3

Semantic Data Integration on Biomedical Data Using Semantic Web Technologies

Roland Kienast and Christian Baumgartner
Institute of Electrical, Electronic and Bioengineering
University for Health Sciences, Medical Informatics and Technology
Austria

1. Introduction

Contemporary life sciences research requires an understanding of systems across wide ranges of scale and distribution. Therefore, there is an urgent need to integrate biomedical knowledge generated by different communities and separate subfields (Shadbolt et al., 2006). Scientific publications and curated databases together hold a vast amount of this useable knowledge. Additionally the number, size, and complexity of life science databases continues to grow (Kei-Hoi et al., 2009). Therefore scientists in the field of genomics, proteomics, metabolomics, clinical medicine and drug discovery need a concept to integrate their data, (Shadbolt et al., 2006) which is a prominent problem (Kei-Hoi et al., 2009). But to generate such a uniform data integration concept there are still some challenges to overcome such as handling the variety and amount of available data, inconsistency with data heterogeneity from the different sources, the autonomy and differing capabilities of the sources and a lack of standards for such an integration concept. Many heterogeneity conflicts remain in data integration due to the lack of semantics (Gagnon, 2007). In order, to efficiently exploit the knowledge from different resources, it will be important to connect the sources in a manner that machine processes can traverse and intelligently identify these links (Neumann et al., 2004). A promising approach to integrate heterogeneous data sources could be the use of *Semantic Web technologies*. They provide a framework to deal with the afore mentioned problems and fulfil the requirements for machine processing.
This book chapter provides an overview of data integration on biomedical data using Semantic Web technologies including existing techniques (standards, specifications and methods), challenges, approaches and projects.

2. Basics of data integration

Data integration is the task of "combining the data residing at different sources, and providing the user with a unified view of the data" (Calì et al., 2001; 2003). But to accomplish the task of combining different heterogeneous sources there are some challenges to be overcome.

2.1 Challenges in integrating information from heterogeneous data sources

In the dictionary[1] heterogeneity is defined as *"the quality of being diverse and not comparable in kind"*. In computer science this inability to compare can be divided into four different classes (Ouksel & Sheth, 1999):

[1] Webster's Online Dictionary `http://www.websters-online-dictionary.org`

- **System heterogeneity** is a result of different hardware platforms and operation systems.
- **Syntactic heterogeneity** is a difference of data representation formats.
- **Structural heterogeneity** rises from different data models or structure in various data sources.
- **Semantic heterogeneity** results from differences in the interpretation of the meaning of different resources.

This heterogeneity leads to some challenges in integrating information from multiple data sources. Some general problems are (Cheung et al., 2007):

- **Locating Resources:** To be able to integrate data it is important to find relevant and inter-operable data sources. But to find such sources it is beneficial to have a widely accepted standard for describing the content of data.
- **Different data formats:** Different resources often provide heterogeneous data formats. For example:
  - *structured data:* e.g. different databases
  - *semi-structured data:* e.g. HTML, XML data
  - *unstructured data:* e.g text documents, images
- **Identify Synonyms and Homonyms:** Before large scale databases where created, researchers independently named biological entities. As a consequence many synonyms exist. The ability to distinguish between synonyms and homonyms is very important for data integration.
- **Detect Ambiguity:** Different terms can be used to represent different concepts. For example the term *insulin* can represent the concept *hormone* or *drug*.
- **Recognize Granularity:** Different biological data sources may provide knowledge at different levels of granularity. For example one source provides information about different genetic diseases and their symptoms. Another source might only contain detailed information about haemophilia[2].
- **Scaling conflicts:** These conflicts occur when different reference systems are used to measure a value e.g., different date formats or size measures.

2.2 Different integration approaches

There are different approaches to integrate different data sources by using *warehousing*, *mediation* or a combination of both.

Warehouse integration consists in catalouging the data from multiple sources into a local database called the *warehouse*. All queries are executed on the data contained in the warehouse (Hernandez & Kambhampati, 2004; Kugler et al., 2008; Pfeifer et al., 2007). The task of importing data from a source into the warehouse is called the *ETL (Extract - Transform - Load)* process.

- *Advantages:* Warehousing eliminates various problems such as network bottlenecks, low response times, and temporarily unavailable sources. It allows to filter, validate, modify, and annotate the data obtained from the sources (Davidson et al., 1995).

[2] Haemophilia is a genetic disease which interferes with blood clotting.

- *Disadvantages:* It is necessary to build and maintain the warehouse and there is a danger of antiquated data. Therefore the warehouse system must regularly check the underlying sources for new or updated data and modify the local copy of the data if required (Davidson et al., 1995).

Mediator based integration concentrates on query translation. A mediator is a system which provides a query translation from a single mediated schema to the local schema of the underlying data source (Hernandez & Kambhampati, 2004). The data flow between mediators and data sources is provided by software components called *Wrappers*. Unlike warehousing, data is not centrally stored but it is accessed directly from the distributed sources.

- *Advantages:* The data is always up to date and there is no need to maintain a storage system.
- *Disadvantages:* Mediator based integration is sensitive to network bottlenecks, low response times and temporarily unavailable sources.

An other possibility is using Semantic Web technologies. The goal of the Semantic Web approach to data integration is to add machine readable metadata to resources and to define and describe relations among them. This makes it easier to automatically process and integrate information available within the different resources (W3C, 2004a) (see figure 1).

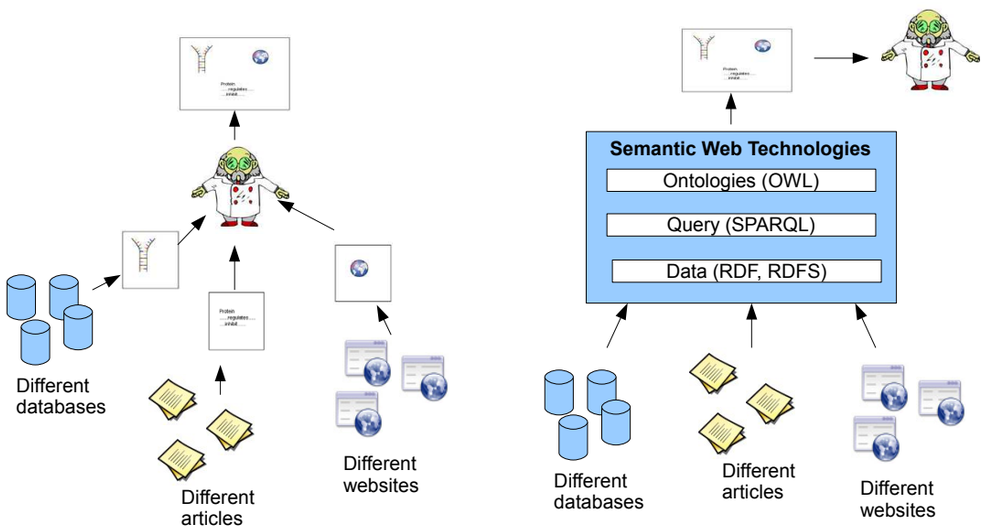


Fig. 1. The goal of data integration using Semantic Web technologies. *Right:* The user must consult several resources individually through different user interfaces to derive a result. *Left:* Semantic Web technology allows the integration of various heterogeneous resources. The system can process the data and provide the results to the user.

3. Semantic Web in a nutshell

Tim Berners-Lee , the director of the World Wide Web Consortium (W3C), coined the term Semantic Web (Berners-Lee et al., 2001) and it is mainly used to describe the model and technologies provided by the W3C which is the main international standards organization for the World Wide Web. The aim of the Semantic Web is to add structured meta-information to

existing documents and data in order to give it a well defined semantic meaning. This enables machines to process semantic information but *"not human speech and writings"* (Berners-Lee et al., 2001). This semantic extension makes it easier for machines to automatically process and integrate information available on the Web (W3C, 2004a).
The basic idea behind the Semantic Web is to add machine readable metadata[3] to resources within the World Wide Web to define and describe relations among them. Semantic Web technologies are able to assimilate this gained information. Furthermore, they do not build a separate web, but function as an extension of the current web. The Semantic Web technology consists of a hierarchical use of various standards and technology in which each layer uses the capabilities of the layers below. The architecture of the semantic web is illustrated in figure 2. A brief description of each layer is summarized below:

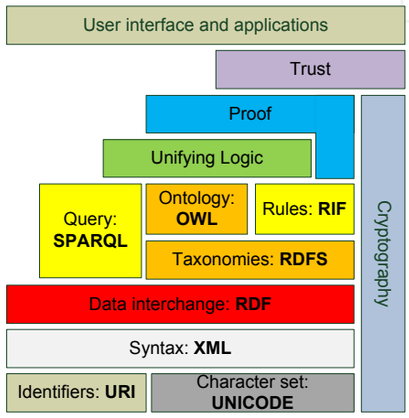


Fig. 2. Semantic Web stack

- **Character Set: UNICODE** defines a fundamental coding standard for data.
- **Identifiers: URI** is a standard for the identification of resources.
- **Syntax: XML** provides a fundamental syntax for structured documents.
- **Data interchange: RDF** is a data model for resources and relations between them. It uses the XML syntax.
- **Taxonomies: RDFS** is an extension of RDF and provides a vocabulary for describing RDF resources.
- **Rules: RIF** defines the rules of semantic data.
- **Ontologies: OWL** offers more opportunities to add semantic information to resources than RDFS.
- **Query: SPARQL** is a protocol and a query language for RDF.
- **Unifying Logic** allows to draw a conclusion.
- **Proof** attempts to verify the conclusions.
- **Trust** provides trusted principles and authentification methods between different agents.

[3] According to the Dictionary of Computing (`http://dictionary.reference.com/browse/meta-data`) metadata is *"definitional data that provides information about or documentation of other data managed within an application or environment."* In relation to the Semantic Web Tim Berners-Lee defines metadata as (Berners-Lee, 1997): *"machine understandable information about web resources or other thing"*. In short, metadata is data about data.

4. Semantic Web approach to data integration

The W3C defines the abilities of the Semantic Web as follows (W3C, 2011):
"The Semantic Web is about two things. It is about common formats for integration and combination of data drawn from diverse sources, where on the original Web mainly concentrated on the interchange of documents. It is also about language for recording how the data relates to real world objects. That allows a person, or a machine, to start off in one database, and then move through an unending set of databases which are connected not by wires but by being about the same thing."
Semantic Web approach to data integration can deal with heterogeneity by providing structured meta-information to existing documents and data. A key feature integrating information is the use of semantics which gives meaning to a word or concept (Gardner, 2005). Semantics can solve the problem of homonyms and synonyms between different sources because it is able to ensure the equivalence of two concepts which might have different names and forms (synonyms) or the dissimilarity of two concepts which might have the same name and form (homonyms). Semantics describe relationships between concepts. This enables a fully descriptive representation of the available information, showing the interaction between concepts and allows inferences. Semantic Web technologies provide a tool to describe such semantic: The use of ***Ontologies***. In order to achieve a beneficial use of ontologies, it is important to link the data to its semantic knowledge. In other words, it is important to annotate instances to ontologies. But these data often have different data formats (relational databases, text files, web sites, etc.). Adding metadata can solve this problem. But to benefit from this metadata, it should be standardized and machine readable. Such a kind of metadata provided by the Semantic Web technology is based on the Extensible Markup Language (XML).

4.1 Important technologies for data integration in greater detail

This section describes the most important technologies which are needed for a semantic data integration based on Semantic Web technologies.

4.1.1 URI (Uniform Resource Identifiers)

A URI is defined in RFC3986 (Berners-Lee et al., 2005): *"A Uniform Resource Identifier (URI) is a compact sequence of characters that identifies an abstract or physical resource."* In the web URIs typically refer to websites or other data. But in general URIs can be used to generate unique identifiers for different resources. For example the namespaces of a XML (Extensible Markup Language) document are identified by URI references. Also, in RDF (Resource Description Framework), URIs are used to refer to resources (Hitzler et al., 2008).

4.1.2 XML (eXtensible Markup Language)

XML is a machine readable, standardized meta-language. It is an important basic technology for the Semantic Web (W3C, 2001) with which it is possible to create structured documents. These documents are text based and provide their data in a hierarchical and logically structured form which can be read by humans and by machines. It is an markup based language and uses tags for this purpose. In Informatics markup languages are used to extend parts of an document with additional information to describe it in more detail. This additional information is also called *metadata*.
Problems with XML and data integration:
XML is standardized, machine readable and defines the syntactical structure of a document. But in the view of the Semantic Web, XML tags are not much better than the natural language

(Hitzler et al., 2008). These tags can be ambiguous, their relationship is not clearly defined and they provide no meaning for machines.

4.1.3 RDF (Resource Description Framework)

Originally RDF was designed for adding metadata to web resources but it has become a framework for adding semantic information to resources. RDF is machine readable. Therefore it enables the encoding, exchange, and reuse of structured metadata and allows structured and semi-structured data to be mixed, exposed and shared across different applications(W3C, 2010a) which can make use of the semantic information (Fensel, 2004).
RDF provides a simple data model for describing relationships between resources in terms of named properties and their values. While XML can only describe documents in a tree structure, RDF is a framework for representing information about resources in the form of a directed graph. An edge of this graph describes the relationship between two resources. RDF documents can be written in Notation 3 (N3) (W3C, 2005), N-Triples (W3C, 2004d), Turtle (W3C, 2008c) syntax or in a XML syntax. This XML syntax is called RDF/XML (W3C, 2004f). But XML can only describe a tree structure whereas RDF represents a graph. Therefore it is necessary to *serialize* these complex data objects into strings. RDF uses so-called *"triples"* (3-tuples) to describe relationships between resources to serialize the graph. A RDF-triple consists of only three elements (W3C, 2004g):

1. *The subject:* Is a RDF URI reference or a blank node.
2. *The predicate:* Is a RDF URI reference.
3. *The object:* Is a RDF URI reference, a blank node or a literal.

A triple is conventionally written in the order subject, predicate, object and can be illustrated by a node and directed arc diagram (see figure 3). A set of these triples form a directed graph.
A problem in RDF is that URI references can not describe a conclusive semantic interpretation

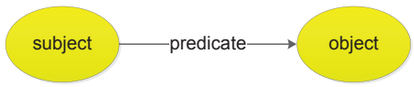


Fig. 3. Illustration of a triple

of RDF coded information (Hitzler et al., 2008) because a URI can also be a homonym or synonym of another URI. This principle is also known as *Non Unique Name Assumption*. A solution to this problem is to use thematic vocabularies such as FOAF (Friend of a Friend) vocabulary which can be used for linking people and information about them (Brickley & Miller, 2010).

4.1.4 Ontologies to share semantic information

(Gruber, 1993) defines an ontology as: *"An ontology is an explicit specification of a conceptualization."* This definition was slightly modified by (Studer et al., 1998): *"An ontology is a formal, explicit specification of a shared conceptualization."*
A *conceptualization* refers to an abstract model of a phenomenon in the world which identifies the relevant concepts of that phenomenon. *Explicit* correlates to the formed types of concepts and their limitations, which are defined explicitly. *Formal* is based on the fact that an ontology should be machine readable. *Shared* means that an ontology should cover matching knowledge. This knowledge is not limited to an individual and is accepted by a group (Fensel, 2004; Studer et al., 1998).

This abstract definition is understandable on the basis of a simple example. It contains a brief abstract of the ontology of animals (see figure 4).
The abstract model includes the terms *animal*, *fish*, *mammal* and *puma*. These terms come from

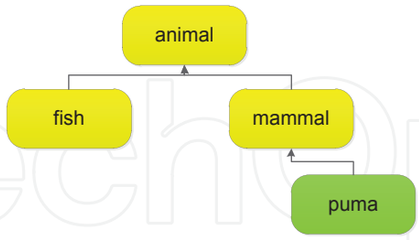


Fig. 4. A brief abstract of the ontology of animals.

the "phenomenon" of animals. Every term is explicit. The term puma is explicitly defined as a animal. It cannot be confused with the clothing brand Puma. Puma also have clear limitations: a puma is an organism which is a mammal and not a fish and belongs to the animals. An ontology is represented as a directed graph. A graph is formal and machine readable. The ontology is also shared because not only one individual can infer knowledge and it is accepted by a group of biologists.
The structure of an ontology is a directed acyclic graph. That makes it possible to support complex relationships which allow terms to have more than one parent. For example the Gene Ontology[4] term *GO:0070229 : negative regulation of lymphocyte apoptosis* is a subclass of *GO:2000107 : negative regulation of leukocyte apoptosis* and *GO:0070228 : regulation of lymphocyte apoptosis*. Ontologies are able to describe the semantic of the information sources in order to make their content explicit. A basic module of ontologies is the so called *"triple"*. Broadly defined, a triple contains two terms and a relation between them[5]. With these elements an ontology can be represented as a directed graph. The terms are the nodes and the relations are the edges of the graph (Smith et al., 2005).

4.1.5 RDFS (RDF Schema)

Like XML, RDF only provides a syntax for exchanging data. RDF properties can be considered as attributes of resources and also represent relationships between them. But it provides no mechanisms for adding a vocabulary to describing these attributes or relationships. RDFS, or also called *RDF Vocabulary Description Language*, extends RDF to describe such vocabularies (W3C, 2004c;e) and add terminological knowledge (schema knowledge) to this vocabulary. For that reason it can be seen as a semantic extension of RDF. RDF Schema vocabulary descriptions are written in RDF syntax (W3C, 2004e). It makes statements about the semantic relationship between terms within an arbitrarily defined vocabulary inside a RDFS document. This ability to define terminological knowledge allows RDFS to create *"light-weight"* ontologies (Hitzler et al., 2008; Volz et al., 2003) to describe semantic dependences within a domain.
Figure 4 shows a simple RDFS document in graph representation. RDFS organize RDF statements hierarchically into classes (terminological knowledge) and instances (assertional knowledge). Properties are used to describe relationships between classes. The terminological part includes the ontology while the assertional part presents conclusions about concrete

[4] see section 5.3.2
[5] see section 4.1.3 for a detailed description

qualities of the subject. This ontology describes, for example, that the class cell is a subclass of the class organ and that every cell consumes energy. Further, it is possible to derive implicit knowledge. If the muscle cell is an instance of the class cell and the ATP (Adenosine Tri-Phosphate) is an instance of high energy chemical bond, then it is possible to infer that a muscle cell is part of a human and ATP is a kind of energy.

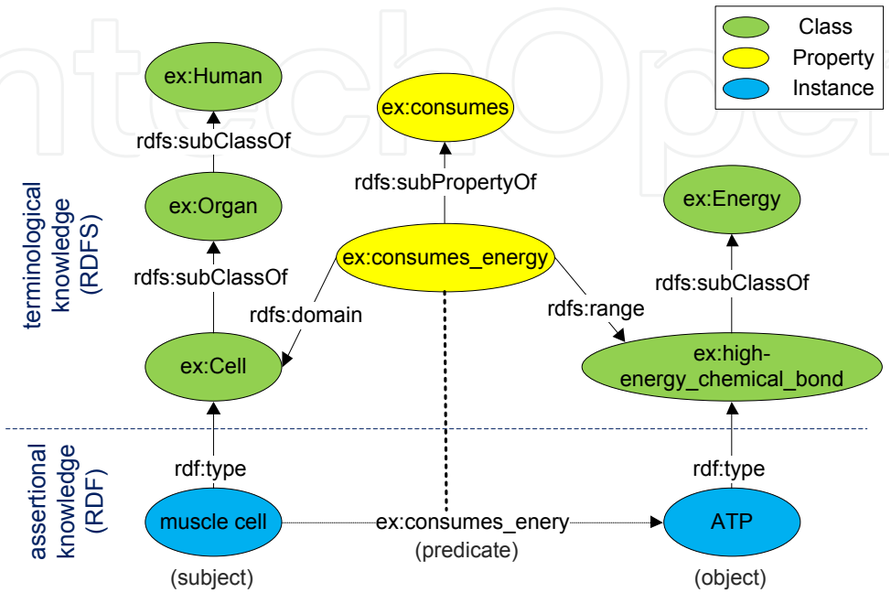


Fig. 5. Simple RDFS-Ontology in graph representation

4.1.6 OWL (Web Ontology Language)

OWL is designed to enable machine processing of information content. OWL can explicitly represent the meaning in terms of vocabularies and their relationship with each other to build an ontology. Since October 2009 the version OWL 2 is recommended from W3C(W3C, 2009a). In contrast to RDFS, OWL has more opportunities to expressing meaning and semantics. Therefore OWL can be seen as an extension of RDFS (W3C, 2004b). An OWL ontology is an RDF graph which consists as a set of triples. It also can be written in different syntactic forms but the most common syntax is RDF/XML for representing these triples.
OWL provides three increasingly expressive sub-languages (Alesso & Smith, 2006):

1. **OWL Lite** to generate a classification hierarchy and simplify constraints. -> *Easily implementable*
2. **OWL DL** (description logic) supports maximum expressiveness while retaining computational completeness and decidability. -> *Mechanizable logic*
3. **OWL Full** provides maximum expression and syntactic freedom of RDF but with no computational guarantees. -> *Complete Logic*

Since OWL is an extension of RDFS and therefore also from RDF, any RDF document will generally be in OWL Full. OWL DL and OWL Lite also extend the RDF vocabulary, but they put restrictions on the use of this vocabulary (W3C, 2004a;b) for better machine processing. These restrictions guarantee computational completeness and decidability of reasoning systems like FaCT++ (Tsarkov & Horrocks, 2006) and the Pellet (Sirin et al., 2007)

which are able to reason over OWL 2 ontologies (Grau et al., 2008). This is achieved because OWL Lite and DL are basically very expressive description logics (DL) where OWL DL is based on the $\mathcal{SHOIN}(D)DL$ (Hitzler et al., 2008) and OWL Lite to the slightly simpler $\mathcal{SHIF}(D)DL$.
Description Logics (DL) stem from semantic networks (Donini et al., 1996). They model *concepts* (equal to a class in OWL), *roles* (equal to a property in OWL) and *individuals* (equal to a object in OWL), and their *relationships*. Therefore they can be used to represent the knowledge of a specific domain in a formal and structured way. Here the context of ontologies is clearly visible. As described in 4.1.4 an ontology consists of axioms, which are used to provide information about classes and properties of a specific domain. The knowledge which is provided by DL is divided into a *TBox* and an *ABox* (Donini et al., 1996). The TBox (terminological box) contains sentences describing concept hierarchies and the ABox (assertional box) contains sentences about the individuals and where they are in the hierarchy (Van Harmelen et al., 2008). For example the statement *"Every protein is made of amino acids"* belongs to the TBox, while the statement *"Leucine is a amino acid"* belongs to the ABox.
The drawing of logical conclusions in OWL are based on the concept of the so-called *Open World Assumption (OWA)*. In contrast to the *Closed World Assumption (CWA)*, this assumption specifies that statements are neither true nor false if they can not be derived from a set of facts based on inference rules. The OWA does not assume that a answer is false unless it can be absolutely proven that the answer is false (Pollock, 2009). Listing 1 shows an example of both assumptions.

Listing 1. Example for the open- and closed world assumption

```
Knowledge Base:          The protein p53 is involved in apoptosis.
Query:                   Is the protein p53 involved in cell repair?
Answer:                          CWA: No.
                                 OWA: Maybe or unknown.
```

4.1.7 SPARQL (Simple Protocol and RDF Query Language)

SPARQL is a protocol and query language for RDF which since January 2008 is an official W3C recommendation (W3C, 2008a). SPARQL queries often contain a set of triple patterns. These patterns, or also called basic graph patterns, look like RDF triples. The difference is that every subject, predicate or object, can be expressed as a variable. A match can be found by replacing variables through substituting RDF terms. If the result of the substitution is equivalent to a subgraph of the RDF data a match is found. For example, to find the meaning of the acronym *ATP* and where it is produced, the SPARQL query would look like Listening 2.

Listing 2. Simple SPARQL query

```
PREFIX ex: <http://example.com/>
SELECT ?longName ?part
WHERE
{ex:ATP ex:hasLongName  ?name.
?name   ex:producedIn   ?part}
```

The SELECT clause defines the variables which appear in the result and the WHERE clause provides the basic graph pattern. In this case the graph pattern consists of two triple patterns with two single variables.
As a simple knowledge basis following RDF data (see Listing3) in Turtle notation is used.

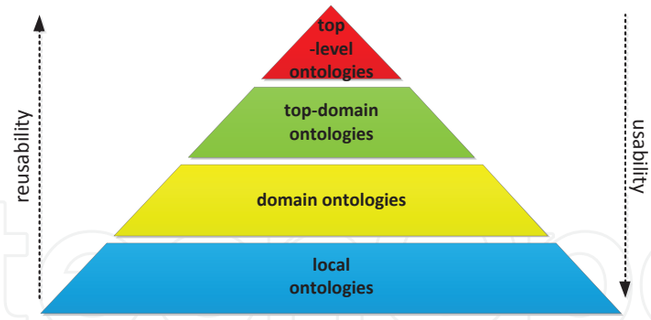


Fig. 6. Classification of ontologies. The reusability decreases with increasing specification. The availability behaves exactly opposite.

Listing 3. Simple RDF data

```
PREFIX ex: <http://example.com/cell>
ex:ATP  ex:hasLongName  "Adenosine␣Tri-Phosphate"
ex:ATP  ex:producedIn   ex:mitochondrion
```

Querying this RDF 3 data with the SPARQL query 2 obtained the result shown in table 1. It

name	part
"Adenosine Tri-Phosphate"	*http://example.org/cell/mitochondrion/*

Table 1. Result of SPARQL Query 2 on RDF Data 3

is also possible to generate complex graph patterns out of a number of simple patterns or to define filters to restrict the result. SPARQL provides four query forms which form a result `SELECT, ASK` set or RDF graphs `CONSTRUCT, DESCRIBE` out of the pattern matching. To serialize a result from a `SELECT` or from an `ASK` query into a XML document the *SPARQL Variable Binding Results XML Format*(W3C, 2008b) can be used.

5. Using ontologies for data integration

Biomedical ontologies play an important role in the process of data integration and support both approaches for data integration: warehousing and meditation (Bodenreider, 2008). Ontologies are a type of controlled vocabulary that attempt to capture the knowledge of a specific domain. This is the standardization required from *warehousing approaches*, where different sources are transformed into a common format and converted to a common vocabulary. On the other hand, the *mediation-based approach* ontologies can be used for defining global schema and mapping between the global schema and local schemes of the sources to integrate. An example of a system using this approach is *ONTOFUSION* (Perez-Rey et al., 2006). The terminological part of ontologies, which contain a list of names for the entities represented in these ontologies, is also an important resource for natural language processing (Altman et al., 2008).
Based on their granularity, ontologies can be divided into four classes (see figure 6):

- **Top-level ontologies** describe very general concepts which are independent of a particular problem or domain (Guarino, 1998) and are highly reusable across specific domains.

- **Top-domain ontologies** contains core concepts of a given domain. For example: *Organism* or *Cell* for a biological domain. They work like an interface between top-level and domain ontologies (Stenzhorn et al., 2008).
- **Domain ontologies** include only domain specific concepts and therefore only describe a certain domain.
- **Local ontologies** describe the semantic of a single information resource.

The ability of ontologies to provide a map of concepts in relationships enables semantic data integration. In this context, ontologies are used to describe the semantics of the data sources in order to make their content explicit (Boury-Brisset, 2003). The integration can take place on an extremely granular level to map data from different resources, no mater if the resources contain structured or unstructured data (Gardner, 2005).
Ontology-based approaches to data integration usually provide a three-layer architecture where a semantic layer working as a mediator is between the presentation layer and the physical layer. This semantic mediator exploits mapping models and transforms queries into execution plans. Wrappers exploit the description of the data sources at the physical layer. This enables a transparent access to diverse data sources by using a unified query language (Boury-Brisset, 2003) like SPARQL. Ontologies are used in the mediator layer because they provide a common vocabulary for the integration of data, where each concept has a unique defined name, associated properties and clearly defined synonyms. Furthermore, an ontology is not a rigid structure, it can grow with time and can be connected to other ontologies.
Wache (Wache et al., 2001) describes three approaches for ontology-based data integration:

- **Single ontology approach:** This approach uses only a single global ontology to integrate different sources. All information sources are related to the global ontology. The global ontology can be a combination of different specialized ontologies. This approach requires data sources with a similar view on the domain and a similar granularity. A disadvantage of this approach is that the integration of new information sources can lead to big changes in the used ontology.
- **Multiple ontologies approach:** The semantic of an source is described by its own local ontology. There is no common vocabulary and therefore inter-ontology mapping is required. An advantage of this approach is that new data sources, and their local ontologies, can be easily integrated. But the lack of common vocabulary can make the mapping between ontologies very difficult to define.
- **Hybrid approach:** This is a combination of the two preceding approaches. As with the multiple ontologies approach, resources are also described by local ontologies. But to avoid the disadvantages and to make these ontologies comparable, they are built from a shared global vocabulary. This vocabulary contains basic terms of a domain and allows querying through a shared vocabulary. The vocabulary can also be an ontology. Then it is also possible to dispense with the mapping between the local ontologies and only define mappings between the shared global ontology and the local ones. New sources can be easily added with no need to modify existing mappings.

An example of using ontologies for data integration in biomedicine is the Gene Ontology Annotation (GOA) [6] project run by the European Bioinformatics Institute (EBI). GOA is based on the single ontology approach and has as target to provide *"high quality electronic and manual"* annotations to the UniProt knowledgebase [7] (UniProtKB)(Barrell et al., 2009). For

[6] http://www.ebi.ac.uk/GOA
[7] http://www.ebi.ac.uk/uniprot

this purpose, GOA uses the standardized vocabulary of the Gene Ontology (GO) 5.3.2 and the International Protein Index (IPI) (Camon et al., 2004). The IPI offers complete, non redundant data sets representing the human, mouse and rat proteomes (Kersey et al., 2004).
Another advantageous feature of ontologies is that terms are organized in a hierarchical manner (Stein, 2003). That means more specific terms are specializations of more general terms. This could help to find the most specific common term shared by two data sources. An example of such a benefit could look like the following:
One research group might create a database in which gene products annotated to the *"negative regulation of T cell apoptosis"*-class of the Gene Ontology. Another group might identify gene products which negatively regulate the programmed cell death. If both groups use the terms of the GO, the two databases can be integrated by finding the most specific common term by traversing up the hierarchy (see figure 7). Without such an organized hierarchy of common concepts, the integration task comes down to tedious and error-prone work by hand (Stein, 2003).

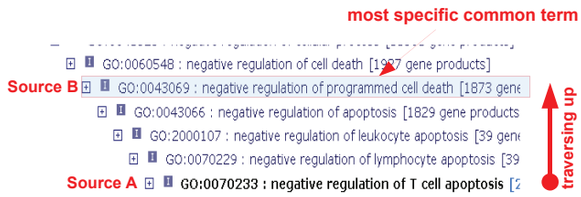


Fig. 7. Find the most specific common term by traversing up the hierarchy.
(This figure shows an extract of the Gene Ontology `http://www.geneontology.org`)

5.1 Examples of existing top-level ontologies

5.1.1 Descriptive Ontology for Linguistic and Cognitive Engineering (DOLCE)

DOLCE is the first module of the WonderWeb [8] foundational ontologies library. *"It aims at capturing the ontological categories underlying natural language and human commonsense."* (Masolo et al., 2003). The Dolce foundational ontology and its extensions provide a domain-independent framework to build ontologies on the basis of highly-reusable patterns.

5.1.2 Basic Formal Ontology (BFO)

The BFO is narrowly focused on the task of providing a genuine upper ontology which can be used in support of domain ontologies developed for scientific research, for example in biomedicine within the framework of the OBO Foundry (IFOMIS, Saarland University, 2010).

5.2 Examples of existing top-domain ontologies

5.2.1 The Unified Medical Language System (UMLS)

Having identified terminology is a key factor for data integration (Bodenreider, 2004) therefore the UMLS was developed by the National Library of Medicine (NLM)[9] and consists of three knowledge Sources which can be used separately or together (U.S. National Library of Medicine, 2010):

- **Lexical resources:** SPECIALIST lexicon: Intends to be a general English lexicon which includes many biomedical terms.

[8] `http://wonderweb.semanticweb.org`
[9] `http://www.nlm.nih.gov`

- **Terminological resources:** Metathesaurus: Includes biomedical and health related source vocabularies, concepts and the relationships between them.
- **Ontological resources:** Semantic Network: Contains categorization of all concepts represented in the UMLS Metathesaurus and relationships between these categories.

The *Semantic Network* (SN) can be seen as a collection of ontologies. In order to use these with Semantic Web technologies it is necessary to convert the SN to OWL DL. There are some approaches to map or convert UMLS SN to RDF (Zeng & Bodenreider, 2007), to OWL (Kashyap & Borgida, 2003; Schulz et al., 2009) or only parts to OWL (Chabalier et al., 2007). But there are formalism problems concerning this task like the complex semantics or the rich attribute set of the UMLS SN.

5.2.2 BioTop

BioTop is a top-domain ontology for the Life Sciences with the goal to provide *"an ontologically sound layer for linking and integrating various specific domain ontologies from the life sciences domain."* (Beisswanger et al., 2008).

5.3 Examples of existing domain ontologies

5.3.1 Open Biological and Biomedical Ontologies (OBO)

The OBO Foundry is a collaborative experiment involving science based ontology developers. The goal is to create orthogonal inter-operable reference ontologies in the biomedical domain (OBO Foundry, n.d.). These ontologies typically have the OBO flat file format. Like OWL, OBO is also an ontology representation languag (Richter, 2006). Ontologies based on the OBO flat file format can be bi-directionally converted to the OWL-DL format (Aranguren et al., 2007; Hoehndorf et al., 2010; Smith B. et al., 2007). The two most significant OBO are the Gene Ontology (GO), which contains the principle attributes of gene products, and the Sequence Ontology, which describes the features of biological sequences.

5.3.2 Gene Ontology (GO)

The GO project[10] contains defined terms which represent gene product properties. The GO covers three aspects of separate ontologies(Gene Ontology, n.d.):

- **Molecular function:** the elemental activities of a gene product at the molecular level, such as binding or catalysis.
- **Biological process:** operations or sets of molecular events with a defined beginning and end, pertinent to the functioning of integrated living units: cells, tissues, organs and organisms.
- **Cellular component**: the parts of a cell or its extracellular environment.

5.3.3 Sequence Ontology (SO)

The SO Project[11] contains defined terms which describe the features and properties of biological sequences. SO is a sister project of the GO and also part of OBO (Eilbeck et al., 2005).

[10] http://www.geneontology.org
[11] http://www.sequenceontology.org

6. Relational database integration using the Semantic Web Approach

A lot of biomedical data is available to the scientific community on the web. Much of this information is stored in a variety of different databases. The content of these databases differ from the type of biological data they provide (Baker & Cheung, 2007). For example:

- *Sequence databases* like EMBL Nucleotide Sequence Database (EBI, n.d.a) or NCBI's GenBank (NCBI, 2004).
- *Microarray gene expression databases* like the EMBL ArrayExpress Archive (EMBL-EBI, n.d.), NCBI's Gene Expression Omnibus (GEO)(NCBI, n.d.) or the Stanford Microarray Database (SMD) (Stanford University, n.d.).
- *Pathway databases* like KEGG (Kanehisa-Laboratories, n.d.) or the Human Protein Reference Database (HPRD) (Keshava Prasad et al., 2008).
- *Proteomic Databases* like the UniProt (EBI, n.d.b).

Computational analyses of biological data often require using multiple datasets. Currently, the integration of different data sets usually happens manually. This approach is very time consuming which requires integrated datasets with rich, flexible and efficient interfaces (Smith A. et al., 2007).

6.1 Problems of heterogeneous database integration

- **Technical heterogeneity** results from different access protocols, file formats, query languages and so on.
- **Data model heterogeneity** arises because of different models storing the same data.
- **Semantic heterogeneity** occurs during combination of different databases with various but related data. For example combine a gene database to a protein database. A gene may have gene products and therefore these two databases are related.

Resolving such heterogeneity and enabling database integration is a key problem which the Semantic Web aims to address (Baker & Cheung, 2007). Therefore a mapping language between RDF and relational databases called RDB2RDF is under development.

6.2 RDB2RDF

A workshop hosted by the W3C on *"RDF accesses to Relational Databases"* in October 2007 resulted in creating a RDB2RDF Incubator Group (W3C, 2010b), which operated from 2008 to 2009. The objective of this group was to create a group to develop a standardized mapping language between RDF and relational databases (W3C, 2009c). The resulting RDB2RDF working group started in 2009 with: *" The mission of the RDB2RDF Working Group, part of the Semantic Web Activity, is to standardize a language for mapping relational data and relational database schemas into RDF and OWL, tentatively called the RDB2RDF Mapping Language, R2RML."* (W3C, 2009b). The results of this working group are scheduled for release September 30^{th},2011.
The RDB2RDF mapping language could be used in two ways (see figure 8):

1. To extract the data from the relational database and store the content in RDF. In this case the data is physically converted to RDF in a ETL (Extract-Transform-Load) and then stored in a RDF triple store. An advantage of this approach is its easy implementation. A disadvantage is that there is always a separate copy of the relational data.
2. To generate virtual mapping between the Semantic Web technologies and the relational database. This virtual mapping queries via SPARQL which will be translated into SQL queries on the underlying relational data.

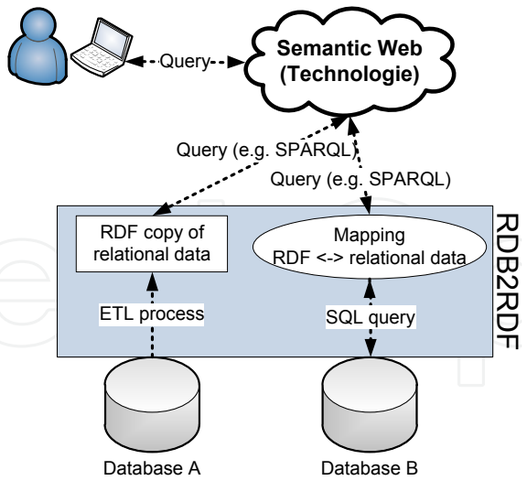


Fig. 8. Two approaches which use the RDB2RDF mapping language

7. Data integration and knowledge acquisition from biomedical literature

The quantity of biomedical literature is steadily growing with a rate of several thousand papers per week (Ananiadou et al., 2006). A large percentage of information is encoded in literature (Krallinger et al., 2008). But for a scientist it is next to impossible to read all relevant literature on a specific topic. Therefore it is important to extract semantic information out of literature to enable machine processing. This section provides an overview of how Semantic Web technologies support this task. Ontologies in particular are able to handle this influx of information and enable the data integration of biomedical literature (Spasic et al., 2005). Basic techniques to extract information from natural language are text mining (TM) and natural language processing (NLP).
Sections of TM are:

1. *Information retrieval (IR):* Retrieve of relevant documents.
2. *Information extraction (IE):* Extraction of relevant information from the document.
3. *Data mining (DM):* Discover of associations between information extracted by IE.

7.1 Information retrieval (IR)

The process of IR can be improved by adding a semantic layer. This layer formulates semantic queries, offering a higher expressive power than keyword matching (Spasic et al., 2005). However, adding semantic information to enhance the process of finding relevant information is generally a main part of Semantic Web technology. An example of such query systems are:

- **GoPubMed** (`www.gopubmed.org`): This system submits keywords to PubMed[12]. The resulting abstracts are matched against Gene Ontology and Medical Subject Headings (MeSH) (Doms & Schroeder, 2005) to be classified. To find a match, a term extraction algorithm based on local sequence alignment is used (Delfs et al., 2004). In other words GoPubMed organize the results of a PubMed search using the GO.

[12] PubMed (`http://www.pubmed.gov`) is a literature database provided by the National Library of Medicine and the National Institutes of Health.

- **Textpresso** (`http://www.textpresso.org`): A tool for neuroscience which has its own literature filled database. It uses a custom ontology to query nine different categories (Müller et al., 2008).

7.2 Information Extraction (IE), Data Mining (DM)

There are two ways to enhance the process of IE respectively, use TM and NLP supporting "literature data integration" based on Semantic Web technologies:

1. Ontology assisted extraction of meta-information from literature.
2. Semi-automatic or automatic engineering of ontologies by a specific domain based on information extracted from literature.

Generally, text mining is used to aid experts in extracting knowledge from a large volume of text by automatically filtering relevant information. A known problem is to find terms which represent specific classes of biomedical entities (e.g. protein names). This process is called *Named Entity Recognition* (NER). The integration of knowledge, supported by ontologies, can improve NER. The goal is to extract terms and map them to concepts of a domain specific ontology. A challenge in this process is the myriad variations of terms used to describe things in natural language. Approximately one third of term occurrences are variants (Jacquemin, 2001) and therefore only synonyms of known terms. Another problem is the specific terminology in biomedical texts. To have terminological knowledge is of vital importance to TM for characterizing knowledge in the domain. This knowledge is stored in ontologies and can enhance the process of IE by (Spasic et al., 2005):

- Using Ontology as a training set for NER by reducing it to a list of classified terms. This can be done in two ways:
  - *Passive ontology use (Ontology-based IE):* The goal of this approach is to map recognized terms in ontology concepts by look-up.
  - *Active ontology use(Ontology-driven IE)*: involves ontologies directly in the process of term recognition.
- Using ontologies to improve machine learning approaches for TM tasks, such as term classification, term clustering and term relation extraction.

7.3 Semi-automatic or automatic ontology engineering

An advanced task is semi-automatic or automatic engineering of ontologies from a specific domain on the basis of information extracted from literature. Currently the development of ontologies *"is largely a manual process, based on personal experience and intuition"* (Alexopoulou et al., 2008). Two primary parts of this process are:

1. Extracting terms which represent a concept in the specific domain.
2. Finding relationships between different concepts.

For an automatic terminology development it is important to extract terms from a text. This automatic identification of possible candidates for terms is called *automatic term recognition (ATR).* At the moment ATR is not able to fully automate the process of ontology design, but it can speed up this process by providing lists of useful domain-specific terms extracted from domain specific literature. Therefore it can support a semi-automatic creation of ontologies (Alexopoulou et al., 2008). Examples of frameworks which support ATR and further identify the semantic relations between them are:

- *Text2Onto:* This is a framework for ontology learning from textual resources. It is based on algorithms calculating the relative term frequency (Cimiano & Volker, 2005).
- *OntoLearn:* OntoLearn is based on a linguistic processor and a syntactic parser. It is able to extract syntactically plausible terminological noun phrases (Navigli & Velardi, 2004; Velardi et al., 2005).

8. Challenges in data integration using Semantic Web technologies

8.1 Uniform naming

A challenge faced by data integration is the individual naming of objects. For example a KEGG[13] entry refers to a collection of proteins involved in a pathway whereas a UniProt entry refers to a class of proteins, a class of variant proteins or some viral protein. To integrate these two resources mapping is required. One approach is to designate an authoritative names commission to manage the definitive list of such names (Stein, 2003). An example is the HUGO Gene Nomenclature Committee [14] for gene names and symbols (short-form abbreviation). But because of the dynamic in the field of biomedical research this approach rarely work in practice (Stein, 2003).
Another way could be the creation of globally unique biological identifiers. For this purpose URIs can be used which allows for the unique identifying of resources. This is central for the use of Semantic Web technologies. Therefore a process is needed which routinely assigns URIs to objects (Shadbolt et al., 2006) to create common, shared identities and names (Goble & Stevens, 2008).

8.2 Extraction of the semantic information out of existing knowledge

For efficient use of Semantic Web technologies, it would be useful to automatically or semi-automatically extract the semantic information from existing sources. Therefore a big challenge is to develop methods which support such a task. This would aid two main tasks in data integration using Semantic Web technologies:

1. **Annotate sources to existing ontologies:** This is a process which extracts information from the data source to automatically or semi-automatically annotate this source to an existing ontology.
2. **Creation process of ontologies:** This is a task which extracts information from different data sources belonging to a specific domain. The goal is to automatically or semi-automatically create an ontology based on the extracted domain information.

A large percentage of information encoded in literature (Krallinger et al., 2008) is in the form of natural language. Some approaches for such "semantic information extraction" from literature can be found in section 7.

8.3 Ontology development, maintenance and quality

Ontologies must be developed, managed and endorsed by committed practice communities (Shadbolt et al., 2006). Furthermore, an ontology is a "living structure" which means that concepts can change constantly because of new knowledge. They can be added, changed, replaced or removed. Therefore ontologies are not fixed for all time and must be constantly maintained. Another problem is the quality assurance (QA) of ontologies. According to Gruber (Gruber, 1995) design and quality criteria for ontologies should be:

[13] KEGG: Kyoto Encyclopedia of Genes and Genomes (`http://www.genome.jp/kegg/`)
[14] `http://www.hugo-international.org/comm_genenomenclaturecommittee.php`

1. *Clarity:* The intended meaning should be clearly defined and the definitions should be objective.
2. *Extendibility:* The effort needed to extend an ontology without invalidating it.
3. *Minimal encoding bias:* No particular symbol-level encoding should be used to specify terms.
4. *Minimal ontological commitment:* An ontology should use as few terms and relationships as possible to describe the domain being modeled.
5. *Coherence:* The content of the ontology should be coherent. In other words inferences should never contradicts a definition.

The quality of an ontology can be checked either collaboratively by users or centrally, by experts. To test the coherence of an ontology *Ontology-Reasoners* like Pellet[15] could be used. Ontology Reasoning is a process of automated logical inference of knowledge with ontologies. It is used to check the consistency of knowledge models and to infer new knowledge in accordance with the laws of logic.

8.4 Mapping, merging, alignment and integration of ontologies

Many individual ontologies are created and therefore the semantic mapping between different ontologies has become a core issue for the Semantic Web and data integration using its technology. To handle the increasing number of ontologies it is necessary to develop semi-automatic or automatic approaches (Ehrig & Sure, 2004).
The problem with the mapping of ontologies is their heterogeneity which can be divided into *metadata heterogeneity* and *instance heterogeneity* (Tang et al., 2006). Metadata heterogeneity is concerned with the intended meaning of the information held in different ontologies and deal with *structural conflicts* and *name conflicts*. Structural conflicts arise from ontologies which cover the same domain but have different taxonomies (Ehrig & Sure, 2004), and naming conflicts concern homonyms and synonyms between concepts of different ontologies. For instance heterogeneity referreds to the variation in notation different e.g. different date formats.
Merging, aligning and integration is an ontology reuse process to create a new ontology. The task of each process is as follows (Choi et al., 2006; Ding et al., 2002):

- **Merging** is the task of generating a single ontology by merging two or more different ontologies of the same domain.
- **Alignment** is a process of creating links between two ontologies when the sources are consistent but kept separate. This addresses the problem of mapping between ontologies.
- **Integration** generates a single ontology by combining two ore more different ontologies in different subjects.

Data which covers different domains can not often be described by only one ontology. Therefore it is necessary to map different ontologies. There are different strategies for mapping various ontologies:

- *Ontology mapping between a global ontology and local ontologies* (Beneventano et al., 2003): Defines mapping between concepts in local ontologies to global ontology.
- *Mapping between local ontologies:* These strategies define mapping between local ontologies.

[15] Pellet is a OWL 2 Reasoner for Java (`http://clarkparsia.com/pellet`).

8.5 Query RDF data

SPARQL overcomes the old problem of different, non standard query languages. Now it is possible to query RDF data using a standard query language (Quilitz & Leser, 2008). But it is important that content providers integrate SPARQL-endpoints to make their data available. Such endpoints provide a machine-friendly interface towards the knowledge base and enables queries using the SPARQL language. One challenge is to query more than just one endpoint at the same time with only one query. There are several approaches which can be divided into two groups (Haase et al., 2010; Kei-Hoi et al., 2009):

- **Warehousing:** This approach stores all RDF data from the different resources in one central database. This database is typically a *triple store* which is designed to efficiently store and handle RDF data.
- **Federated query**: A query engine decomposes a single query into sub-queries. Each of these queries can be answered by an individual endpoint. After that, all results are combined again into one and represented to the user.

Two examples of Java frameworks are *Sesame*[16] witch supports the warehouse approach and the *ARQ*[17] extension of the *Jena Ontology API*[18] which provides the federated query approach.

8.6 Visualization

The semantic integration of different resources results in increasing the amount of semantically linked data. Semantic Web technologies use RDF, defining links between data. Therefore the challenge is to create an interface to visualize and navigate a massive RDF graph without information overload. The visualization should help the user to easily explore and quickly find relevant information (Le Grand & Soto, 2002) in the structure.

8.7 Availability

There are two issues: The availability of ontologies and content. A key to integrating data using Semantic Web technologies is the availability of ontologies. Many ontologies are freely available but concerns arise if an ontology is commercial or only partially released. For example a license is necessary to access UMLS[19]. On the other side it is important to access content which is annotated to ontologies. But this may cause problems if this content is not available due to technical problems, deleted static web sites and legal restrictions, etc.

8.8 Different ontology formats

The Semantic Web defines ontologies in the OWL format. But other ontologies exists with different formats (for example the UMLS Rich Release Format (RRF) or the OBO format). Therefore, mapping must be defined to convert these different formats to OWL.

8.9 Multilingualism

A challenge is also multilingualism when using Semantic Web technologies (Börner, 2006). It plays a role in ontology development, annotation of data and representing multilingual informations in user interfaces (Benjamins et al., 2002). For example, a scenario that leads to a problem because of multilingualism:
User A annotates a document in French to *Term A* of an ontology designed in English. *User B*

[16] `http://www.openrdf.org/`
[17] http://jena.sourceforge.net/ARQ/
[18] `http://jena.sourceforge.net`
[19] This license is freely available for research purposes

searches for *Term A* in English and finds a document related to what he is interested in, but it is written in French.

9. Discussion

The idea behind the Semantic Web is to transform the Web into a global knowledge base (Kei-Hoi et al., 2009). The key to make this possible is data integration. Therefore Semantic Web technologies offer a more or less standardized hierarchical framework for data integration and enable a decentralized semantic integration of different heterogeneous data sources. For this integration, it is not necessary to change the structure of the data to assemble knowledge from structured and unstructured sources. This technology extends the source by adding machine readable semantic metadata using the Resource Description Framework (RDF). This metadata contains sets of relations between data and concepts. This will enable people to clearly and commonly define the concepts and logic within any document (Neumann et al., 2004). Furthermore, Semantic Web technologies support an automatic traverse of the connected resources. This queries the integrated sources or even infers new knowledge using the standard query language SPARQL. The prerequisite for meaningful semantic data integration is the presence of ontologies. They enable a unique identification of entities in heterogeneous information systems and provide semantic data integration on different granular levels. Semantic Web technologies provide standard languages including the RDF Schema (RDFS), and the Web Ontology Language (OWL) for creating ontologies. The quality of the data integration is tightly correlated with the quality of the used ontologies. But in recent years, many high quality open access biomedical ontologies have been created, such as the Gene Ontology, the Open Biological and Biomedical Ontologies.
In summary, Semantic Web technologies are a promising tool for data integration but there are still some challenges to be overcome such as uniform naming, extraction of the semantic information out of existing knowledge, ontology development, ontology maintenance or query RDF data (see section 8).

10. Additionally

A public available example software, termed *OBOBrowsA*, can be downloaded following the link `http://www.umit.at/page.cfm?vpath=departments/technik/iebe/tools/obobrowsa&switchLocale=en_US`. It is able to load and display OBO files[20] in tree or graph representation. The software further allows the user to interactively browse through the ontology, search for ontology classes and annotate textual data. The manual and application examples are included in the help function.

11. References

Alesso, H. & Smith, C. (2006). *Thinking on the Web: Berners-Lee, Gödel, and Turing*, Wiley-Interscience.

Alexopoulou, D., Wächter, T., Pickersgill, L., Eyre, C. & Schroeder, M. (2008). Terminologies for text-mining; an experiment in the lipoprotein metabolism domain, *BMC Bioinf* 9(Suppl 4): S2.

[20] Link to download OBO formated ontologies: `http://www.obofoundry.org`

Altman, R., Bergman, C., Blake, J., Blaschke, C., Cohen, A., Gannon, F., Grivell, L., Hahn, U., Hersh, W. & Hirschman, L. (2008). Text mining for biology-the way forward: opinions from leading scientists, *Genome Biol* 9(Suppl 2): S7.

Ananiadou, S., Kell, D. & Tsujii, J. (2006). Text mining and its potential applications in systems biology, *Trends Biotechnol* 24(12): 571–579.

Aranguren, M., Bechhofer, S., Lord, P., Sattler, U. & Stevens, R. (2007). Understanding and using the meaning of statements in a bio-ontology: recasting the gene ontology in owl, *BMC bioinformatics* 8(1): 57.

Baker, C. & Cheung, K. (2007). *Semantic Web: Revolutionizing knowledge discovery in the life sciences*, Springer Verlag.

Barrell, D., Dimmer, E., Huntley, R., Binns, D., O'Donovan, C. & Apweiler, R. (2009). The goa database in 2009–an integrated gene ontology annotation resource, *Nucleic Acids Res.* 37(Database issue): D396–D403.

Beisswanger, E., Schulz, S., Stenzhorn, H. & Hahn, U. (2008). Biotop: An upper domain ontology for the life sciences. a description of its current structure, contents and interfaces to obo ontologies, *Applied Ontology* 3(4): 205–212.

Beneventano, D., Bergamaschi, S., Guerra, F. & Vincini, M. (2003). Synthesizing an integrated ontology, *IEEE Internet Comput* 7(5): 42–51.

Benjamins, V., Contreras, J., Corcho, Ó. & Gómez-Pérez, A. (2002). Six challenges for the semantic web, *KR2002 Semantic Web Workshop*.

Berners-Lee, T. (1997). Metadata architecture, URL: `http://www.w3.org/DesignIssues/Metadata`. 18.03.2011.

Berners-Lee, T., Fielding, R. & Masinter, L. (2005). Uniform resource identifier (uri): Generic syntax, URL: `http://tools.ietf.org/rfc/rfc3986.txt`. 18.03.2011.

Berners-Lee, T., Hendler, J. & Lassila, O. (2001). The semantic web, *Sci. Am.* 284(5): 28–37.

Bodenreider, O. (2004). The unified medical language system (umls): integrating biomedical terminology, *Nucleic Acids Res.* 32(Database Issue): D267.

Bodenreider, O. (2008). Ontologies and data integration in biomedicine: Success stories and challenging issues, *in* A. Bairoch, S. Cohen-Boulakia & C. Froidevaux (eds), *Data Integration in the Life Sciences*, Vol. 5109 of *Lecture Notes in Computer Science*, Springer Berlin / Heidelberg, pp. 1–4.

Börner, K. (2006). Semantic association networks: Using semantic web technology to improve scholarly knowledge and expertise management, *in* V. Geroimenko & C. Chen (eds), *Visualizing the Semantic Web*, Springer London, pp. 183–198.

Boury-Brisset, A. (2003). Ontology-based approach for information fusion, *Proceedings of the Sixth International Conference on Information Fusion*, pp. 522–529.

Brickley, D. & Miller, L. (2010). Foaf vocabulary specification 0.98, URL: `http://xmlns.com/foaf/spec/`. 18.03.2011.

Calì, A., Calvanese, D., De Giacomo, G. & Lenzerini, M. (2001). Accessing data integration systems through conceptual schemas, *Conceptual Modeling - ER 2001*, Vol. 2224 of *Lecture Notes in Computer Science*, Springer Berlin / Heidelberg, pp. 270–284.

Calì, A., Calvanese, D., De Giacomo, G. & Lenzerini, M. (2003). On the expressive power of data integration systems, *in* S. Spaccapietra, S. March & Y. Kambayashi (eds), *Conceptual Modeling - ER 2002*, Vol. 2503 of *Lecture Notes in Computer Science*, Springer Berlin / Heidelberg, pp. 338–350.

Camon, E., Magrane, M., Barrell, D., Lee, V., Dimmer, E., Maslen, J., Binns, D., Harte, N., Lopez, R. & Apweiler, R. (2004). The gene ontology annotation (goa) database:

sharing knowledge in uniprot with gene ontology, *Nucleic Acids Res.* 32(Database Issue): D262–D266.

Chabalier, J., Dameron, O. & Burgun, A. (2007). Integrating and querying disease and pathway ontologies: building an owl model and using rdfs queries, *ISMB conference*, Citeseer.

Cheung, K., Smith, A., Yip, K., Baker, C. & Gerstein, M. (2007). Semantic web approach to database integration in the life sciences, *in* C. J. O. Baker & K.-H. Cheung (eds), *Semantic Web*, Springer US, pp. 11–30.

Choi, N., Song, I. & Han, H. (2006). A survey on ontology mapping, *SIGMOD Rec.* 35(3): 34–41.

Cimiano, P. & Volker, J. (2005). A framework for ontology learning and data-driven change discovery, *Proceedings of the 10th International Conference on Applications of Natural Language to Information Systems (NLDB)*, Springer, pp. 227–238.

Davidson, S., Overton, C. & Buneman, P. (1995). Challenges in integrating biological data sources, *J. Comput. Biol.* 2(4): 557–572.

Delfs, R., Doms, A., Kozlenkov, A. & Schroeder, M. (2004). Gopubmed: ontology-based literature search applied to geneontology and pubmed, *Proceedings of German Bioinformatics Conference. LNBI*, pp. 169–178.

Ding, Y., Fensel, D., Klein, M. & Omelayenko, B. (2002). The semantic web: yet another hip?, *Data Knowl Eng* 41(2-3): 205–227.

Doms, A. & Schroeder, M. (2005). Gopubmed: exploring pubmed with the gene ontology, *Nucleic Acids Res.* 33(Web Server Issue): W783–W786.

Donini, F., Lenzerini, M., Nardi, D. & Schaerf, A. (1996). *Reasoning in description logics*, Center for the Study of Language and Information, Stanford, CA, USA.

EBI (n.d.a). Embl nucleotide sequence database, URL: `http://www.ebi.ac.uk/embl`. 18.03.2011.

EBI (n.d.b). Uniprot, URL: `http://www.ebi.ac.uk/uniprot`. 18.03.2011.

Ehrig, M. & Sure, Y. (2004). Ontology mapping - an integrated approach, *in* C. Bussler, J. Davies, D. Fensel & R. Studer (eds), *The Semantic Web: Research and Applications*, Vol. 3053 of *Lecture Notes in Computer Science*, Springer Berlin / Heidelberg, pp. 76–91.

Eilbeck, K., Lewis, S., Mungall, C., Yandell, M., Stein, L., Durbin, R. & Ashburner, M. (2005). The sequence ontology: a tool for the unification of genome annotations, *Genome Biol* 6(5): R44.

EMBL-EBI (n.d.). Array express archive, URL: `http://www.ebi.ac.uk/microarray-as/ae`. 18.03.2011.

Fensel, D. (2004). *Ontologies: A Silver Bullet for Knowledge Management and Electronic Commerce*, Springer.

Gagnon, M. (2007). Ontology-based integration of data sources, *10th international Conference on Information Fusion, Quebec, Canada*.

Gardner, S. (2005). Ontologies and semantic data integration, *Drug Discov Today* 10(14): 1001–1007.

Gene Ontology (n.d.). An introduction to the gene ontology, URL: `http://www.geneontology.org/GO.doc.shtml`. 18.03.2011.

Goble, C. & Stevens, R. (2008). State of the nation in data integration for bioinformatics, *Journal of biomedical informatics* 41(5): 687–693.

Grau, B., Horrocks, I., Motik, B., Parsia, B., Patel-Schneider, P. & Sattler, U. (2008). Owl 2: The next step for owl, *Web Semantics: Science, Services and Agents on the World Wide Web* 6(4): 309–322.

Gruber, T. (1993). A translation approach to portable ontology specifications, *Knowl Acquis* 5: 199–220.

Gruber, T. (1995). Toward principles for the design of ontologies used for knowledge sharing, *Int J Hum-Comput St* 43(5): 907–928.

Guarino, N. (1998). Formal ontology in information systems, *Formal ontology in information systems: proceedings of the first international conference (FOIS'98), June 6-8, Trento, Italy*, IOS Press.

Haase, P., Mathäß, T. & Ziller, M. (2010). An evaluation of approaches to federated query processing over linked data, *Proceedings of the 6th International Conference on Semantic Systems*, ACM, pp. 1–9.

Hernandez, T. & Kambhampati, S. (2004). Integration of biological sources: current systems and challenges ahead, *SIGMOD Rec.* 33(3): 51–60.

Hitzler, P., Krötzsch, M., Rudolph, S. & Sure, Y. (2008). *Semantic Web: Grundlagen*, Springer.

Hoehndorf, R., Oellrich, A., Dumontier, M., Kelso, J., Rebholz-Schuhmann, D. & Herre, H. (2010). Relations as patterns: bridging the gap between obo and owl, *BMC Bioinf* 11(1): 441.

IFOMIS, Saarland University (2010). Basic formal ontology (bfo), URL: `http://www.ifomis.org/bfo`. 18.03.2011.

Jacquemin, C. (2001). *Spotting and discovering terms through natural language processing*, MIT press Cambridge, MA.

Kanehisa-Laboratories (n.d.). Kegg: Kyoto encyclopedia of genes and genomes, URL: `http://www.genome.jp/kegg`). 18.03.2011.

Kashyap, V. & Borgida, A. (2003). Representing the umls semantic network using owl, *The SemanticWeb - ISWC 2003*, Vol. 2870 of *Lecture Notes in Computer Science*, Springer Berlin / Heidelberg, pp. 1–16.

Kei-Hoi, C., Robert, F., Scott, M., Matthias, S., Jun, Z. & Adrian, P. (2009). A journey to semantic web query federation in the life sciences, *BMC Bioinf* 10(Suppl 10): S10.

Kersey, P., Duarte, J., Williams, A., Karavidopoulou, Y., Birney, E. & Apweiler, R. (2004). The international protein index: an integrated database for proteomics experiments, *Proteomics* 4(7): 1985–1988.

Keshava Prasad, T., Goel, R., Kandasamy, K., Keerthikumar, S., Kumar, S., Mathivanan, S., Telikicherla, D., Raju, R., Shafreen, B., Venugopal, A. et al. (2008). Human protein reference database–2009 update, *Nucleic Acids Res.* 37(Database issue): D767–D772.

Krallinger, M., Valencia, A. & Hirschman, L. (2008). Linking genes to literature: text mining, information extraction, and retrieval applications for biology, *Genome Biol* 9(Suppl 2): S8.

Kugler, K., Tejada, M., Baumgartner, C., Tilg, B., Graber, A. & Pfeifer, B. (2008). Bridging data management and knowledge discovery in the life sciences, *Open Bioinformatics Journal* 2: 28–36.

Le Grand, B. & Soto, M. (2002). Visualisation of the semantic web: Topic maps visualisation, *Sixth International Conference on Information Visualisation, 2002. Proceedings*, pp. 344–349.

Masolo, C., Borgo, S., Gangemi, A., Guarino, N. & Oltramari, A. (2003). Wonderweb deliverable d18, URL: `http://www.loa-cnr.it/Papers/DOLCE2.1-FOL.pdf`. 18.03.2011.

Müller, H., Rangarajan, A., Teal, T. & Sternberg, P. (2008). Textpresso for neuroscience: searching the full text of thousands of neuroscience research papers, *Neuroinformatics* 6(3): 195–204.

Navigli, R. & Velardi, P. (2004). Learning domain ontologies from document warehouses and dedicated web sites, *Comput Linguist* 30(2): 151–179.

NCBI (2004). Genbank overview, URL: http://www.ncbi.nlm.nih.gov/genbank/GenbankOverview.html. 18.03.2011.

NCBI (n.d.). Gene expression omnibus, URL: http://www.ncbi.nlm.nih.gov/geo. 18.03.2011.

Neumann, E., Miller, E. & Wilbanks, J. (2004). What the semantic web could do for the life sciences, *Drug Discov Today* 2(6): 228–236.

OBO Foundry (n.d.). The open biological and biomedical ontologies, URL: http://www.obofoundry.org. 18.03.2011.

Ouksel, A. & Sheth, A. (1999). Semantic interoperability in global information systems, *SIGMOID Rec.* 28(1): 5–12.

Perez-Rey, D., Maojo, V., Garcia-Remesal, M., Alonso-Calvo, R., Billhardt, H., Martin-Sanchez, F. & Sousa, A. (2006). Ontofusion: Ontology-based integration of genomic and clinical databases, *Comput Biol Med* 36(7-8): 712–730.

Pfeifer, B., Aschaber, J., Baumgartner, C., Dreiseitl, S., Modre-Osprian, R., Schreier, G. & Tilg, B. (2007). A life science data warehouse system to enable systems biology in prostate cancer, *4th International Workshop, p 9ff DILS 2007, Pennsylvania, USA*.

Pollock, J. (2009). *Semantic Web for Dummies*, For Dummies.

Quilitz, B. & Leser, U. (2008). Querying distributed rdf data sources with sparql, *ESWC'08: Proceedings of the 5th European semantic web conference on The semantic web*, Springer-Verlag, Berlin, Heidelberg, pp. 524–538.

Richter, J. (2006). The obo flat file format specification, version 1.2, URL: http://www.geneontology.org/GO.format.obo-1_2.shtml. 18.03.2011.

Schulz, S., Beisswanger, E., Van Den Hoek, L., Bodenreider, O. & Van Mulligen, E. (2009). Alignment of the umls semantic network with biotop: methodology and assessment, *Bioinformatics* 25(12): i69–i76.

Shadbolt, N., Hall, W. & Berners-Lee, T. (2006). The semantic web revisited, *IEEE Intell Syst App* 21(3): 96–101.

Sirin, E., Parsia, B., Grau, B., Kalyanpur, A. & Katz, Y. (2007). Pellet: A practical owl-dl reasoner, *Web Semantics: science, services and agents on the World Wide Web* 5(2): 51–53.

Smith, A., Cheung, K., Yip, K., Schultz, M. & Gerstein, M. (2007). Linkhub: a semantic web system that facilitates cross-database queries and information retrieval in proteomics, *BMC Bioinf* 8(Suppl 3): S5.

Smith, B., Ashburner, M., Rosse, C., Bard, J., Bug, W., Ceusters, W., Goldberg, L., Eilbeck, K., Ireland, A., Mungall, C. & others (2007). The obo foundry coordinated evolution of ontologies to support biomedical data integration, *Nat Biotechnol* 25(11): 1251–1255.

Smith, B., Ceusters, W., Klagges, B., Köhler, J., Kumar, A., Lomax, J., Mungall, C., Neuhaus, F., Rector, A. & Rosse, C. (2005). Relations in biomedical ontologies, *Genome Biol* 5: R46.

Spasic, I., Ananiadou, S., McNaught, J. & Kumar, A. (2005). Text mining and ontologies in biomedicine: making sense of raw text, *Brief Bioinform* 6(3): 239–251.

Stanford University (n.d.). Stanford microarray database, URL: http://smd.stanford.edu. 18.03.2011.

Stein, L. (2003). Integrating biological databases, *Nat Rev Genet* 4(5): 337–345.

Stenzhorn, H., Schulz, S., Beißwanger, E., Hahn, U., Van Den Hoek, L. & Van Mulligen, E. (2008). Biotop and chemtop–top-domain ontologies for biology and chemistry, *7th International Semantic Web Conference (ISWC)*, Vol. 401, Citeseer.

Studer, R., Benjamins, V. & Fensel, D. (1998). Knowledge engineering: Principles and methods, *Data Knowl Eng* 25: 161–197.

Tang, J., Li, J., Liang, B., Huang, X., Li, Y. & Wang, K. (2006). Using bayesian decision for ontology mapping, *Web Semantics: Science, Services and Agents on the World Wide Web* 4(4): 243–262.

Tsarkov, D. & Horrocks, I. (2006). Fact++ description logic reasoner: System description, *in* U. Furbach & N. Shankar (eds), *Automated Reasoning*, Vol. 4130 of *Lecture Notes in Computer Science*, Springer Berlin / Heidelberg, pp. 292–297.

U.S. National Library of Medicine (2010). About the umls, URL: http://www.nlm.nih.gov/research/umls/about_umls.html. 18.03.2011.

Van Harmelen, F., Lifschitz, V. & Porter, B. (2008). *Handbook of knowledge representation*, Elsevier Science Ltd.

Velardi, P., Navigli, R., Cucchiarelli, A., Neri, F., Buitelaar, P., Cimiano, P. & Magnini, B. (2005). *Evaluation of OntoLearn, a methodology for automatic learning of domain ontologies*, IOS Press.

Volz, R., Oberle, D. & Studer, R. (2003). Implementing views for light-weight web ontologies, *Database Engineering and Applications Symposium, International* 0: 160–169.

W3C (2001). Xml in 10 points, URL: http://www.w3.org/XML/1999/XML-in-10-points.html.en. 18.03.2011.

W3C (2004a). Owl web ontology language overview, URL: http://www.w3.org/TR/2004/REC-owl-features-20040210. 18.03.2011.

W3C (2004b). Owl web ontology language reference, URL: http://www.w3.org/TR/owl-ref. 18.03.2011.

W3C (2004c). Rdf primer, URL: http://www.w3.org/TR/2004/REC-rdf-primer-20040210. 18.03.2011.

W3C (2004d). Rdf test cases, URL: http://www.w3.org/TR/rdf-testcases. 18.03.2011.

W3C (2004e). Rdf vocabulary description language 1.0: Rdf schema, URL: http://www.w3.org/TR/rdf-schema. 18.03.2011.

W3C (2004f). Rdf/xml syntax specification (revised), URL: http://www.w3.org/TR/rdf-syntax-grammar. 18.03.2011.

W3C (2004g). Resource description framework (rdf): Concepts and abstract syntax, URL: http://www.w3.org/TR/2004/REC-rdf-concepts-20040210. 18.03.2011.

W3C (2005). Primer: Getting into rdf & semantic web using n3, URL: http://www.w3.org/2000/10/swap/Primer. 18.03.2011.

W3C (2008a). Sparql protocol for rdf, URL: http://www.w3.org/TR/rdf-sparql-protocol. 18.03.2011.

W3C (2008b). Sparql query results xml format, URL: http://www.w3.org/TR/rdf-sparql-XMLres. 18.03.2011.

W3C (2008c). Turtle - terse rdf triple language, URL: http://www.w3.org/TeamSubmission/turtle. 18.03.2011.

W3C (2009a). Owl 2 web ontology language document overview, URL: http://www.w3.org/TR/owl2-overview. 18.03.2011.

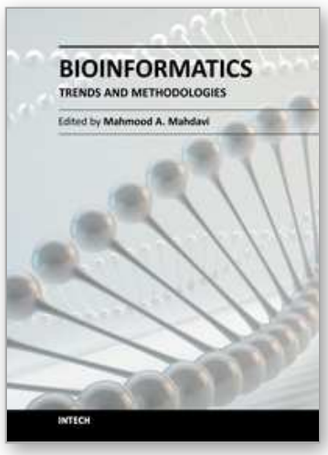
W3C (2009b). Rdb2rdf working group charter, URL: http://www.w3.org/2009/08/rdb2rdf-charter.html. 18.03.2011.

W3C (2009c). W3c rdb2rdf incubator group report, URL: http://www.w3.org/2005/Incubator/rdb2rdf/XGR-rdb2rdf-20090126. 18.03.2011.

W3C (2010a). Resource description framework (rdf), URL: http://www.w3.org/RDF. 18.03.2011.

W3C (2010b). W3c rdb2rdf incubator group, URL: http://www.w3.org/2005/Incubator/rdb2rdf). 18.03.2011.

W3C (2011). W3c semantic web activity, URL: http://www.w3.org/2001/sw. 18.03.2011.
Wache, H., Voegele, T., Visser, U., Stuckenschmidt, H., Schuster, G., Neumann, H. & Hübner, S. (2001). Ontology-based integration of information-a survey of existing approaches, *IJCAI-01 Workshop: Ontologies and Information Sharing*, Vol. 2001, Citeseer, pp. 108–117.
Zeng, K. & Bodenreider, O. (2007). Integrating the umls into an rdf-based biomedical knowledge repository, *AMIA Annu Symp Proc.*, p. 1170.



Bioinformatics - Trends and Methodologies
Edited by Dr. Mahmood A. Mahdavi

ISBN 978-953-307-282-1
Hard cover, 722 pages
Publisher InTech
Published online 02, November, 2011
Published in print edition November, 2011

Bioinformatics - Trends and Methodologies is a collection of different views on most recent topics and basic concepts in bioinformatics. This book suits young researchers who seek basic fundamentals of bioinformatic skills such as data mining, data integration, sequence analysis and gene expression analysis as well as scientists who are interested in current research in computational biology and bioinformatics including next generation sequencing, transcriptional analysis and drug design. Because of the rapid development of new technologies in molecular biology, new bioinformatic techniques emerge accordingly to keep the pace of in silico development of life science. This book focuses partly on such new techniques and their applications in biomedical science. These techniques maybe useful in identification of some diseases and cellular disorders and narrow down the number of experiments required for medical diagnostic.

INTECH
open science | open minds

InTech Europe
University Campus STeP Ri
Slavka Krautzeka 83/A
51000 Rijeka, Croatia
Phone: +385 (51) 770 447
Fax: +385 (51) 686 166
www.intechopen.com

InTech China
Unit 405, Office Block, Hotel Equatorial Shanghai
No.65, Yan An Road (West), Shanghai, 200040, China
中国上海市延安西路65号上海国际贵都大饭店办公楼405单元
Phone: +86-21-62489820
Fax: +86-21-62489821

IntechOpen