Semistructured Data Management
Part 3 - (Towards the) Semantic Web
Today's Question

1. What is the "Semantic Web"

2. Semantic Annotation using RDF

3. Ontology Languages
1. The Vision of W3C: Semantic Web

- The Semantic Web is an extension of the current Web in which information is given well-defined meaning, better enabling computers and people to work in cooperation. The mix of content on the Web has been shifting from exclusively human-oriented content to more and more data content. The Semantic Web brings to the Web the idea of having data defined and linked in a way that it can be used for more effective discovery, automation, integration, and reuse across various applications. For the Web to reach its full potential, it must evolve into a Semantic Web, providing a universally accessible platform that allows data to be shared and processed by automated tools as well as by people. The Semantic Web is an initiative of the World Wide Web Consortium (W3C), with the goal of extending the current Web to facilitate Web automation, universally accessible content, and the ‘Web of Trust’.

(http://www.w3.org/2001/sw/Activity)
Phase Zero: Textual Search (reminder)

- Web search engines (Google)
  - Searching for data on "anglerfish"
    - Results will be precise
  - This seems easy, but the same for "leech"
    - Organism leech
    - Authors: "Bleech", "Leechman", ...
    - Protein sequences: ...MNTSLEECHMPKD...
  - Search for "257" ...

Actually Web search engines have made the Web already semantically interoperable, based on what can be supported on pure text processing (essentially text-based search - we will introduce technical details on that later in the part on information retrieval). The problem is that this method is simply not precise enough, in particular for automated processing. It relies on the human understanding of natural language.
Phase 1: XML

- User-defined markup (schemas): provides possibility to share interpretation of data across various applications

- Different databases - Different schemas
  - SwissProt: Find `<Species> leech </Species>`
  - EMBL: Find `<Organism> leech </Organism>`

- Problem: Semantic Heterogeneity ∴ Phase 2: Semantic Web

  ![Diagram]

- Cost of semantic integration: 1 Trillion$ / year (Mike Brodie, GTE)

The next step taken towards a more meaningful Web was the introduction of additional structure in web documents in the form of user-defined markup. This is a first step towards semantic interoperability. The markup provides in a machine-processable form the necessary interpretation of the data or content. Applications that are able to correctly interpret the “semantic” markup are thus enabled to interpret data correctly. The problem is that different applications use different markups, and that the meaning of the markup is encoded into the terms used for element names, into associated documentation or simply kept in the heads of developers. Again, as soon as different markup schemas are used the relationship between the different interpretations have to be established in the heads of the users or developers, thus semantic interoperation does not occur in an automated manner.
What Do You Think?

• How to overcome semantic heterogeneity?
Three Ways to Overcome Semantic Heterogeneity

1. Standardization: agree on common user-defined markup (schemas)
   - great if no pre-existing applications
   - great if power player enforces it

2. Translation: create mappings among different schemas
   - requires human interpretation and reasoning
   - mappings can be difficult, expensive to establish

3. Annotation: create relationships to agreed upon conceptualizations
   - requires human interpretation and reasoning
   - annotation can be difficult, expensive to establish
   - reasoning over the conceptualization can provide added value

There exist three principal possibilities to tackle the issue of automating semantic processing of the data on the Web:

1. Standardization avoids the problem of semantic heterogeneity at the level of schemas. This approach is used where there exists already (historically) a wide agreement on the structure of relevant information and their interpretation, such as in business. Terminology in the financial world, for example, is pretty standardized, and therefore it is not a major problem to come up with agreed upon formal specifications of the terminology. This is even more the case as there exist in business typically strong players that can enforce the standards.

2. Translation or Mapping between different schemas is the second possibility to establish interoperability. This is the approach that has been extensively studied for integration problems in relatively small and controlled domains, such as business and large organizations. The requirements are typically changing not too quickly, thus much effort can be invested into developing the necessary mappings in order to properly map data from one representation into another, or to map data from multiple representations into one common global representation.

3. The third possibility is slightly different from the second: instead of engineering mappings between heterogeneous schemas for each integration problem, one first agrees on a common conceptualization of the world (or the relevant aspects of the world for a large class of applications). This conceptualization is normally called an ontology. Thus ontologies are fairly application independent and since they are formalized they are machine-processable. Once such an ontology is in place existing information sources can relate the structural elements they use for expressing certain concepts (e.g. element names) to concepts from the the ontology. This then (ideally) allows other applications to properly interpret the contents of the information systems. In addition ontologies in the general case should include reasoning capabilities, which would allow not only to use hard-coded, or pre-canned knowledge (e.g. in form of annotations), but also to derive new knowledge from combining existing knowledge in different ways.
This simple example illustrates the point of how ontologies might help to increase semantic interoperability. Take our earlier example of biological databases. These typically use different schemas to model related facts. For example, Database 1 uses the term Organism to denote an organism, and database 2 uses the term Species to do the same. Two annotators, who share the same ontology, now inspect the document and each of them associates the elements with terms taken from the ontology. So Annotator 1 will decide that the element Organism corresponds to information related to an organism, whereas annotator 2 recognizes actually from the content that the element is related to information about a fish and annotates correspondingly (by establishing a is-instance-of relationship). From that point on the fact that both annotators used the same ontology and that reasoning is possible in this ontology comes into play. Since in the ontology a Fish is a subconcept of Organism (a fact represented formally by a ISA relationship in the ontology) an automated processing tool (e.g. for searching for information) might exploit this relationship and correctly identify for a request for information on Organisms in both databases the related elements.
Ontologies

- Ontologies are an explicit specification of a conceptualization of the real world (Gruber, 1993)
- Ideally
  - different information systems agree on the same ontology
  - relate their model/schema/data elements to the ontology
  - mapping can be constructed via the ontology
- Issues
  - ontology languages (e.g. RDF), mappings and engineering
- Problem: requires agreement on the conceptualization of the real world!

We recall what we have said in the introduction on ontologies:
In order to obtain a better handle on the problem of dealing with different interpretations of data, a possibility is to make the interpretation a formally described entity. This requires the use of some “proxy” for the real world that is formally describable. These “proxies” are called ontologies.
Issues for Ontologies

- Agreement on meaning of terms
  - concepts: what does organism mean? what does fish mean?

- Modeling and encoding of ontologies
  - modeling primitives: what does an arrow mean?
  - what does "instance-of" mean? what does ISA mean?
  - different encodings of the model

What are the important aspects on ontologies?

1. They require an agreement on the meaning of terms (symbols). In practice, there exists no other way to establish such an agreement than by extensively collecting (from humans) this information and represent it within a formal specification. This however is done and there exist extensive ontologies, some of them have been built up over many years (see the example on the next slide)

2. In order to store this information a representation mechanism (a model) is needed, and the model needs to be encoded into data. The choice of the model is an important issue, since it should be very expressive (we want to model the world!) and easy to use (everyone should be able to use the ontology) at the same time. The encoding is equally important, as it should be done in a standardized form. If this where not the case we would immediately loose the advantage of having a common conceptualization of the world at the abstract level, since we were not able to exchange it properly with others. The examples illustrate of how easy it would be to create confusion at the level of encoding.
Researchers in artificial intelligence first developed ontologies to facilitate knowledge sharing and reuse. Since the beginning of the 1990s, ontologies have become a popular research topic, and several AI research communities—including knowledge engineering, natural language processing, and knowledge representation—have investigated them.

More recently, the notion of an ontology is becoming widespread in fields such as intelligent information integration, cooperative information systems, information retrieval, electronic commerce, and knowledge management.

Ontologies are becoming popular largely because of what they promise: a shared and common understanding that reaches across people and application systems.

One example of an ontology that has been developed as part of these efforts is WordNet. WordNet explains the meaning of English terms in a way as general and as precise as possible. The example gives the different meanings of the term “information” in English. WordNet can be freely accessed and downloaded over the Internet and has become very popular among researchers for that reason.
With respect to the modeling and encoding of ontologies to be used for the semantic Web, there exist a number of requirements, some of which follow from what we have discussed earlier:

1. **Simplicity**: the success of the Web was always founded on the principle of simplicity of concepts to encourage wide-spread use. Therefore complex models will not be successful. This is an important criterion, since some of the existing ontologies (one example is Cyc) are expressed in fairly complex knowledge representation models.

2. **Exchangeability**: Since the web is a communication environment, any kind of data that is processed must be easily exchangeable. This is what motivated the use of XML as a data representation format in the first place and should hold for metadata and ontology data as well.

3. **Non-intrusive annotation**: as the example on annotation we gave earlier demonstrated machine-processable knowledge required for the interpretation of data will be associated with the data typically a-posteriori. Also there exists no always a unique interpretation for the same data. Therefore any attempt to encode the knowledge required for interpretation directly into the data (as it would be the case if we use XML elements for annotation, to give an example) is not workable.

4. **Domain-specific vocabularies**: the model must provide a mechanism that allows to introduce vocabulary or terminology that is specific to a domain, in other words the possibility to specify schemas for different domains.

5. **Modeling primitives**: since any ontology model will be used in many different, and potentially very complex contexts (applications) they have to offer a sufficiently rich set of possibilities to model complex situations (e.g. complex structures or complex relationships). There exists a rich experience in modeling (e.g. from data modeling in databases, e.g. the entity-relationship model) and models for ontologies can draw from them.

6. **Reasoning Capabilities**: the example we discussed earlier already illustrate that also simple forms of reasoning within the ontology layer can make the interpretation of the data much more powerful (and thus the processing in the Semantic Web).

In the table we evaluate HTML, XML, RDF and OIL with respect to each of these aspects. RDF is the Resource Description Framework and is the first WWW standard proposed for the Semantic Web. OIL is the ontology interchange language, an extension of RDF proposed to enrich it with more reasoning capabilities and providing a well-defined semantics. We will introduce both models subsequently.

A one way to interpret the introduction of ontology models to support automated processing of semantics on the Web is the following: it can be seen analogously to the step from HTML to XML, were the issues of structuring data where separated from the issues related to the presentation of data. With the Semantic Web an attempt is made to separate the issues related to structuring of data from the issues related to interpreting the data, or in other words separating meaning from structure.

<table>
<thead>
<tr>
<th>Requirement</th>
<th>HTML</th>
<th>XML</th>
<th>RDF</th>
<th>OIL</th>
</tr>
</thead>
<tbody>
<tr>
<td>Simplicity</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>Exchangeability</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>Non-intrusive annotation</td>
<td>+</td>
<td>-</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>Domain-specific vocabularies</td>
<td>-</td>
<td>+</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>Modeling primitives</td>
<td>-</td>
<td>-</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>Reasoning capabilities</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>+</td>
</tr>
</tbody>
</table>

With the introduction of ontology models to support automated processing of semantics on the Web, it can be seen analogously to the step from HTML to XML, were the issues of structuring data where separated from the issues related to the presentation of data. With the Semantic Web an attempt is made to separate the issues related to structuring of data from the issues related to interpreting the data, or in other words separating meaning from structure.
This is the view the World Wide Web consortium develops on the Semantic Web:

Ontologies applied to the World Wide Web are creating the Semantic Web. When building the semantic web, several layers are required. At the lowest level a generic mechanism for expressing machine readable semantics of data is required. Originally, the Web grew mainly around HTML, which provides a standard for structuring documents that browsers can translate in a canonical way to render those documents. On the one hand, HTML’s simplicity helped spur the Web’s fast growth; on the other, its simplicity seriously hampered more advanced Web applications in many domains and for many tasks. This led to XML, which lets developers define arbitrary domain- and task-specific extensions (even HTML appears as an XML application— XHTML).

The Resource Description Framework (RDF) is the foundation for processing metadata providing a simple data model and a standardized syntax for metadata. Basically, it provides the language for writing down factual statements.

The next layer, the schema layer introduces means to define vocabulary, structure and constraints for expressing meta data. This is provided by the RDF Schema specification.

The fourth layer, then is the logical layer. We need ways of representing logic in documents to allow for, for example, rules that represent the deduction of one type of document from a document of another type; the checking of a document against a set of rules of self-consistency; and the resolution of a query by conversion from terms unknown into terms known.
Summary

• Why is the possibility to markup content (as in XML) not sufficient in order to establish semantic interoperability?

• What are ontologies and what is their purpose?

• What advantage has the "annotation" approach to semantic interoperability as compared to the "translation" approach?

• What is the difference between an ontology model and the encoding of an ontology?

• What are reasons not to change documents in order to make them semantically interoperable?
2. RDF Resource Description Framework

- **RDF (Instances)**
  - Statements about Resources (addressable by an URI) and literals (XML data)
  - Statements are of the form: subject property object
  - Like simple natural language sentences
  - RDF statements are themselves resources (reasoning about RDF)
  - Properties define relationships to other resources or atomic values

- **RDF-Schema**
  - Data model to specify schemas for RDF instances
  - Which properties are applicable for which objects with which subjects
  - Defines "grammar" and "vocabulary" for semantic domains (ontology language)

- **Relation RDF vs. RDFS comparable to well-formed vs. valid XML**

RDF is a standard supported by the W3C (http://www.w3.org/RDF/) to represent metadata. It is a fairly simple, graph-oriented data model to annotate any kind of XML document. RDF consists of two parts: a language for the metadata instances (RDF), which allows to connect simple « sentences » with document parts that are addressed by Universal Resource Identifiers (URI – we will call these document parts in the following resources), of which URL’s are the most important example. And a language for specifying schemas for RDF instances, that defines the possible vocabularies and the grammar of how these sentences may be formed. So the situation is very similar as with well-formed XML (instances) and XML-DTD (schemas).

The RDF model is in some sense similar to the entity-relationship (ER) model. Entities correspond to resources and relationships correspond to properties. The main difference is that RDF requires that relationships are directed, and carry the semantics that the resource from which the (directed) relationship emerges is assigned a property with the value to which the relationship points. This reflects the intention to use RDF to associated metadata (the value) with data (the source of the relationship).

Typical applications include PICS (annotating documents with information on the suitability of the content for certain groups, e.g. like the movie rating system) and Dublin Core (annotating documents with basic bibliographic information).

It is important to keep in mind that RDF is was only specified in a « semi-formal » manner, that means the XML encoding is well specified, however the underlying meaning of the syntax is given informally. Tools for processing (parsing, editing, storing, querying) RDF are developing at a fast pace.
RDF instances are RDF statements. We can view RDF instances in three different ways: we can view them like natural language sentences, where the subject is a URI and the object can be either URI’s or Strings, we can view them as directed graphs, where the subject and object are considered as nodes and the predicate is considered as link, or we can view them as XML documents where the RDF statement is encoded into an XML document. For each purpose (simply understanding the meaning, processing and algorithms for manipulating the structure, or representation for exchange and storage), one of the forms is most suitable.

In the graph representation an ellipsis is used to represent resources (identified by a URI) and rectangles to represent literals (atomic XML values).
For encoding RDF there exist many syntactic variations which make reading RDF documents sometimes rather confusing. Here we summarize the most important variants. The basic pattern of encoding is as follows: the subject is referenced in an element called rdf:Description. This element is the root of the document fragment representing the RDF statement. In the content of this element one finds one (or more) predicates, represented by elements, e.g. s:Creator. The content of this element in turn is the object of the statement. If the object is not a literal, one can alternatively represent the object as an attribute of the predicate element. Also, both the predicate and the object can be encoded into the element representing the statement, as it is shown for s:Date.
XML Namespaces

• An XML namespace is a collection of names (markup vocabulary)
  - identified by a URI reference
  - agreed element and attribute names

```xml
<rdf:RDF xml:lang="en"
  xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
  xmlns:rdfs="http://www.w3.org/2000/01/rdf-schema#">
```

• URIs are used just as unique identifiers, nothing else
  - in particular they do not refer to a DTD or schema

• Uses
  - universally agreed names
  - combination of names from different DTDs without name conflicts
  - but not combination of different DTDs

XML namespaces are a convention to distinguish different vocabularies for different applications in order to avoid names collisions. Consider just of how many XML applications would use the element name "name" and what happens if an application encounters this element name in documents originating from different sources without being able to distinguish the origin of the name. Therefore element and attribute names are prefixed by a label that is unique for a namespace. The problem is of course of how to obtain these unique labels. The solution is very simple: one uses Universal Resource Identifiers (typically URLs). But one must be careful here: the only purpose of using a URI (URL) is to have a unique prefix. There is nothing else associated with the URL, in particular the corresponding URL may not exist, it does not contain any information like a DTD or a Schema (though often useful information pertaining to the namespace is found under its specific URL).

Namespaces are important in order to provide universally agreed names with unambiguous semantics that can be exploited by applications. They also allow to combine element names from different DTDs which would otherwise be in conflict which each other. But the namespace concept is not related to the problem of combining definitions from different DTDs in any way.
Declaration

• Declaration of
  - Default namespace: xmlns (all names without prefix are supposed to be from there)
  - Identification by prefix: xmlns:ns (ns is the prefix)

• Can be declared in different positions
  - In internal DTD by using default attributes
  - In document by using attributes

```xml
<?xml version= "1.0"?>
<!-- element names without prefix belong to "books" -->
<book xmlns='urn:loc.gov:book'
     xmlns:isbn= 'urn:ISBN.0-395-36341-6'>
  <title>XML Handbook</title>
  <isbn:number>1591240349</isbn:number>
</book>
```

```xml
<!DOCTYPE doc [ 
  <!ELEMENT doc (x)> 
  <!ELEMENT x EMPTY> 
  <!ATTLIST x xmlns CDATA #FIXED "http://www.jclark.com/"> ]>
<doc><x/></doc>
```

Within the scope of one element one can always declare one default namespace, to which then all then names belong that do not have a prefix. In addition, an arbitrary number of other namespaces can be distinguished by a prefix. The namespace declaration can be given both in the DTD, then it is a default attribute, or within the document by using attributes. We see in the example all the cases illustrated:

In the example document an attribute xmlns is used to declare the default namespace, such that for example the element title is from the namespace “urn:loc.gov:bok” and a second namespace isbn is declared.

In the example DTD the element type x is equipped with the default namespace http://www.jclark.com/. Note that the keyword FIXED is used in the attribute declaration in order to express that it is an unchangeable attribute value.
RDF Complex Values

- Use an intermediate resource

\[
\text{http://www.doc.ch/} \quad \text{dc:Creator} \quad \text{Person://1234/1}
\]

\[
\text{name} \quad \text{email}
\]

\[
\text{Urs Giger} \quad \text{urs.giger@epfl.ch}
\]

Single values for the objects of an statement are just the simplest possible case. In general, if we have a complex object about which we want to make a statement involving multiple properties we create a new intermediate resource that combines the multiple properties and we can make then statements about this new complex object. In the XML encoding this can be represented by directly inlining the complex object into the content of the statement.
RDF Containers

- Containers
  - Bag (unordered)
  - Seq (ordered)
  - Alt (alternatives)
- Quantifiers
  - about: Urs is author of the talk (consisting of many slides)
  - aboutEach: Urs is author of each slide of the talk

Statements can be made not only on single resources, but as well on collections of resources. For that purpose RDF provides the container concept. Containers are special resources of one of three container types that are specified in the RDF standard. A container resource is then associated with a set of other resources. By creating a statement using the container object as "object", one can express statements made about the set of objects. More precisely, one can specify whether the statement is a statement of the set of objects "as a whole" or a statement that applies to each element of the set individually (what the consequences of this distinction are is not further specified in RDF).

There exist three different types of containers: bags which are unordered multi-sets (= sets with multiple occurrences of the same resources), sequences which are ordered sets (i.e. lists) of resources and alternatives which is a single resource that is to be chosen out of a given set. The property labels can be used to impose an order of elements of the set, by using labels _1, _2 etc. If the order is irrelevant one can use the alternative syntax rdf:li instead of rdf:_1, rdf:_2 etc.
Typing Resources

- Resources can be associated with a type by using `rdf:type` (a special property)

RDF allows to categorize documents into different classes. For doing that one associates with the resource that should be categorized another resource, that represents the category, with a special property `rdf:type` which is defined as part of the RDF specification. We will see later, when introducing RDF schema, what are possible consequences of typing. With respect to encoding, the type property can either be represented as any other property, or one can use a special abbreviated syntax, where the name of the type becomes the element name of the element embracing the statement.
Creating New Resources

- New RDF resources are created by using `rdf:ID` (a special property)

```
<rdf:RDF>
  <rdf:Description rdf:ID="12345">
    <rdf:type>Document</rdf:type>
    <dc:Title>Harry Potter</dc:Title>
  </rdf:Description>
</rdf:RDF>
```

```
<rdf:RDF>
  <dc:Document rdf:ID="12345">
    <dc:Title>Harry Potter</dc:Title>
  </dc:Document>
</rdf:RDF>
```

In this example the `rdf:Description` element has an `rdf:ID` attribute instead of an `rdf:about` attribute. Using `rdf:ID` indicates that we are describing a new resource, identified by the value of the `rdf:ID` attribute ("12345" in this case), rather than referring to an existing resource defined somewhere else. This resource can from there be referenced later by other RDF statements.
In RDF everything is a resource. In particular, RDF statements are resources and therefore should be annotated. From the viewpoint of the Semantic Web this is in fact extremely important. Since annotations do not express absolute truth but rather interpretations of data, and interpretations may be different, it is important to foresee the capability of annotating annotations (such as illustrated in the simple example). This allows to comment on annotations, to agree or dispute them etc.

Structurally, when inspecting the graph representation of RDF, it is not immediately clear of how to treat a statement as a resource, since a statement consists of three structural elements, the subject, the object, and the predicate. But we can apply the same "trick" as we did already for complex objects and collections. We introduce an abstract object which serves as representative for the statement, and make this object a representative by properly specifying it's properties. This process is called reification. As for the properties of the reified statement it is straightforward to connect it to it's subject and object as they are both resources. Also the type of the object is determined through a property rdf:type pointing to the special resource rdf:statement representing the category of statements. The predicate requires the introduction of a further object representing it. As we will see later this new resource representing the predicate is in fact an object that belongs to an RDF schema. For reification RDF introduces special properties rdf:object, rdf:subject and rdf:property, which are part of the RDF specification.

By reifying a statement one creates a new anonymous resource (this can be recognized from the fact that no identifier is found in the node representing the reified statement).
RDF Reification - Syntax

- The statement has an anonymous resource as subject, namely the reified statement which is fully characterized by its properties.

```xml
<rdf:RDF>
  <rdf:Description>
    <rdf:subject resource="http://www.doc.ch"/>
    <rdf:predicate resource="http://description.org/schema#Creator"/>
    <rdf:object>Urs Giger</rdf:object>
    <rdf:type resource="http://www.w3.org/TR/WD-rdf-syntax#Statement"/>
    <dc:Creator>Irma Müller</dc:Creator>
  </rdf:Description>
</rdf:RDF>
```

The XML encoding of reified statements follows the principles that have been introduced before. The annotation statement is represented as a complex, anonymous object.
We now introduce RDF Schema. RDF Schema provides two basic concepts.

1. the possibility to categorize RDF resources, into classes.
2. the possibility to constrain the usage of properties for classes, i.e. which classes can participate as subjects and objects in statements using a specific property.

First we introduce the classification mechanism. Classes are represented themselves as resources, which are of type rdfs:Class, a special class introduced in the RDFS specification. The type property rdf:type is used as before to indicate the type of class resources. If an application resource is of a specific application type then the resource is connected to the class resource via the rdf:type predicate. In the example two cases can be seen, the resource with ID "Karl Aberer" belongs to class (or is of type) Person and the resource "Married" is of type MaritalStatus. Between different classes a subclass relationship can be specified, by using the attribute rdfs:subClassOf. The intended semantics is that any resource belonging to the subclass also belongs to the superclass (containment relationship).

An interesting point is to see of how RDFS treats it's own modeling constructs within an RDF meta-schema that specifies the RDFS model in terms of the RDFS model itself. In RDF everything is a resource. Therefore in particular also resources that are classes are resources, and thus the rdfs:Class resource is a subclass of rdfs:Resource. Also application classes, such as "Animal", contain resources, and thus also these classes are subclass of rdfs:Resource. On the other hand, the type of a resource is indicated by connecting it by a rdf:type property to the resource that represents the type. Since rdfs:Resource is a class it is connected via the rdf:type property to the rdfs:Class resource. This produces the cyclic structure that we can observe within the RDF meta-class schema (in fact in the figure only a small fragment of the RDFS meta-class schema is displayed).
RDF Schema - Properties

- RDF properties
  - Connect resources
  - The RDF instance must have the properties that are declared for the class
  - rdfs:domain: classes of which the instances may have a property
  - rdfs:range: classes of which the instances may be the value of a property

The second important concept of RDFS is the possibility to define properties and their usage. RDF properties connect resources. As everything in RDF, RDF properties are resources themselves. Thus, they are represented in an RDF schema as resources. For example maritalStatus is a resource representing a property. In RDF schema it is now possible to constrain the usage of properties as follows: by connecting the property resource through the property rdfs:domain to a class resource, one specifies that the subject when using this property must originate from that class, i.e. be of the type of this class. Similarly for the object, the range can be constrained using rdfs:range. Ranges can also be atomic type, in that case one connects the property resource to (predefined) resources representing the data type of the atomic type, in the example above, this is STRING.

The RDFS model bears a lot of similarity with object-oriented models (or type specifications in the context of OO programming languages). However a fundamental difference is that properties (attributes in OO terminology) are defined independently of classes.
RDF Schema - Syntax

```xml
<rdf:RDF xml:lang="en"
  xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
  xmlns:rdfs="http://www.w3.org/2000/01/rdf-schema#">
  <rdfs:Class rdf:ID="Person">
    <rdfs:subClassOf rdf:resource="http://www.w3.org/classes#Animal"/>
  </rdfs:Class>
  <rdf:Property ID="maritalStatus">
    <rdfs:range rdf:resource="#MaritalStatus"/>
    <rdfs:domain rdf:resource="#Person"/>
  </rdf:Property>
  <rdf:Property ID="name">
    <rdfs:range rdf:resource="http://www.w3.org/classes#String"/>
    <rdfs:domain rdf:resource="#Person"/>
  </rdf:Property>
  <rdf:Property ID="age">
    <rdfs:range rdf:resource="http://www.w3.org/classes#Integer"/>
    <rdfs:domain rdf:resource="#Person"/>
  </rdf:Property>
  <rdfs:Class rdf:ID="MaritalStatus"/>
    <MaritalStatus rdf:ID="Married"/>
    <MaritalStatus rdf:ID="Divorced"/>
    <MaritalStatus rdf:ID="Single"/>
    <MaritalStatus rdf:ID="Widowed"/>
</rdf:RDF>
```

Since RDF schemas are expressed as RDF statements they can be encoded into XML along the same principles that we have introduced for RDF statements earlier. This example shows the complete encoding of the RDF schema we have used in our example before. Throughout the schema the abbreviated syntax for statements is used, replacing the element name "Description" by the corresponding class name of the subject of the description. Note of how using the ID attribute in the description elements, in the RDF schema new resources are introduced. They can be referred to from other statements using the newly introduced identifier and prepending #. The specification of the class "MaritalStatus" includes the creation of the complete extension of the class, enumerating all possible values. Strictly speaking, the statements creating the instances of class are not part of the schema level but part of the instance level of RDF.
RDF Schema Inheritance

- **rdfs:subClassOf**
  - A subclassOf B: Every instance of A is also instance of B
  - transitive, not reflexive, anti-symmetric (no cycles!)
  - M:N: a class can have arbitrarily many subclasses and superclasses
  - Subclass has all properties of the superclass

- **rdfs:subPropertyOf**
  - P1 subPropertyOf P2: If A has Property P1 with value B then it has also value B with Property P2
  - Example: Irma has Property Father with value Urs, and Father subProperty AParent implies Irma has Property AParent with value Urs

Similar to object-oriented modelling, RDF provides inheritance mechanisms in order to support the reuse of specifications. RDF Schema provides both for type inheritance and value inheritance.

1. **Type inheritance** is indicated by the rdfs:subClass property. If a class A is subclass of a class B this implies two things: first, every instances of A is also an instance of B, and second, every property that is specified for A can also be used for B (or is also specified for B). Since the subclass relationship has the subset semantics it is transitive, but must not contain cycles.

2. **Property inheritance** is indicated by connecting two properties by the property rdfs:subPropertyOf. If a property P1 is subproperty of another property P2, this has the following meaning. Every resource A that has property P1 (i.e. is the subject of a statement involving that property), and this property has value B (i.e. B is the object of the statement), then it has also property P2, an property P2 has the value B.
## Comparison XML - RDF

<table>
<thead>
<tr>
<th>Aspect</th>
<th>XML</th>
<th>RDF</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Document/Instance</strong></td>
<td>Nested String, Sequence</td>
<td>Graph, Sequence</td>
</tr>
<tr>
<td><strong>Main application</strong></td>
<td>Syntax Document content</td>
<td>&quot;Semantics&quot; Meta Data</td>
</tr>
<tr>
<td><strong>Schema</strong></td>
<td>Complex content models (Alternative, Sequence, Repetition)</td>
<td>(Complex) Objecttypes Resource = Object Property = Relationship</td>
</tr>
<tr>
<td><strong>Inheritance</strong></td>
<td>Not existent</td>
<td>SubClass SubProperty</td>
</tr>
<tr>
<td><strong>Main application of schemas</strong></td>
<td>Document Type Conformance</td>
<td>Vocabulary Specification</td>
</tr>
</tbody>
</table>

©2003/4, Karl Aberer, EPFL-SSC, Laboratoire de systèmes d'informations répartis
Summary

• What is the difference between derived and associative metadata?
• How is an RDF statement structured?
• Which are the three views of RDF statements?
• What is reification?
• How is an RDF statement represented in XML?
• What can be specified with RDF schema?
• Which inheritance mechanisms does RDF schema provide?
3. Towards Ontology Languages

• Limitations of RDF Schema
  - Limited expressive power (subclass, property, subproperty)
  - Unclear semantics (e.g. subproperty never has been precisely defined)
  - No reasoning support

• Requirements on ontology languages
  - Well designed
    • Intuitive to human users
    • Adequate expressive power
  - Well defined
    • Clearly specified syntax (obviously)
    • Formal semantics (equally important)
    • Adequate expressive power
  - Compatible with existing (web) standards
    • in particular RDF

RDF has been recognized to have some clear deficiencies with respect to it's use as ontology model. The set of basic modeling primitives is not very rich, even as compared e.g. to an ER model. Even worse, the semantics of RDF Schema is not well defined and it is therefore, for example, not clear when an RDF schema is consistent. For illustration consider the following example:

If P1 subPropertyOf P2 then we can conclude that in case that a class C1 has property P1, then C1 has also property P2. Assume now that the domain of P1 is C1, the domain of P2 is C2, and that in the schema C2 is declared to be a subclass of C1. From this it follows that only C2 can be domain of P2, which contradicts the fact that P1 is subproperty of P2. Since there exists no well-defined semantics of RDF, this kind of reasoning can not be performed formally, and thus the meaning of such schemas remains unclear. Thus an ontology language is desired satisfying the following three requirements:

• It must be highly intuitive for the human user. For example, the object-oriented modeling paradigm is highly successful because of it's intuitive nature, thus an ontology language should achieve the same. This must however not be at the cost of sacrificing the expressive power.
• It must have a well-defined formal semantics with established reasoning properties to ensure completeness, correctness, and efficiency in order to avoid problems as mentioned in the example before.
• It must have a proper linkage with existing Web languages such as XML and RDF to ensure interoperability.
OIL matches these criteria and unifies the three important aspects that different communities provide: rich modeling primitives as provided by the frame community (knowledge-based systems), formal semantics and efficient reasoning support as provided by description logics (a decidable fragment of first order logic), and a standard proposal for syntactical exchange notations as provided by the Web community.

- **Description Logics** describes knowledge in terms of concepts and role restrictions that are used to automatically derive classification taxonomies. They provide theories and systems for expressing structured knowledge, for accessing it and reasoning with it. OIL inherits from DL its formal semantics and the efficient reasoning support.

- **Frame-based systems** provide as central modeling primitive classes (i.e., frames) with attributes. These attributes do not have a global scope but are only applicable to the classes they are defined for. A frame provides a certain context for modeling one aspect of a domain. OIL incorporates the essential modeling primitives of frame-based systems.

- **Web standards: XML and RDF**. We must formulate a syntax of an ontology exchange language with existing Web standards for information representation. First, OIL has a well-defined syntax in XML based on a document type definition. Second, OIL is an extension of RDF and RDFS.
This is an example of OIL expressed in its surface language (similarly as the graph representation of RDF, only that OIL is richer in expressivity such that a graph-based representation is no longer feasible).

In OIL as in RDFS classes are defined to categorize resources and classes can be subclasses of other classes. The class definition for "plant" gives however already an example of the richer expressivity of OIL. One has the possibility to declare classes as subclasses of sets that are defined by complex so-called class expression, which are arbitrary combinations of other classes formed through the basic set operations of intersection, union and complement.

Properties are specified in OIL using the key-word "slot-constraint" (this is a heritage from frame-based systems where attributes are called slots). By using "has-value" one associates ranges, which again can be composed class expressions, as shown in the class "herbivore". Additional constraints can be formulated such as cardinality constraints, but also more complex constraints such as stating transitivity or symmetry of properties.

Without going further into details of OIL modeling this example should illustrate that OIL is in fact a richer model than RDFS.
OIL as RDF Extension

- The language to specify OIL models is given as an RDF Schema
  - similarly as the RDF schema language is given within an RDF schema
  - therefore OIL models can be expressed in RDF
  - some of the OIL modeling primitives and RDF modeling primitives overlap and are re-used in OIL

```xml
<rdfs:Class rdf:ID="herbivore">
  <rdf:type rdf:resource="#DefinedClass"/>
  <rdfs:subClassOf rdf:resource="#animal"/>
  <rdfs:subClassOf>
    <oil:NOT>
      <oil:hasOperand rdf:resource="#carnivore"/>
    </oil:NOT>
  </rdfs:subClassOf>
</rdfs:Class>
```

Defining an ontology language as an extension of RDFS means that every RDFS ontology is a valid ontology in OIL. As a consequence, for example, an OIL processor will also understand RDFS. However, the other direction is also possible: Defining an OIL extension as closely as possible to RDFS allows maximal reuse of existing RDFS-based applications and tools. However, because the ontology language usually contains new primitives (and therefore a new vocabulary, which an RDFS processor does not know), 100 percent compatibility is impossible.

Let’s look at an example. The following OIL expression defines herbivore as a class, which is a subclass of animal and disjoint to all carnivores:

An application limited to pure RDFS can still capture some aspects of this definition:

It captures that herbivore is a subclass of animal, however cannot understand that herbivore is a subclass of the complement of carnivore.
OIL Semantics

• The semantics of OIL is given in Description Logics (DL)
  - Description Logics is a fragment (a sublanguage) of first order predicate logic
  - reasoning (deriving/proofing statements) can be done (more) efficiently than in first order predicate logic (FOL)

• Example

<table>
<thead>
<tr>
<th>OIL expression</th>
<th>disjoint herbivore, carnivore</th>
</tr>
</thead>
<tbody>
<tr>
<td>DL expression</td>
<td>herbivore (\neg) carnivore</td>
</tr>
<tr>
<td>FOL expression</td>
<td>(\forall x(\text{herbivore}(x) \rightarrow \neg \text{carnivore}(x)))</td>
</tr>
</tbody>
</table>

• Using FOL reasoning (and thus DL reasoning) we could obtain, for example

\[
\text{if herbivore} \rightarrow \neg \text{carnivore, and lion} \rightarrow \neg \text{herbivore}
\]

A crucial aspect of OIL is its formal semantics. The formal semantics of OIL is given in terms of Description Logics. Without discussing description logics in detail, it is possible to understand the principle:

Description Logic is a fragment of first order logic (FOL) that has been developed with two things in mind: first, it should allow to reason about classes of objects and their relationships, and, second, reasoning in description logic should be decidable and efficient, i.e. we should be able to provide an algorithm, that decides for every statement that we make in DL whether it is true or false.

DL is exactly as OIL (and RDF) based on two basic constructs: classes and properties. The syntax of DL is an abbreviated syntax of FOL, which is for the non-expert awkward to read at the beginning.

We illustrate first by an example of how classes are represented in DL and then correspondingly also in FOL (which allows us to properly understand the interpretation of OIL statements).

In this example we look at the OIL expression that was used to state the condition that herbivores and carnivores are disjoint. The corresponding expression in DL is given below, and with some imagination one can interpret it properly. For a precise interpretation we give the corresponding expression in first order logic below.

Since we know which are the laws for reasoning in first order logic, we can see immediately that the reasoning given below in description logic is in fact correct.
### Property Semantics in OIL

- Classes in OIL (RDF) are unary predicates
- Slot constraints (= properties) in OIL are binary predicates

<table>
<thead>
<tr>
<th>OIL expression</th>
<th>class-def defined herbivore subclass-of animal slot-constrain eats value-type plant</th>
</tr>
</thead>
<tbody>
<tr>
<td>DL expression</td>
<td>herbivore = animal / ∀ eats .plant</td>
</tr>
<tr>
<td>FOL expression</td>
<td>∀x(herbivore (x) ↔ ∀y(eats(x, y) → plant(y)))</td>
</tr>
</tbody>
</table>

The second basic concept in DL besides classes are properties. We again give an example of how a property definition in OIL is translated into a description logic expression (no more so clear what it means!) and the corresponding FOL expression. One observes that properties are represented in FOL as binary predicates. This should be obvious, if we consider that properties where used to connect resources (and where represented as edges of a graph in RDF).
Summary

- Which are the reasons that RDF is not considered as being suitable as an ontology language?

- Which are the requirements an ontology language should satisfy?

- Which are examples of extensions in expressive power OIL introduces when compared to RDFS?

- Which approach is used to provide a precise semantics for OIL (and thus implicitly also for RDF)?
References

• WebSite
  - RDF: http://www.w3.org/
  - OIL: http://www.ontoweb.org/

• Articles