et al* (2006), there is great potential for enabling biological discoveries by integrating literature and databases. Smith *et al* (2007a) demonstrate how to leverage the Semantic Web graph structure (linking protein identifiers between different papers) to improve information retrieval in the proteomic domain. A small number of known gold-standard examples were used to improve supervised exploration of unclassified documents. Although database mining has been enhanced by machine-friendly standards such as minimum information about a microarray experiment (MIAME) (Brazma *et al*, 2001) and MIAPE (Taylor *et al*, 2007), literature mining has been hampered by the inability to make use of these standards.

To help reconcile the dichotomy between database mining and literature mining, the structured digital abstract (SDA) (Gerstein *et al*, 2007) has been proposed as a first step toward using a database-like format to ease literature mining. This proposal has stirred up debate and several discussions in the scientific community (Editorial, 2007; Leitner and Valencia, 2008). It suggests that a machine-readable structure is provided for summarizing the paper. The structured abstract consists of three main elements:

1. The first element is a translation table or 'cast of characters,' which lists all named biological entities like genes, proteins, metabolites, and other objects in the article, and relates their human-readable names to precise database identifiers.
2. The second element is a list of the main results described in simple ontologies using a controlled vocabulary—for example, interactions ('protein A binds to protein B'), phenotypes ('mutation C suppresses deletion D'), and

protein modifications ('protein E is phosphorylated at residue F by protein kinase G').

3. The third element is standard evidence codes for the main techniques underlying the results—for example, 'affinity purification' or 'mass spectrometry.' Variants of experimental techniques are numerous and the design of such an ontology is beyond the scope of this proposal, but rather propose the use of general, high-level terms such as 'affinity purification' or 'mass spectrometry' to describe data acquisition.

FEBS Letters (http://www.elsevier.com/wps/find/journaldescription.cws_home/506085/description#description), one of the leading journals for short reports in molecular biosciences, has recently launched an experiment to add the SDA as an enhancement to its articles (http://www.eurekalert.org/pub_releases/2008-04/e-fls040208.php).

Although the SDA is being adopted on a pilot basis, this paper provides a perspective as to how to incrementally expand the SDA concept to capture data and results presented in the tabular format. This expansion helps pave the way for a structured digital literature project where abstracts and tables are the first components of the paper to be digitized. A prototype version of a structured digital paper including citations has been described and demonstrated (Shotton *et al*, 2009). Tables (including those that are exposed to the Web in HTML format) are very common presentation schemes for researchers to describe biological entities and their relationships, organize data, summarize experimental results, etc. They have been a target of text mining (Tengli *et al*, 2004; Gatterbauer and Bohunsky, 2006). The accuracy of table extraction, however, has been limited by the heterogeneity of table formats. To this end, we will describe how to use the Semantic Web (Berners-Lee *et al*, 2001) as a standard way to express table contents as machine-readable triples.

The Semantic Web (Berners-Lee *et al*, 2001; Feigenbaum *et al*, 2007) transforms the Web into a global database or knowledge base by providing: (1) globally unique names through the uniform resource identifiers (URIs) (http://www.w3.org/Addressing/), (2) standard languages including the resource description framework (RDF) (http://www.w3.org/TR/REC-rdf-syntax/), RDF schema (RDFS) (http://www.w3.org/TR/rdf-schema/), and the Web Ontology Language (OWL) (http://www.w3.org/TR/owl-features/) for modeling data and creating ontologies, and (3) a standard query language—SPARQL (http://www.w3.org/TR/rdf-sparql-query/). There is a strong interest in exploring the intersection between Semantic Web and life science data integration (Sagotsky *et al*, 2008).

The RDF provides a triple format for representing a statement that consists of *subject*, *property*, and *object*. Each component of the triple is identified by a URI. For example, the following triple (statement) asserts that the 'Wnt signaling pathway' (subject) 'has a participant' (property) 'frizzled' (object).

Subject: http://en.wikipedia.org/wiki/Wnt_signaling_pathway

Property: http://www.obofoundry.org/ro/#OBO_REL:has_participant

Object: http://en.wikipedia.org/wiki/Frizzled

Notice in the above example that URLs (subtype of URIs) are used to locate the subject and object (this is only for demonstration purposes). The relationship (identified by a URI) between the subject and the object is explicitly specified as 'has participant.' As indicated by its URI, this relationship is defined in the Relation Ontology developed by the Open Biomedical Ontology Foundry based on community standards and best practices. There are other triple-based data models. Among them, the entity-attribute-value (EAV) model has been used in building biomedical and neuroscience databases (Nadkarni *et al*, 1998; Marenco *et al*, 2003) in circumstances where the number of attributes (properties or parameters) that can be used to describe an entity (or object) is potentially very large, but the number that will actually apply to a given entity is relatively modest (i.e. the relationship matrix is sparse). One of the common examples of EAV modeling is seen with the clinical findings (e.g. past history, present complaints, physical examination, laboratory tests, special investigations, and diagnoses) that can apply to a patient. Across all specialties of medicine, these can range in the hundreds of thousands (with new tests still being developed). The majority of patients who visit a doctor, however, have relatively few findings.

A collection of RDF triples/statements forms a 'directed acyclic graph.' Such a graph can be identified by a URI. Named graphs (Carroll *et al*, 2005) allow provenance (metadata) to be associated with an entire RDF graph structure by treating the graph as a subject that can have one or more properties. For example, the RDF statement above can be treated as an RDF graph identified by a URI (g_1). Then we can define a new RDF statement as follows:

Subject: g_1

Property: http://purl.org/dc/terms/source

Object: http://www.genome.jp/kegg/pathway/hsa/hsa04310.html

The above statement asserts that the *source* (which is a standard term of the Dublin Core Metadata Initiative (http://dublincore.org/)) of the statement: '*Wnt_signaling_pathway, has-participant, frizzled*' comes from KEGG (Kanehisa and Goto, 2000) that provides information about the Wnt signaling pathway (http://www.genome.jp/kegg/pathway/hsa/hsa04310.html).

Several named graph syntaxes have been proposed, which include TriX (http://www.w3.org/2004/03/trix/), RDF/XML (http://www.w3.org/TR/rdf-syntax-grammar/), and TriG (http://www4.wiwiss.fu-berlin.de/bizer/TriG/) that is an extension of Turtle (http://www.w3.org/TeamSubmission/turtle/).

To capture richer data semantics to support inferencing and reasoning, RDFS and the OWL have been used to encode expressive ontologies in the life sciences (Golbreich *et al*, 2006; Lam *et al*, 2006; Smith *et al*, 2007b; Bug *et al*, 2008). RDFS provides constructs to define classes and their subclass relationships. A resource may be defined as a class (e.g. protein) that can include other resources (e.g. 'frizzled' and 'fibronectin') as its members. As an example of class hierarchy, *Glycoprotein* is a subclass of *Protein*. Most of the RDFS components are included in the more expressive language OWL. OWL has the 'same as' property that allows a synonymous relationship to be defined between resources (e.g. 'Fz-3,' 'hFz3,' and 'frizzled 3'). Also, cardinality constraints can be applied to properties (e.g. the *has_participant* property can have a minimum cardinality of one and a maximum cardinality of some positive integer). Although

OWL is semantically richer than RDF and RDFS, it can be expressed using the RDF syntax. In other words, an OWL ontology can be written in the form of RDF triples.

In a scientific paper, tables are often used to present results including summarized data and observations obtained from a study. A table typically consists of the following elements: (1) a title that provides a brief description of the table, (2) column headers and values, and (3) an optional caption or legend giving a more detailed explanation (including annotation) of the table. We generally classify tables into the following canonical types, as shown in Figure 1A and B.

1. *Properties table*. This type of table is commonly found in the biomedical literature. It is used to describe the properties (or attributes) of a particular type of biological entity (e.g. gene or protein). For example, a gene table may consist of properties including symbol, name, chromosomal location, biological function, and so on. In this case, each row corresponds to a gene, whereas each column corresponds to a property.
2. *Network table*. This represents interactions between biological entities, which form a network. Biological networks constitute an important type of data in the field of systems biology. A network may involve a single type of entity (e.g. proteins) or two different types of entities (e.g. protein kinases and substrates). One common representation of a network is achieved in the form of a matrix table where the first row and column of the matrix are used to indicate the entities involved, whereas other matrix elements are used to indicate the degrees of interaction. An alternative table

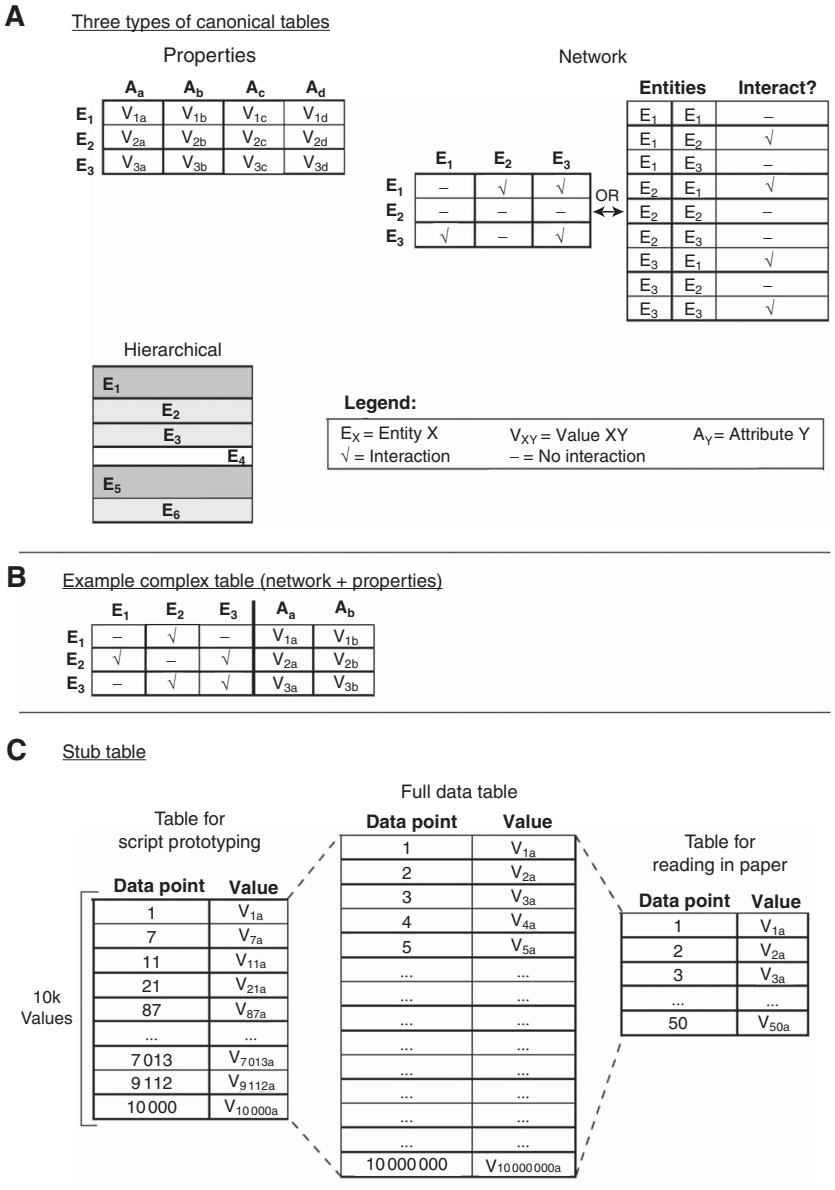


Figure 1 (**A**) Three types of canonical tables. (**B**) A complex table consisting of network and properties tables. (**C**) Stub tables: a small one (listing the top 50 rows of the full table) for reading in paper and a medium one (consisting of 10 000 randomly selected rows from the full table) for scripting purposes.

format contains columns like *entity1*, *entity2*, and *interaction*, where the third column indicates the degree of interaction.

3. *Hierarchical table*. This type of table is an outgrowth of outline structure. For example, a table may be used to outline families and subfamilies of genes (e.g. cytochrome P450 families and its subfamilies).
4. *Complex table*. This represents an ontological structure involving complex relationships among different types of biological entities. In some cases, a complex table represents a combination of properties, network, and hierarchical tables.

In addition, an important aspect of structured digital tables is the concept of a 'stub' table, as shown in Figure 1C. Increasingly, people want to publish extremely large data sets comprising millions of entries and potentially billions and even trillions of bytes. These huge tables are often kept in databases and dealt with separately from publications. This often causes the precise conclusions and specific numbers in publications to become out of sync with what exactly is in the big data table. Structured digital tables can help bridge this divide. One can, in a sense, put all of the information in a huge table into a structured format, very similar to that in the smaller structured digital tables discussed above with, for instance, an overall canonical table layout and standardized columns. Then for the actual publication of a paper one can have a particular subsampling of this huge table displayed, for instance the first 50 rows or the top 50 entries, in a visually appealing way. This would allow the publication to serve as a precise documentation of the actual database data. A second subsampling of the table is often useful for developing computer scripts. One can, for instance, have a subsampling of a randomly chosen 1000 or 10 000 rows, which are good for quickly prototyping scripts that would then run in a long period of time over a whole table with millions of rows.

A multiplicity of formats has been used to publish tables of different types. Although these table formats are readable by humans, they are not readily parseable by machines. To address this problem, we propose a common machine-readable table format for each of the table types shown in Figure 1. Such a common table format is readable by both humans and computer. A similar approach called 'MAGE-TAB' (Rayner *et al*, 2006) has been applied as a canonical table format for representing and sharing DNA microarray data. To further enhance machine readability, we convert the common table format into the triple format for supporting data mining on the Semantic Web.

Table triplification

In this section, we discuss how to transform published tables into a standard digital format using the Semantic Web. We call this process 'table triplification.' To this end, we leverage the work that has been done in the context of translating relational tables into the Semantic Web format. Various tools such as triplify (http://triplify.org), D2R Map (http://www4.wiwiss.fu-berlin.de/bizer/d2rmap/D2Rmap.htm), and FeDeRate (http://www.w3.org/2007/05/SPARQLfed) have been developed to convert relational databases into the RDF structure. At the community level, the RDB2RDF (conversion from relational database to RDF) (http://www.w3.org/2005/Incubator/rdb2rdf) is an incubator activity that has been carried out as part of the World Wide Web Consortium (W3C) activities. Our approach introduces machine-readable table formats to which published tables are mapped. These common table formats can be implemented using the relational database technology as well as other technologies including the XML. Below we describe the steps to convert tables into Semantic Web triples.

1. Map the published table to the corresponding canonical table format. There are two scenarios for mapping a published table to its canonical table counterpart.
   a. The mapping is simple in the sense that the published table and the canonical table have the same structure (the column names may differ).
   b. The mapping is more complicated in the sense that some restructuring is reflected in the machine-readable table (e.g. additional columns may be introduced to capture more semantics needed for triplification).
2. *Translate the* canonical *table into triples*. How to convert from the canonical table to triples depends on the types of the common tables.
   a. *Properties table*. Work has been done to convert a properties table in the EAV model (Marenco *et al*, 2003). In this case, each row is mapped to an entity, each column header is mapped to an attribute, and each column value is mapped to an attribute value. The EAV model corresponds naturally to the *subject–property–object* triples of RDF.
   b. *Network table*. For converting this type of table into triples, we create an 'Interaction' class with the following properties: *entity1* (holding a value from the first column), *entity2* (holding a value from the first row), and *interaction value* (holding the value of the cell that *entity1* and *entity2* intersect).
   c. *Hierarchical table*. The common format of a hierarchical table consists of columns corresponding to the super types and subtypes of entities. For example, the 'families' and 'subfamilies' columns may be used to indicate the superfamilies and subfamilies of genes in a gene table. These additional columns are used to automatically construct the hierarchy of classes to which the resulting triples belong.
   d. *Complex table*. A complex table is formed based on some combination of the properties, network, and hierarchical tables. Although properties, network, and hierarchical tables can be syntactically converted into triples, the conversion of complex tables into ontologies needs to be done semantically. This often involves a manual process. In other words, the conversion code is table specific.
3. *Use named graph to store provenance and metadata*. This is the last step of table triplification in which provenance and metadata associated with the tables are stored using named graphs. Some representative types of provenance and metadata include the following:
   a. Creator (who created the triples).
   b. Creation date (when the triples were obtained).
   c. Source (e.g. the source publication containing the table).

d. Title (a short description of the table).
e. Table captions or legends (they serve as a detailed description and annotation of the table).
f. Summary information (size of the table including number of rows and number of columns).
g. Table type (e.g. properties table, network hierarchical table, or complex table).
h. Types of entities represented by the table.
i. Interpretation of nulls—what do missing values really mean? For example, they may refer unknown or uncertain values. Their meaning may be specific to individual columns.
j. Column-specific metadata (footnotes):
   - Precision—mathematically, it refers to the number of digits to which a column value may be measured reliably. It also reflects the ability of a measurement to be reproduced consistently.
   - Units of measurement (e.g. μg and mg are units of mass measurement).
   - Footnotes may sometimes be applied to an individual column value instead of a whole column. This is particularly true if these individual values represent outliners or exceptions.

Below we provide a number of examples to illustrate how different types of tables are triplified.

Table I Yeast strains used in the study by Hermann *et al* (1997)

Name	Genotype*	Source
FY10	*MATα leu2Δ1 ura3-52*	F Winston
FY22	*MATα his3Δ200 ura3-52*	F Winston
GHY1	*MATα leu2Δ1 his3Δ200 ura3-52 mdm20-1*	This study
JSY707	*MATα his3Δ200 ura3-52 tpm1D::HIS3*	This study
JSY948	*MATα leu2Δ1/leu2Δ1 ura3-52/ura3-52*	This study
JSY999	*MATα leu2Δ1 his3Δ200 ura3-52*	This study
JSY1065	*MATα leu2Δ1 his3Δ200 ura3-52 mdm20D::LEU2*	This study
JSY1084	*MATα leu2Δ1 his3Δ200 ura3-52 tpm1D::HIS3*	This study
JSY1138	*MATα leu2Δ1/leu2Δ1 his3Δ200/his3Δ 200 ura3-52/ura3-52 tpm1D::HIS3/+ mdm20D::LEU2/+*	This study
JSY1285	*MATα leu2Δ1 his3Δ200 ura3-52 tpm2D::HIS3*	This study
JSY1340	*MATα leu2Δ1 his3Δ200 ura3-52 mdm20D::LEU2*	This study
JSY1374	*MATα leu2Δ1/leu2Δ1 his3Δ200/his3Δ200 ura3-52/ura3-52 tpm2D::HIS3/+ mdm20D::LEU2/+*	This study
ABY1249	*MATα leu2-3,112 ura3-52 lys2-801 ade2-101 ade3 bem2-10*	A Bretscher
IGY4	*MATα leu2-3,112 his3Δ200 ura3-52 lys2-801 ade2 sac6D::LEU2*	A Adams
SLY63	*MATα leu2-3,112 ura3-52 trp1-1 his6 myo2-66*	S Brown

*All of the GHY and JSY strains used in this study are isogenic to FY10 (Winston *et al*, 1995).

Properties table

Table I is an example of a properties table (its canonical table counterpart has the same structure). This table was obtained from a study to test whether the yeast gene, MDM20, is necessary for mitochondrial inheritance and organization of the actin cytoskeleton (Hermann *et al*, 1997). It lists the different yeast strains that were used in the study. The table has three columns (name, genotype, and source). Each table row corresponds to a specific yeast strain. We can apply the following rules to convert this table into RDF triples:

1. Each row is mapped to a subject
2. Each column header is mapped to a property
3. Each column value (cell) is mapped to a property value

Figure 2 depicts the mapping process and some of the mapping results. For the subject of each triple, we may check to see if it is an instance of an existing ontology class (represented using OWL or RDFS). For example, each subject (e.g.'FY10') derived from Table I is an instance of (represented by a dotted line) the class 'yeast strain' in some organism ontology. Although the column name can be used to name the property, we may want to map it to some standard property name, if available. The generated triples represent a RDF graph. To this end, we use the named graph technique to identify the RDF graph generated from the table and to store the provenance information including the title, description (e.g. the table caption), creator, source (e.g. the paper), and so on. The properties (e.g. title, description, creator and source) are derived from the Dublin Core metadata standard (http://dublincore.org/).

Although Table I can be converted into the canonical form without restructuring, there are canonical tables that are created by restructuring the published tables. Figure 3 depicts how such a table restructuring takes place. The published table shown in Figure 3A lists several of the palmitoylated proteins identified by two proteomic experiments using the multi-dimensional protein identification technology (MudPIT) (Martin and Cravatt, 2009). These experiments involve different numbers of fractionations. In the first experiment (number of fractions=6), cells treated with the reagent 17-ODYA (labeled cells) were compared with those treated with the reagent palmitate (controlled cells). In the second experiment (number of fractions=5), the cells labeled with 17-ODYA were compared with those treated with hydroxylamine. The amounts of protein (protein abundance) were estimated by spectral counting. These spectral counts are listed under the columns: '17-ODYA' and 'palmitate' for experiment 1; '17-ODYA' and 'hydroxylamine' for experiment 2. Although the grouping of columns is shown visually, it needs to be represented in the canonical table in a machine-readable way. To this end, additional columns are created in the canonical table, as shown in Figure 3B, for explicitly describing each experiment (e.g. experiment ID and number of fractions) and the experimental results (e.g. label/control reagents used in each experiment and the spectral counts corresponding to each protein). In addition, a new column named 'N-Met-Gly' is created to identify proteins that have a consensus myristoylation site (such proteins are annotated in the published table—the annotation is described in the table caption). The mappings between the published table and canonical table are represented by the dotted lines. For converting into triples, the same rules that were applied in the previous example can then be applied to the canonical table here. The difference is that two named graphs (one for each experiment) are generated in this case. Figure 3C shows one of the named graphs.

Network table

Figure 4 gives an example illustrating how a network table is converted into triples. Figure 4A depicts a table representing the interactions between two types of entities: proteins and neuroreceptor subunits. The first column lists the PDZ-containing proteins, whereas the column headers (columns two to five) represent different NMDA receptor subunits. These are part of the results generated by a study that investigates the molecular interactions between PDZ domains of PSD-95 proteins and C termini of NMDA receptor subunits in neuronal synapses (Cui *et al*, 2007). Such interactions may have a function in downstream neurotoxic signaling molecules. Figure 4B depicts a set of triples representing a particular interaction (between TIP1 and NR2B). An *interaction* subject has the following properties: *entity1*, *entity2*, and *value*. In this case, *entity1* and *entity2* belong to two different entity types. The dotted lines show how some of the triples correspond to the table elements.

Hierarchical table

Figure 5 depicts how a hierarchical table is triplified into the corresponding hierarchy. The table lists the caterpiller subfamilies of proteins that have been identified to be involved in inflammation (Tschopp *et al*, 2003). As shown in Figure 5A, the *synonyms* column contains multiple types of information: subfamily names, common protein names, and synonyms. Such composite information is decomposed into multiple columns (*caterpiller subfamilies*, *protein*, and *synonyms*) in the canonical table, as shown by the dotted lines in Figure 5B. On the basis of the hierarchically related columns *caterpiller subfamilies* and *protein* (enclosed by a dotted rectangle), the protein family hierarchy is automatically constructed, as shown in Figure 5C.

Complex table

Figure 6 depicts how an ontology (which can be viewed as a set of interrelated triples) is created based on the semantic relationships between different entities represented in a complex table. The table shown in Figure 6A captures information about different types of receptors (in the *Type* column), drugs, and drug actions on the receptors. This complex table is composed of a hierarchical table (receptor hierarchy), network table (interactions between drugs and receptors), and properties table (properties related to the interactions). This table is part of an article that gives a comprehensive review of different drugs and their different types of targets including receptors (Imming *et al*, 2006).

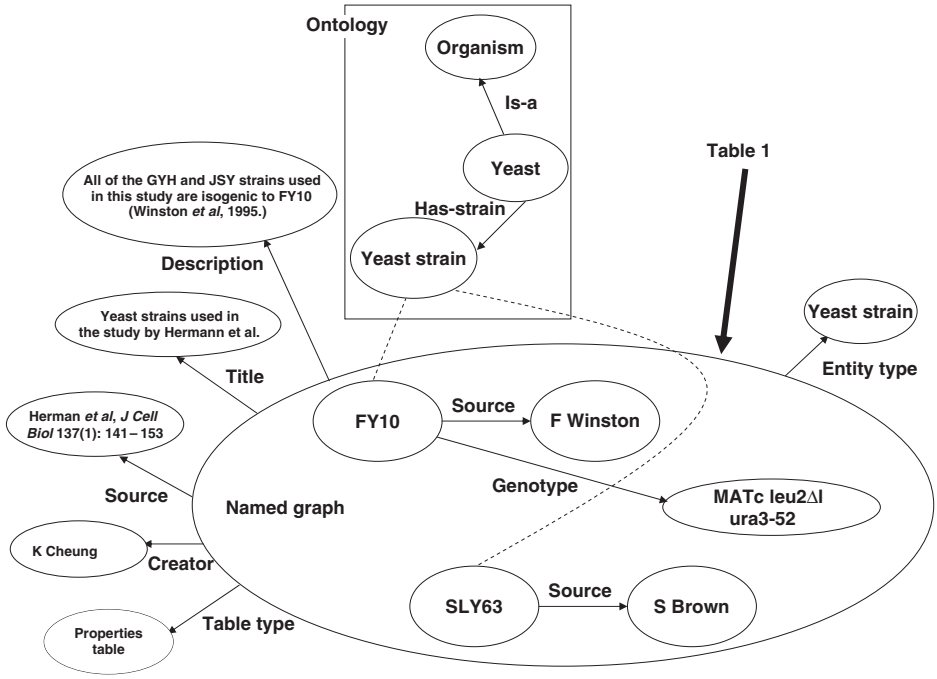


Figure 2 Conversion of Table I into triples contained in a named graph. Source data is available for this figure at www.nature.com/msb.

Figure 3 (**A**) A published table featuring a repeated group of columns (experiment 1 and experiment 2; Martin and Cravatt, 2009, reproduced with permission, *Nature Methods* © 2009). (**B**) The corresponding canonical tables featuring some restructure of the published table. For example, two canonical tables are derived from the single-published table according to the two experiments. As described in the paper, *17-ODYA* is the name of the reagent that was used for labeling the sample, whereas *palmitate* and *hydroxylamine* are the names of the reagents used for treating the controls, additional columns are created in the canonical table for storing these reagent names. (**C**) Triple graph representation of the common table. Notice that separate named graphs are defined for the control and label groups (of identified proteins), with the reagent property associating with each named graph. This approach helps reduce the number of triples (the reagent property does not need to be defined for each identified protein). Source data is available for this figure at www.nature.com/msb.

Protein name	Protein description	Experiment 1 (*n* = 6) 17-ODYA	Experiment 1 (*n* = 6) Palmitate	Experiment 2 (*n* = 5) 17-ODYA	Experiment 2 (*n* = 5) Hydroxylamine
CANX	Calnexin	95±11	11 ± 2	122 ± 30	2 ± 1
HLA-A, HLA-B, HLA-C	HLA class I histocompatibility antigen	68±10	13 ± 6	68 ± 12	0 ± 0
LCK	Proto-oncogene tyrosine-protein kinase LCK	75±13	1 ± 1	25 ± 4	4 ± 1

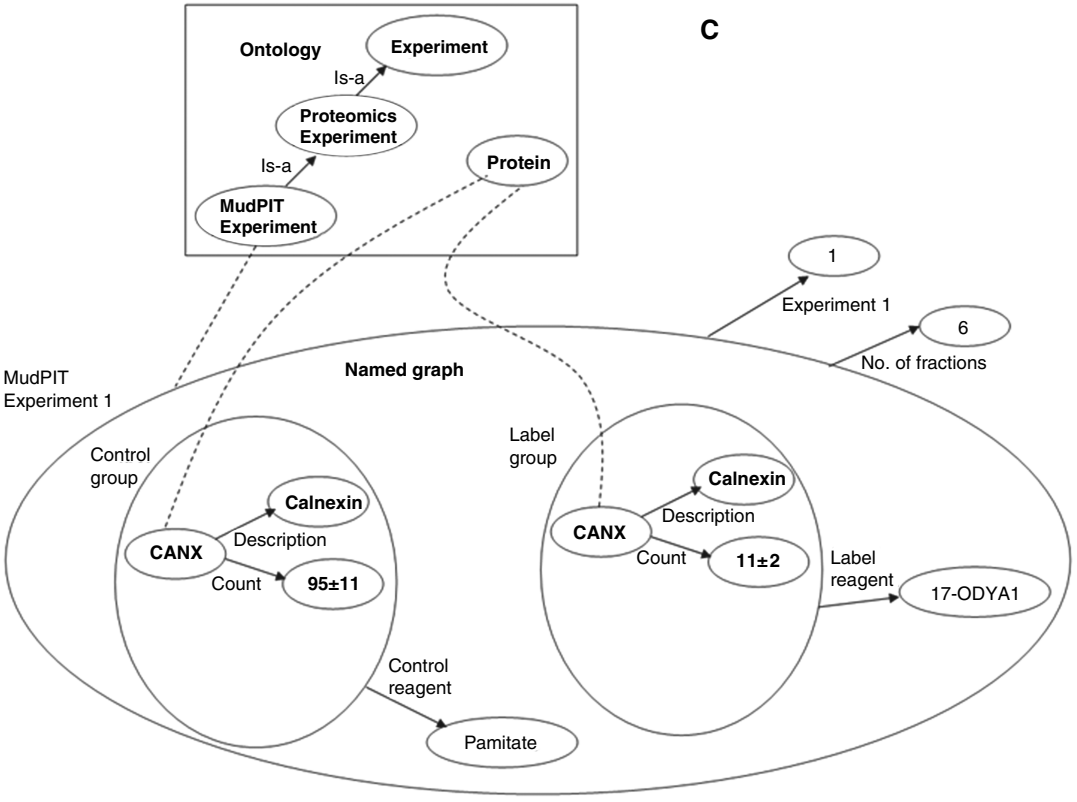
Proteins are listed according to highest total spectral counts identified in both experiment 1 and experiment 2, which compare 17-ODYA-labeled membrane particulate proteomes to palmitic acid-treated and hydroxylamine-treated controls, respectively. Spectral count data represent average values ± s.e.m.

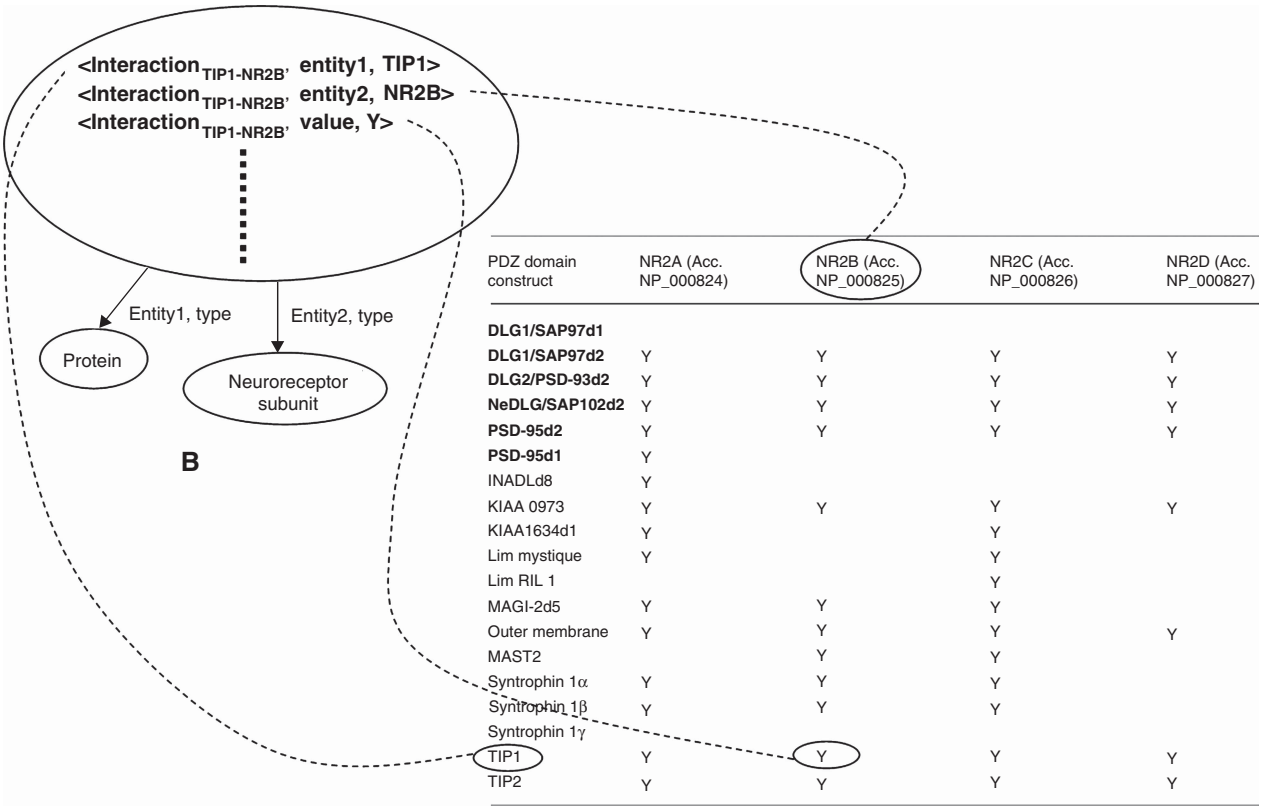
*Proteins with a consensus myristoylation site (N-Met-Gly)

A

B

	Protein name	N-Met-Gly	Protein description	Label count	Label reagent	Control count	Control reagent	Experiment	Number of fractions
Experiment 1	CANX		Calnexin	95±11	17-ODYA	11 ± 2	Palmitate	1	6
	HLA-A, HLA-B, HLA-C		HLA class I histocompatibility antigen	68±10	17-ODYA	13 ± 6	Palmitate	1	6
	LCK	Y	Proto-oncogene tyrosine-protein kinase LCK	73±13	17-ODYA	1 ± 1	Palmitate	1	6
Experiment 2	CANX		Calnexin	122 ±30	17-ODYA	2 ± 1	Hydroxylamine	2	5
	HLA-A, HLA-B, HLA-C		HLA class I histocompatibility antigen	68±12	17-ODYA	0 ± 0	Hydroxylamine	2	5
	LCK	Y	Proto-oncogene tyrosine-protein kinase LCK	25±4	17-ODYA	4 ± 1	Hydroxylamine	2	5





PDZ domain construct	NR2A (Acc. NP_000824)	NR2B (Acc. NP_000825)	NR2C (Acc. NP_000826)	NR2D (Acc. NP_000827)
DLG1/SAP97d1				
DLG1/SAP97d2	Y	Y	Y	Y
DLG2/PSD-93d2	Y	Y	Y	Y
NeDLG/SAP102d2	Y	Y	Y	Y
PSD-95d2	Y	Y	Y	Y
PSD-95d1	Y			
INADLd8	Y			
KIAA 0973	Y	Y	Y	Y
KIAA1634d1	Y		Y	
Lim mystique	Y		Y	
Lim RIL 1			Y	
MAGI-2d5	Y	Y	Y	
Outer membrane	Y	Y	Y	Y
MAST2		Y	Y	
Syntrophin 1α	Y	Y	Y	
Syntrophin 1β	Y	Y	Y	
Syntrophin 1γ				
TIP1	Y	Y	Y	Y
TIP2	Y	Y	Y	Y

Acc., GenBank accession number, Y, yes (there is an interaction). PSD-95 family members are in bold lettering.

A

Figure 4 (**A**) PDZ domain-containing protein and NMDA receptor subunit interactions (Cui *et al*, 2007, reproduced with permission *Journal of Neuroscience* © 2007). (**B**) A set of triples (triple graph) corresponding to the interaction between TIP1 and NR2B. Source data is available for this figure at www.nature.com/msb.

Figure 6B shows the corresponding canonical table that includes two hierarchically related columns: *receptor type* and *receptor* (their mappings to the published table are shown by the dotted lines). Figure 6C depicts the resulting ontology that includes the receptor hierarchy as well as the relationships between the receptors and drugs (through the *drug–receptor interaction* class). As shown in Figure 6C, each oval represents a class and each arrow represents a relationship. There are two broad types of relationships: *is-a* relationship and user-defined relationships. For example, 'GABAA receptors' *is-a* 'G-protein-coupled-receptors.' In addition, the class 'drug–receptor interaction' has the following properties: *receptor*, *drug*, *drug activity*, and *source*. Unlike the previous table-to-triple conversions, the ontological conversion here is handcrafted.

Discussion

Although we have given some representative examples illustrating how to convert different types of tables into triples, there are other types of tables that are not included in this paper because of space limitation. For example, Supplementary Tables are sometimes accompanied with a paper. Unlike published tables, Supplementary Tables are usually quite large in size. Converting such large tables into triples requires more storage and efficient query support.

We have identified two general approaches to table triplification, namely, syntactic triplification and semantic triplification. The former represents a bottom–up approach with minimal or no input from existing ontologies, whereas the latter represents a top–down approach featuring use of available standard ontologies in directing the triplification process. These approaches are not mutually exclusive to each other. Instead, they are complementary in the sense that the bottom–up approach can serve as a starting point for achieving unification at the syntactic level, whereas the top–down approach can be used to implement broad semantic interoperability in an incremental manner. The table shown in Figure 3 can potentially be converted into an experiment-centric ontology based on existing ontologies such as Ontology for Biomedical Investigations (OBI) (http://obi-ontology.org/page/Main_Page) and EXPO (Soldatova and King, 2006). In addition, there is a growing collection of biomedical ontologies that are developed by different groups based upon standard guidelines and best practices that are set forth by the ontology community (e.g. OBO Foundry (Smith *et al*, 2007b)). Such ontologies have an important function in the semantic triplification of tables.

To reduce the effort of semantic table triplication, text-mining techniques may be used to help identify mappings between table elements and elements of existing ontologies. For example, there are tools/services available for automati-

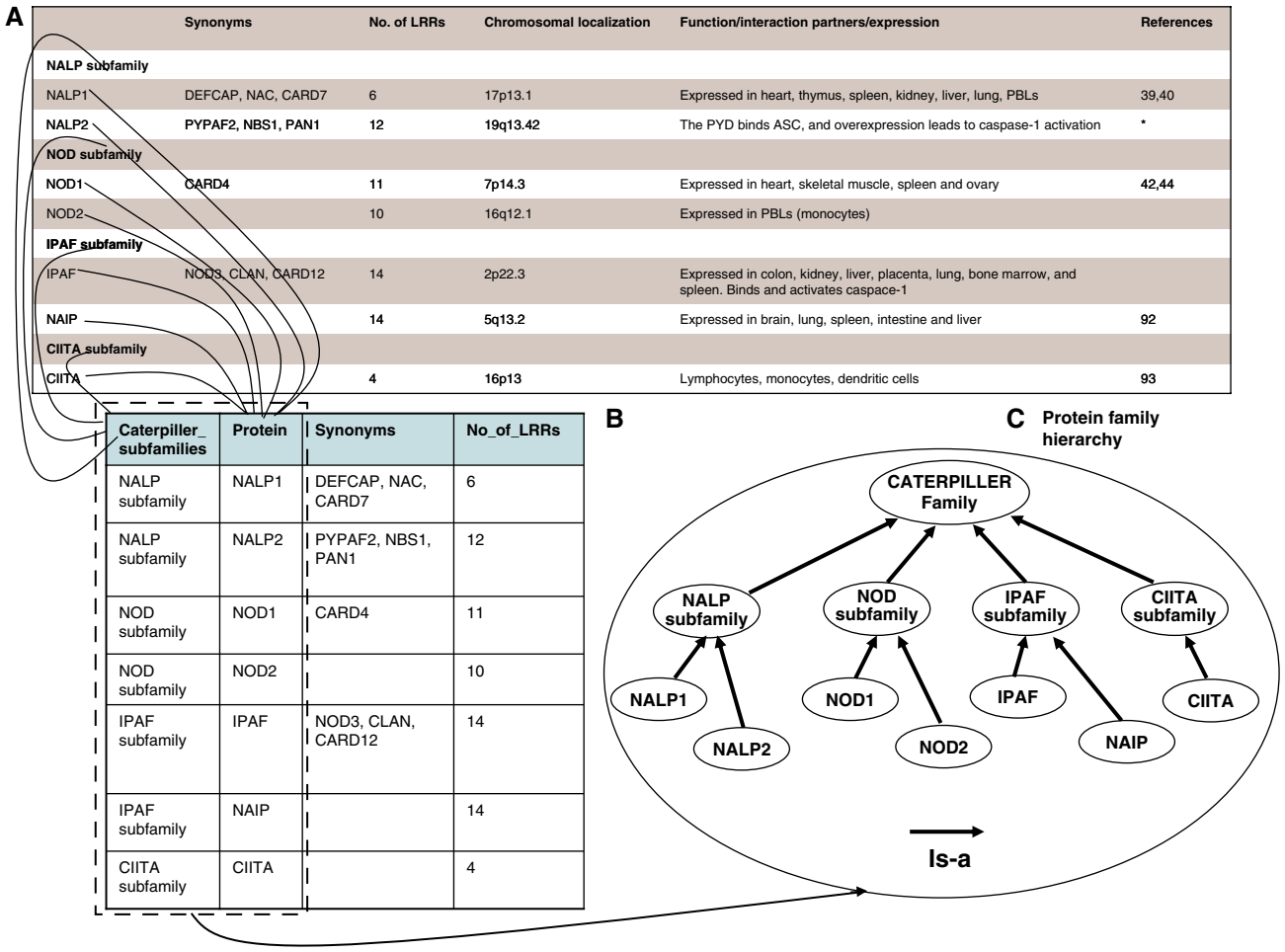


Figure 5 (**A**) A published table listing the caterpiller subfamilies of proteins involved in inflammation. (**B**) A portion of the corresponding canonical tables showing the two columns (caterpiller subfamilies and protein) extracted from the synonyms column of the published table (Tschopp *et al*, 2003, reproduced with permission *Nat Rev Mol Cell Biol* © 2003). (**C**) The hierarchical graph structure of the caterpiller protein families/subfamilies. Source data is available for this figure at www.nature.com/msb.

cally extracting concepts/relationships from text including tables. Some of these tools (e.g. Open Biomedical Annotator (Shah *et al*, 2009) developed by the National Center for Biomedical Ontology and WikiProtein's Knowlets (Mons *et al*, 2008)) can be driven by existing ontologies.

One problem that might be encountered during the extraction of data from tables is that, for space reasons, authors often use acronyms to refer to entities (e.g. genes and proteins) within the tables. In many cases, these acronyms are defined locally within each publication, which would become unintelligible outside of the context of the original document. This problem can be addressed by using algorithms that expand acronyms that have been defined locally within the document to their spelled-out variants. For example, the Schwartz and Hearst algorithm (http://biotext.berkeley.edu/software.html) could be used for this purpose quite straightforwardly. In addition, there are text-mining tools specifically developed for processing biomedical literature (e.g. Chilibot and iHOP (Chen and Sharp, 2004; Good *et al*, 2006)). These tools are capable of recognizing acronyms such as gene/protein symbols. They may aid in the automatic (semi-automatic) conversion of published tables into canonical tables.

As more tables are available on the Web in HTML format, technologies such as RDF attributes (RDFa) (http://www.w3.org/TR/xhtml-rdfa-primer/) can be used to embed RDF triples into the existing HTML-formatted tables. Such embedded triples are invisible to humans, but they are readable by machines. In addition, tools such as Vispedia (Chan *et al*, 2008) have been developed to automatically identify, query, and integrate table data published in different Wikipedia articles. Such table extraction tools may be extended to access tables published in any Web articles in general. To integrate RDF triples derived from tables published in the same paper or different papers, it is important to make sure that global URIs are used whenever possible. For example, the same protein referenced in different tables should be identified by the same URI. Otherwise, synonymous relationships will need to be established, which may create maintenance and performance problems. The 'cast of characters' (list of canonical names) that is part of the original SDA also partially addresses the acronym problem. The unique URIs can be included among the cast of characters in the SDA. In addition, the named graph URI identifying a table can be associated with a snippet of text describing the table in the paper.

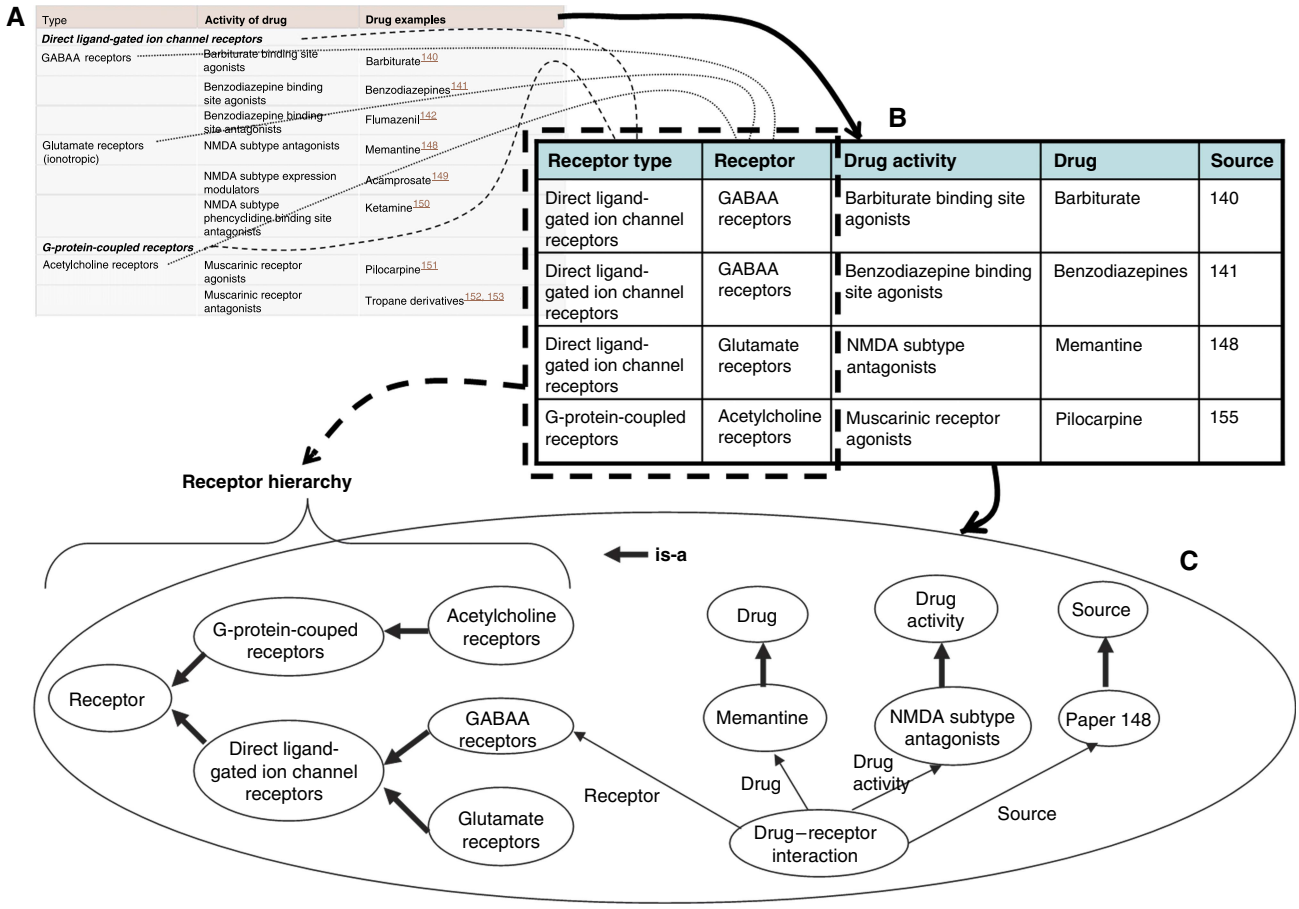


Figure 6 (**A**) A published table listing different drugs and their activities on different categories of receptors (Imming *et al*, 2006, reproduced with permission *Nat Rev Drug Discov* © 2006). (**B**) The corresponding canonical table. (**C**) The ontology graph is created based on the canonical table (the RDF representation of the graph is available in Supplementary information). Source data is available for this figure at www.nature.com/msb.

For publishers to embrace table triplification, tools may be provided for the authors to create common tables. Some word processors (e.g. Word) provide tools for creating tables. It would be beneficial if these table creation tools can be extended to support the triplification task. Recently, there is a Microsoft Word 'add-in' that supports ontology recognition (http://ucsdbiolit.codeplex.com/). It allows text to be automatically tagged with existing ontological terms. These kind of author-oriented tools should also be helpful to achieving table triplification. In addition, different templates may be provided for authors to choose to convert the canonical table into a pretty table format for human readability. One could imagine readers even choosing to render a set of underlying triples in alternative visual formats, depending on the application—in a similar way to how an underlying latex or XML formatted table can be differently presented depending on a style sheet.

We would like to point out that although it will be a while before all tables in papers are laid out in a structured form, even a small number of structured tables in the literature can potentially be quite beneficial. They can function as a gold standard to train and test text-mining engines, which can then be better used to classify the remaining unstructured parts of the literature.

Although tables are a common part of a scientific paper, there are other components that can also potentially be triplified. One such component is the *image*. For example, SLIF (Qian and Murphy, 2008) is a system that identifies panels containing fluorescence microscope images among figures in online journal articles as a prelude to further analysis of the subcellular patterns in such images. It introduces the use of a type of probabilistic graphical model, a factor graph, to represent the structured information about the images in a figure, and permit more robust and accurate inference about their types. Such a factor graph may be represented as an RDF graph.

Another component is the *hypothesis*. Biomedical ontologies such as the SWAN have been developed to represent and relate hypotheses as part of the scientific discourse (Ciccarese *et al*, 2008). The SWAN ontology has been used in the context of neurological diseases such as Alzheimer's disease (AD). In the SWAN model, a hypothesis contains one or more claims. Each claim is supported by evidence (e.g. citations and findings). A hypothesis has a descriptive title (e.g. 'function of intramembraneous a-Beta dimers in AD pathogenesis'). A claim is expressed in the form of a simple statement (e.g. 'a-Beta peptides alter membrane properties'). These claims/statements can potentially be broken into related triples.

In addition to hypotheses, there are a number of data specifications and ontologies that have been developed to describe different types of *experiments*. For example, the MIAME (Brazma *et al*, 2001) is a standard guideline specifically designed for describing microarray (transcriptomics) experiments so that data generated from different microarray experiments can be reproduced, compared, and integrated more easily. The MGED ontology has been created based on the MIAME guideline as an annotation resource for microarray data. Similar data specifications and ontologies have been developed for other biomedical domains including proteomics (e.g. MIAPE (Taylor *et al*, 2007)) and functional genomics (e.g. FuGO (Whetzel *et al*, 2006)). In addition to domain-specific ontologies, domain-independent ontologies have been proposed to describe scientific experiments in general. Examples are EXPO (Soldatova and King, 2006) and the OBI. Distilled from these experiment ontologies, the following concepts/terms can be used to describe experiment-related elements.

- Title of the experiment
- Type of experiment: (e.g. high-throughput sequencing, DNA microarray, proteomics, etc)
- Technology type/platform: (e.g. chip seq, RNA seq, mass spectrometry)
- Instrument (e.g. oligo spotted array, affymetrix gene chip, centrifuge)
- Experiment design (e.g. controls, variables, replication)
- Organism (e.g. NCBI taxonomy)
- Biological materials:

  (a) Biological source (e.g. cell line, tissue)
  (b) Sample extracted from the source. The following sample descriptors may apply: (eukaryote prokaryote) sex, age, developmental stage, organism part (tissue), cell type, animal/plant strain or line, genetic variation (e.g. gene knockout, transgenic variation), individual genetic characteristics (e.g. disease alleles, polymorphisms), disease state or normal
  (c) Sample extraction/separation method (e.g. none, trimming, microdissections, fluorescence-activated cell sorting (FACS))

- Sample treatment:

  (a) Growth conditions
  (b) *In vivo* treatment (e.g. organism or individual treatments)
  (c) *In vitro* treatment (e.g. cell culture conditions)
  (d) Treatment type (e.g. small molecule, heat shock, cold shock, food deprivation)
  (e) Compound

- Experiment protocol for conducting the experiment
- Data protocol:

  (a) Data pre-processing (e.g. normalization)
  (b) Data analysis (e.g. principal components analysis dimensionality reduction, mean calculation)

Recently, there have been several technological developments that can help push the Semantic Web to a new level of data interoperability. Among these developments, linked data (http://linkeddata.org/) is a method of exposing, sharing, and connecting data via dereferenceable HTTP URIs on the Semantic Web. A dereferenceable HTTP URI serves as both an identifier and a locator. The key idea is that useful information should be provided to data consumers when its URI is dereferenced. For example, useful information about a protein (e.g. amino-acid sequence and 3D structure) should be provided by dereferencing the URI of the protein. Using the linked data approach, not only do data providers make their data available in the form of RDF graphs (table triples), but data linkers can also create new RDF graphs that consist of links between independently generated RDF graphs provided by different sources. Examples of linked data (e.g. DBpedia (http://dbpedia.org/)) are listed on linking open data (http://linkeddata.org/).

Last but not least, publishing on the Semantic Web requires a community effort. Efforts such as the Semantic Web for Health Care and Life Sciences Interest Group (http://www.w3.org/2001/sw/hcls/) and the National Center for Biomedical Ontologies (http://bioontology.org/) have important functions in advocating/promoting the use of standard technologies in scientific data sharing. In addition, major data providers (e.g. NCBI and EBI) and publishers (e.g. Nature Publishing Group, Science Magazine, and Elsevier) need to work together to come up with ways to bridge the digital gap between databases and literature.

Conclusion

Our paper envisions the need to create an information infrastructure that allows a gradual shift from the current purely human-oriented reading of content to the one that enables some machine-based interpretation. As part of this transition, we seek ways to digitize/triplify legacy data that have been presented in the form of human-readable tables. Another part of the transition is that triples will be automatically produced by using advanced editing tools as part of the paper-writing process in the future paradigm of semantic publishing.

We have presented approaches to turning human-readable tables published in the literature into structured digital tables published on the Semantic Web (in the form of machine-readable triples). Such machine-readable tables (produced by individual authors/curators/editors) can be automatically/semantically linked to each other and then can be mined by programs developed by other researchers (possibly in some other discipline). Below are several use cases of globally linking structured digital tables on the Semantic Web.

- *Annotation*. There are tables providing curated annotations of biological entities such as genes and proteins. By combining these annotation tables containing the same entities, we can acquire more complete annotation for these entities.
- *Network analysis*. Tables representing biological networks (e.g. gene and protein networks) can be combined to form more comprehensive networks for increasing the power of network analysis.
- *Integrative mining based on combining different types of data*. A large number of data-mining approaches have been developed based on integration of different types of data.

For example, the table (protein/receptor interactions) shown in Figure 4 can potentially be linked to the table (drug–receptor interactions) shown in Figure 6. By linking these two tables obtained from separate studies, we can get more comprehensive information (drug activities and protein interactions) about the NMDA receptor (glutamate receptor). Someone may develop tools to mine such integrated information to discover more drug targets because the proteins interacting with the receptor may also be drug targets.

These use cases demonstrate that the creation of a social networking environment fostering collaborations that would otherwise not have been possible. Table triplification on the Semantic Web makes scientific publishing more intelligent, collaborative, and collective. Although tables are the main discussion in this paper, other components including figures, hypotheses, and experiments can potentially be triplified along the journey toward a structured digital literature. There has been some progress in standardizing digital figures (Ahmed *et al*, 2009; Chen and Murphy, 2009), and we imagine the next components of a paper to be readily standardized would be equations and then the methods.

Supplementary information

Supplementary information is available at the *Molecular Systems Biology* website (http://www.nature.com/msb).

Acknowledgements

We acknowledge support from the NIH and from the AL Williams Professorship funds. KC is supported in part by NIH grant P01 DC04732.

Conflict of interest

The authors declare that they have no conflict of interest.

References

Ahmed A, Xing E, Cohen W, Murphy R (2009) Structured correspondence topic models for mining captioned figures in biological literature. *Proc 15th ACM SIGKDD Int Conf Knowledge Discov Data Mining*, pp 39–47

Berners-Lee T, Hendler J, Lassila O (2001) The Semantic Web. *Sci Am* **284:** 34–43

Brazma A, Hingamp P, Quackenbush J, Sherlock G, Spellman P, Stoeckert C, Aach J, Ansorge W, Ball C, Causton H, Gaasterland T, Glenisson P, Holstege F, Kim I, Markowitz V, Matese J, Parkinson H, Robinson A, Sarkans U, Schulze-Kremer S *et al* (2001) Minimum information about a microarray experiment (MIAME)—toward standards for microarray data. *Nat Genet* **29:** 365–371

Bug WJ, Ascoli GA, Grethe JS, Gupta A, Fennema-Notestine C, Laird AR, Larson SD, Rubin D, Shepherd GM, Turner JA, Martone ME (2008) The NIFSTD and BIRNLex vocabularies: building comprehensive ontologies for neuroscience. *Neuroinformatics* **6:** 175–194

Carroll J, Bizer C, Hayes P, Stickler P (2005) Named graphs. *Web Semant* **3:** 247–267

Chan B, Wu L, Talbot J, Cammarando M, Hanrahan P (2008) Vispedia: interactive visual exploration of Wikipedia data via search-based integration. *IEEE Trans Comput Graph* **14:** 1213–1220

Chen H, Sharp B (2004) Content-rich biological network constructed by mining PubMed abstracts. *BMC Bioinformatics* **5:** 147

Chen S-C, Murphy R (2009) A graphical model approach to automated classification of protein subcellular location patterns in multi-cell images. *BMC Bioinformatics* **7:** 90

Ciccarese P, Wu E, Wong G, Ocana M, Kinoshita J, Ruttenberg A, Clark T (2008) The SWAN biomedical discourse ontology. *J Biomed Inform* **41:** 739–751

Cui H, Hayashi A, Sun H-S, Belmares M, Cobey C, Phan T, Schweizer J, Salter M, Wang Y, Tasker R, Garman D, Rabinowitz J, Lu P, Tymianski M (2007) PDZ protein interactions underlying NMDA receptor-mediated excitotoxicity and neuroprotection by PSD-95 inhibitors. *J Neurosci* **27:** 9901–9915

Editorial (2007) The database revolution. *Nature* **445:** 229–230

Feigenbaum L, Herman I, Hongsermeier T, Neumann E, Stephens S (2007) The Semantic Web in action. *Sci Am* **297:** 64–71

Gatterbauer W, Bohunsky P (2006) *Table Extraction Using Spatial Reasoning on the CSS2 Visual Box Model 21st AAAI Conference on Artificial Intelligence*. Boston, MA: AAAI Press

Gerstein M, Seringhaus M, Fields S (2007) Structured digital abstract makes text mining easy. *Nature* **447:** 142

Golbreich C, Zhang S, Bodenreider O (2006) The foundational model of anatomy in OWL. *J Web Semant* **4:** 181–195

Good B, Kawas E, Kuo B, Wilkinson M (2006) iHOPerator: user-scripting a personalized bioinformatics Web, starting with the iHOP website. *Curr Protoc Bioinformatics* **7:** 534

Hermann G, King E, Shaw J (1997) The yeast gene, MDM20, is necessary for mitochondrial inheritance and organization of the actin cytoskeleton. *J Cell Biol* **137:** 141–153

Imming P, Sinning C, Meyer A (2006) Drugs, their targets and the nature and number of drug targets. *Nat Rev Drug Discov* **5:** 821–834

Jensen L, Saric S, Bork P (2006) Literature mining for the biologist: from information retrieval to biological discovery Nature Reviews. *Genetics* **7:** 119–129

Kanehisa M, Goto S (2000) KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res* **28:** 27–30

Lam HY, Marenco L, Shepherd GM, Miller PL, Cheung KH (2006) Using web ontology language to integrate heterogeneous databases in the neurosciences. *AMIA Annu Symp Proc*, pp 464–468

Leitner F, Valencia A (2008) A text-mining perspective on the requirements for electronically annotated abstracts. *FEBS Lett* **582:** 1178–1181

Marenco L, Tosches N, Crasto C, Shepherd G, Miller PL, Nadkarni PM (2003) Achieving evolvable web-database bioscience applications using the EAV/CR Framework: recent advances. *J Am Med Inform Assoc* **10:** 444–453

Martin B, Cravatt B (2009) Large-scale profiling of protein palmitoylation in mammalian cells. *Nat Methods* **6:** 135–138

Mons B, Ashburner M, Chichester C, Mulligen Ev, Weeber M, Dunnen Jd, Ommen Gv, Musen M, Cockerill M, Hermjakob H, Mons A, Packer A, Pacheco R, Lewis S, Berkeley A, Melton W, Barris N, Wales J, Meijssen G, Moeller E *et al* (2008) Calling on a million minds for community annotation in WikiProteins. *Genome Biol* **9:** R89

Nadkarni P, Brandt C, Frawley S, Sayward F, Einbinder R, Zelterman D, Schacter L, Miller P (1998) Managing attribute—value clinical trials data using the ACT/DB client-server database system. *J Am Med Inform Assoc* **5:** 139–151

Qian Y, Murphy R (2008) Improved recognition of figures containing fluorescence microscope images in online journal articles using graphical models. *Bioinformatics* **24:** 569–576

Rayner T, Rocca-Serra P, Spellman P, Causton H, Farne A, Holloway E, Irizarry R, Liu J, Maier D, Miller M, Petersen K, Quackenbush J, Sherlock G, Stoeckert C, White J, Whetzel P, Wymore F, Parkinson H, Sarkans U, Ball C *et al* (2006) A simple spreadsheet-based, MIAME-supportive format for microarray data: MAGE-TAB. *BMC Bioinformatics* **7:** 486

Sagotsky JA, Zhang L, Wang Z, Martin S, Deisboeck T (2008) Life sciences and the web: a new era for collaboration. *Mol Syst Biol* **4:** 201

Shah N, Bhatia N, Jonquet C, Rubin D, Chiang A, Musen M (2009) Comparison of concept recognizers for building the open biomedical annotator. *BMC Bioinformatics* **10**(Suppl 9)**:** S14

Shotton D, Portwin K, Klyne G, Miles A (2009) Adventures in semantic publishing: exemplar semantic enhancements of a research article. *PLoS Comput Biol* **5:** e1000361

Smith A, Cheung K, Krauthammer M, Schultz M, Gerstein M (2007a) Leveraging the structure of the Semantic Web to enhance information retrieval for proteomics. *Bioinformatics* **23:** 3073–3079

Smith B, Ashburner M, Rosse C, Bard J, Bug W, Ceusters W, Goldberg LJ, Eilbeck K, Ireland A, Mungall CJ, Consortium O, Leontis N, Rocca-Serra P, Ruttenberg A, Sansone SA, Scheuermann RH, N NS, Whetzel PL, Lewis S (2007b) The OBO foundry: coordinated evolution of ontologies to support biomedical data integration. *Nat Biotechnol* **25:** 1251–1255

Soldatova L, King R (2006) An ontology of scientific experiments. *J R Soc Interface* **3:** 795–803

Taylor CF, Paton NW, Lilley KS, Binz PA, Julian RK, Jones AR, Zhu W, Apweiler R, Aebersold R, Deutsch EW, Dunn MJ, Heck AJ, Leitner A, Macht M, Mann M, Martens L, Neubert TA, Patterson SD, Ping P, Seymour SL *et al* (2007) The minimum information about a proteomics experiment (MIAPE). *Nat Biotechnol* **25:** 887–893

Tengli A, Yang Y, Ma N (2004) Learning table extraction from examples. Proceedings of the 20th International Conference on Computational Linguistics. Geneva, Switzerland, Association for Computational Linguistics: Article no. 987

Tschopp J, Martinon F, Burns K (2003) NALPs: a novel protein family involved in inflammation. *Nat Rev Mol Cell Biol* **4:** 95–104

Whetzel P, Brinkman R, Causton H, Fan L, Field D, Fostel J, Fragoso G, Gray T, Heiskanen M, Hernandez-Boussard T, Morrison N, Parkinson H, Rocca-Serra P, Sansone S-A, Smith DSB, Stevens R, Stoeckert C, Taylor C, White J, Wood A (2006) Development of FuGO: an ontology for functional genomics investigations. *OMICS* **10:** 199–204

Winston F, Dollard C, Ricupero-Hovasse SL (1995) Construction of a set of convenient *Saccharomyces cerevisiaestrains* that are isogenic to S228C. *Yeast* **11:** 53–55