Comparing application profiles and ontologies for describing experiment data

João Silva
Faculdade de Engenharia da Universidade do Porto, Portugal
joaorosilva@gmail.com

Cristina Ribeiro and João Correia Lopes
DEI — Faculdade de Engenharia da Universidade do Porto / INESC Porto, Portugal
{mcr,jlopes}@fe.up.pt

Abstract. Digital data curation is currently becoming an essential part of knowledge management; this holds especially true for scientific data assets, since preserved data can be used for secondary research efforts. Regardless of making their data public or not, some U.Porto researchers show the need for tools which would allow them to deposit their scientific data assets in a secure environment and performing simple analysis, such as temporal series or data sub-setting.

It is in this context that an experiment is being developed at U.Porto, which aims to compare different data models. We compare application profiles against ontologies for the purpose of representing and describing a dataset created from a series of water and sediment pollution control experiments at U.Porto.

We argue that selecting the most appropriate data model and corresponding data exchange format is the first step in offering researchers a system which can provide them with a more consolidated view of their otherwise disperse datasets. As such, the scope of this study is to devise a machine-processable format for the data and its representation information, which are currently only present in human-readable documents produced during the experiment. Such documents include not only description metadata such as geospatial coverage and scientific methodology – variable descriptions, instrument measurement tolerances – but also structural metadata such as the ordering of samples and experiments.

1 Introduction

The ease of access to more powerful means of experimental analysis has made it possible (and sometimes even necessary) to produce more data during the course of scientific research efforts. As a consequence, researchers are producing increasingly large amounts of data, a challenge which falls under the e-Science[1]

[1] E-Science is computationally intensive science that is carried out in highly distributed network environments, or science that uses immense data sets that require grid computing[1]

domain. However, as the effort put into the production of data increases, so does the need for the adoption of adequate curation practises for this data.

Recent studies[2] have shown the need for the adoption of proper data curation practises and also the need for the implementation of data curation plans. In the USA, for example, NSF[2] grant applicants are required to annex data management plans to their research grant proposals[3].

This document starts with a description of some challenges in the field of scientific data curation. The production context of the data used as basis for this work is explained, followed by two alternatives developed for its representation. Finally, a comparison of these solutions is presented in the conclusions section.

2 The challenges for Data Curation

Data Curation poses several challenges regarding political, social and ethical concerns[4]. This study focuses on one of the most complex technical issues for data curation solutions: the need to preserve datasets in maintainable, exchangeable formats. Maintainability can be improved through the creation of a public specification of the used exchange format, which must not be dependent on any specific technology. Technologies such as XML[3], XSD[4], RDF[5] and OWL[6] are ideal candidates for this purpose because they can not only cope with these needs but also be used to build semantic standards for interoperability which can be reused, either in part or as a whole.

2.1 The pollutant analysis workflow

The Department of Chemistry of the Faculty of Engineering of the University of Porto performs routine analysis and experiments regarding the concentration of certain pollutants in water and sediments, which must be kept under strict limits specified by Portuguese law.

During these analysis runs, samples are taken and analysed using the appropriate apparatuses and experimental methods. This data is then saved in Excel spreadsheets, where it is statistically processed, and the final results are written in Word documents. This workflow is fairly common in many research efforts, and poses obvious preservation concerns. The data is dispersed among several sheets and reports. Careful organisation of these data files must be carried out, often by the researchers themselves. Seeing the problems that arise from the dispersion of their research materials, researchers have expressed the need for solutions which help them centralise and search their data.

[2] National Science Foundation

[3] eXtensible Markup Language

[4] XML Schema

[5] Resource Description Framework

[6] Web Ontology Language

2.2 Alternatives for the representation of research data

In an attempt to mitigate the issues faced by the researchers performing this analysis, we have used two alternatives for the representation of the data produced by these experiments: an application profile, represented as an XML Schema[5] document and an ontology, represented in the OWL[9] format. This article documents this process and compares the two alternatives.

3 Representing Experiment Data using an Application Profile

The first option for the representation of the gathered data is the development of an XML Schema. XML Schemas can be used to enforce a specific structure for the representation of data. This is important for correct data exchange and essential for the creation of a robust system capable of querying it.

In the development of the schema for this language — named Water Pollution Analysis and Experimental Data Modelling Language (WPAEDML) — some elements from existing schemas were reused, namely: the Dublin Core base elements, Qualified Dublin Core and the CML[7] schema.

Dublin Core[6] is a widely used format for the representation of Descriptive Metadata — a flat metadata schema for describing a resource. It is used to identify general metadata such as the authors, contributors, or the creation date of a specific resource. A set of 15 elements comprises the Core of the specification.

In addition to the core elements, another Dublin Core specification — DCMI Metadata Terms[7] — was used in the proposed schema. This second specification includes qualifiers which can be used, for example, to specify resources related to the one being described, or technical aspects such as the resource's file format.

To provide a domain-specific element for the analysed chemical substances, the `molecule` qualifier from the CML specification was used. The CML specification is much more complex and offers sophisticated element such as the description of the atoms and bonds that make up the structure of a molecule[14]. In this case, however, such level of detail is not required, since only the molecule name is required to uniquely identify a given substance. The advantage of using part of an existing specification whenever possible is that by sharing the same qualifiers, better interoperability between data models can be attained. The final result is shown in Figure 1.

3.1 The developed schema

In this schema, a `run` is a set of `experiments`, which can be seen as a table with its associated metadata. The `run` element is at the root of this hierarchy along with its own descriptive metadata. This metadata is represented by a

[7] Chemical Markup Language[13]

convenience group specified in the DC Core schema, which provides a link to its set of 15 base elements.

Each `experiment` has a series of `results`, which can be seen as the rows of a table; the header names for this conceptual table are the attributes of the `result` element. The `analysed_molecule` attribute's datatype, `moleculeIDType`, is specified in the CML schema to identify molecules by their name. The unit attribute also comes from the CML schema and is used to specify the unit of measurement for the result.

The `coverage` element is taken from the *Dublin Core* specification and means to specify the geospatial position where the data for this experiment was gathered.

The `related_things` node contains a list of related resources referenced by the experiment — for example, the Portuguese laws that specify the legal limits for any analysed substances — and any other resources which reference the annotated resource.

The `tested_molecules` element contains the list of substances analysed throughout the whole `run` and is taken from the CML Schema. The `methodology` element contains an identifier to another resource containing the information about the scientific procedures and methods followed during the `run`.

Finally, the `formats` node lists all the file formats in which the experiment results are available.

3.2 Metadata Granularity

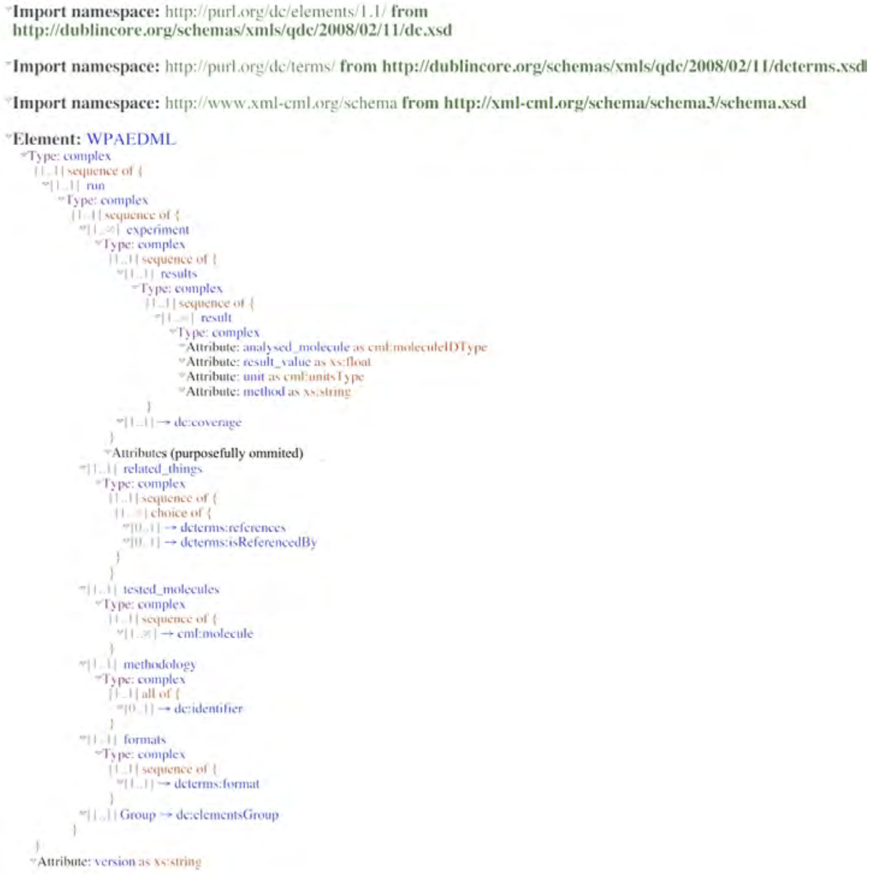
As the schema implies, there are different levels of metadata at different granularities inside the dataset. There is metadata at the `run` level, represented by the `formats`, `methodologies`, `tested_molecules` and `related_things` elements. At the `experiment` level there are several attributes which are simple datatypes and domain-dependant — listing them would not add to this study. Lastly, the columns names — symbolised by the attributes — which are specified at the `result` level can be considered metadata themselves.

4 Representation of Experiment Data through an Ontology

The specification of an XML schema can be used to specify a syntax for the exchange of this type of data. However, ontologies can add richer semantic content to the data representation. In this case, concepts like properties can help establish the semantics of all relationships between different parts of datasets — something which is not present in an XML Schema.

Following the principles of the Semantic Web[8], we have reused `Classes` and `Properties` present in three public ontologies, linking these concepts with those which are specific to the representation of this dataset.

The reused ontologies are the Dublin Core RDFS specification, the *Measurement Units Ontology* and the *ChemInf* Ontology, provided by *semantic chemistry*[10], an open project to support semantic chemistry.



```
Import namespace: http://purl.org/dc/elements/1.1/ from
http://dublincore.org/schemas/xmls/qdc/2008/02/11/dc.xsd

Import namespace: http://purl.org/dc/terms/ from http://dublincore.org/schemas/xmls/qdc/2008/02/11/dcterms.xsd

Import namespace: http://www.xml-cml.org/schema from http://xml-cml.org/schema/schema3/schema.xsd

Element: WPAEDML
  Type: complex
    [1..1] sequence of {
      [1..1] run
        Type: complex
          [1..1] sequence of {
            [1..∞] experiment
              Type: complex
                [1..1] sequence of {
                  [1..1] results
                    Type: complex
                      [1..1] sequence of {
                        [1..∞] result
                          Type: complex
                            Attribute: analysed_molecule as cml:moleculeIDType
                            Attribute: result_value as xs:float
                            Attribute: unit as cml:unitsType
                            Attribute: method as xs:string
                      }
                  [1..1] → dc:coverage
                }
              Attributes (purposefully ommited)
            [1..1] related_things
              Type: complex
                [1..1] sequence of {
                 [1..∞] choice of {
                   [0..1] → dcterms:references
                   [0..1] → dcterms:isReferencedBy
                 }
                }
            [1..1] tested_molecules
              Type: complex
                [1..1] sequence of {
                  [1..∞] → cml:molecule
                }
            [1..1] methodology
              Type: complex
                [1..1] all of {
                  [0..1] → dc:identifier
                }
            [1..1] formats
              Type: complex
                [1..1] sequence of {
                  [1..1] → dcterms:format
                }
            [1..1] Group → dc:elementsGroup
          }
    }
  Attribute: version as xs:string
```

Fig. 1. The structure of the developed XML schema

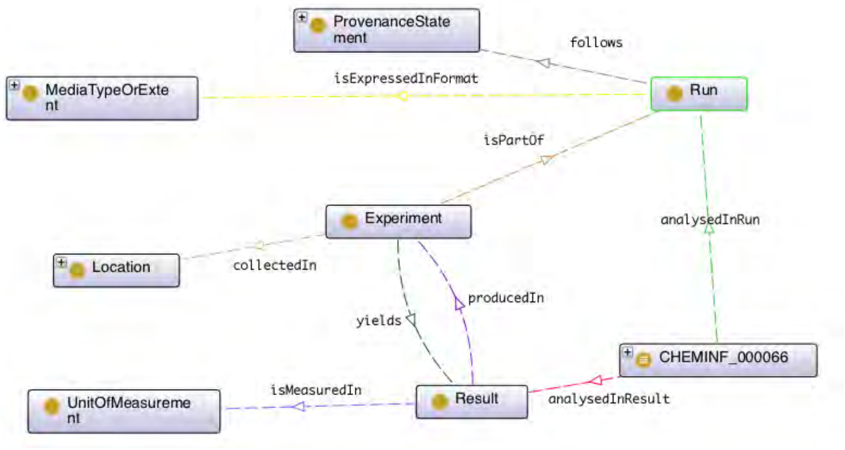


Fig. 2. Representation of the ontology developed for the studied dataset

Figure 2 offers a visual representation of the developed ontology.

In this ontology, the `Experiment`, `Run` and `Result` classes are specific to this study. All others are reused classes from existing ontologies.

A `Run` represents the group of `Expertiments`. It is materialised in a data file which has a format represented by the `MediaTypeOrExtension` class from the *Dublin Core Terms* ontology, encapsulating a MIME type. It also includes analysis on a set of chemical substances. These substances are represented by the `CHEMINF_000066` class, which is specified in the *Cheminf* ontology as the representation of "an information entity which is about a polyatomic molecular entity"[11]. `Runs` must state the method through which all their experiments are produced, providing a placeholder for researchers to annotate their research, including relevant parameters such as equipment measurement tolerances or analysis procedures. This method is represented through the `Provenance Statement` class from the Dublin Core ontology.

An *Experiment* is the representation of the analysis' results on a single sample of water or sediment. It is part of a *Run*, and yields a series of results. To perform the appropriate matching between the `experiment` and the place in which the samples were collected, the `Location` class from the *Dublin Core Terms* ontology.

Finally, a `Result` is always produced in the context of an `Experiment`. It represents the measured concentration of a substance, represented by the `CHEMINF_000066` class and expressed in an `UnitOfMeasurement`, taken from the Measurement Unit Ontology[12].

4.1 Similarities between the XML Schema and the Ontology

XML Schemas and ontologies cannot be seen as comparable entities since they reside in different levels of abstraction. However, there is an implicit correspondence between some Schema elements and some ontology classes. The most relevant example is the `result` - `experiment` — `run` element hierarchy present in the XML schema, which is similar to the `Result` — *producedIn* — `Experiment` and `Experiment` — *isPartOf* — `Run` properties in the ontology. This kind of relationship has been analysed in studies looking to extract semantic information from XML Schemas[17].

5 Conclusions and Future Work

An XML Schema such as the one presented in this study can help represent the analysed experimental data in a machine-processable format — something Excel spreadsheets are not suited for. This can be is a critical first step in the creation of a system capable of proper data curation.

The creation of an ontology for this data can open the data to the world, since Classes and Qualifiers can be taken from existing ontologies, allowing for the use of shared semantics in the representation of the experimental data.

The presented approaches are useful in the context of Data Curation since they can help solve two of its main issues: the need for shared, domain-specific data models to correctly represent data and the need for agreement on the semantics of data representation.

Research was performed during the course of this work to find existing ontologies and schemas which could be reused. From this research, we have concluded that schemas are more easily found than ontologies. The `semanticweb.org` website[15], for example, does not yet list the ontologies used for this work.

Former research[16] has shown that in many cases, it is possible to establish relationships between ontologies and XML Schemas. These matches are also present in this study, since there are similarities between parts of the XML Schema hierarchy and a set of Properties and Classes in the developed ontology.

We conclude that these two solutions complement each other in a data curation environment. An XML Schema is useful to specify a data exchange format for the raw data, which is a part of deposit policies for a data curation solution. It is also easier to specify and use than an ontology, but lacks its semantics.

Future work on this subject includes the analysis of potential applications of the two presented approaches in an hypothetical production environment — for example, a scientific data repository.

References

1. Wikimedia Wikipedia `http://en.wikipedia.org/wiki/E-Science` - Consulted on April 2011

2. Phillip Lord, Alison Macdonald et al.: From Data Deluge to Data Curation, 2004 http://www.ukoln.ac.uk/ukoln/staff/e.j.lyon/150.pdf - Consulted on April 2011
3. National Science Foundation: Grants.gov Application Guide, 2011 http://www.nsf.gov/pubs/policydocs/grantsgovguide0111.pdf - Consulted on April 2011
4. The Cornell University Library (CUL) Data Working Group (DaWG): Digital Research Data Curation: Overview of Issues, Current Activities, and Opportunities for the Cornell University Library, May 2008 http://ecommons.cornell.edu/bitstream/1813/10903/1/DaWG_WP_final.pdf - Consulted on April 2011
5. W3C: XML Schema http://www.w3.org/XML/Schema - Consulted on April 2011
6. Dublin Core Metadata Initiative: DCMI Specifications http://dublincore.org/specifications/ - Consulted on April 2011
7. Dublin Core Metadata Initiative: DCMI Metadata Terms http://dublincore.org/documents/dcmi-terms/ - Consulted on April 2011
8. John Hebeler, Matthew Fisher et al.: Semantic Web Programming, 2009 http://www.w3.org/TR/owl2-overview/ -
9. W3C: OWL 2 Web Ontology Language - Document Overview http://www.w3.org/TR/owl2-overview/ - Consulted on April 2011
10. Semantic Chemistry: The Cheminf Ontology - Homepage. http://code.google.com/p/semanticchemistry/ Consulted on April 2011
11. Semantic Chemistry: The Cheminf Ontology - OWL representation. http://code.google.com/p/semanticchemistry/source/browse/trunk/ontology/cheminf.owl?r=45 Consulted on April 2011
12. MORFEO Project: The Measuremet Units Ontology http://forge.morfeo-project.org/wiki_en/index.php/Units_of_measurement_ontology Consulted on April 2011
13. CML: Chemical Markup Language Home. http://xml-cml.org/index.php Consulted on April 2011
14. Peter Murray-Rust, Henry S. Rzepa et al.: CML - Chemical Markup Language, 1995 http://www.ch.ic.ac.uk/rzepa/cml/ Consulted on April 2011
15. semanticweb.org: Ontologies on semanticweb.org. http://semanticweb.org/wiki/Ontology Consulted on April 2011
16. Michael Klein, Dieter Fensel et al.: The relation between ontologies and XML schemas, 2001 http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.14.1037&rep=rep1&type=pdf Consulted on April 2011
17. Matthias Ferdinand, Christian Zirpins et al. Lifting XML Schema to OWL, 2004 http://vsis-www.informatik.uni-hamburg.de/getDoc.php/publications/204/fzt-lxs-04.pdf Consulted on April 2011