Managing Information Quality in e-Science with Semantic Web Technology*

Alun Preece†, Paolo Missier‡, Suzanne Embury‡, Binling Jin†, & Mark Greenwood‡
†University of Aberdeen, Computing Science, Aberdeen, UK
‡University of Manchester, School of Computer Science, Manchester, UK

Abstract

We outline a framework for managing information quality (IQ) in e-Science, using ontologies, semantic annotation of resources, and rules. Scientists define the quality characteristics that are of importance in their particular domain by extending an OWL DL IQ ontology, which classifies and organises these domain-specific quality characteristics within an overall quality management framework. RDF is used to annotate resources with IQ indicators. Rules are used to specify scientists' IQ preferences. As an illustration of our approach, we present an example Web service that computes IQ annotations for experiment datasets in biology.

1 Introduction

Information is viewed as a fundamental resource in the discovery of new scientific knowledge. Scientists expect to make use of information produced by other labs and projects in validating and interpreting their own results. A key element of e-Science is the development of a stable environment for the conduct of information-intensive forms of science. Problems arise due to variations in the quality of the information being shared [English, 1999]. Data sets that are incomplete, inconsistent, or inaccurate can still be useful when scientists are aware of these deficiencies.

The Qurator project is developing techniques for managing information quality (IQ) using Semantic Web technology. In contrast to previous IQ research which has tended to focus on the identification of generic, domain-independent quality characteristics (such as accuracy, currency and completeness) [Wang and Strong, 1996], we allow scientists to define the quality characteristics that are of importance in their particular domain. For example, one group of scientists may record "accuracy" in terms of some calculated experimental error, while others might define it as a function of the type of equipment that captured the data. Domain-specific IQ indicators are defined by extending a core IQ ontology defined in OWL DL, and the ontology classifies new indicators within the overall IQ framework. This allows scientists to *use* the definitions, by creating executable metrics based on them, and also to *reuse* definitions created by others, by browsing and querying an organised collection of definitions.

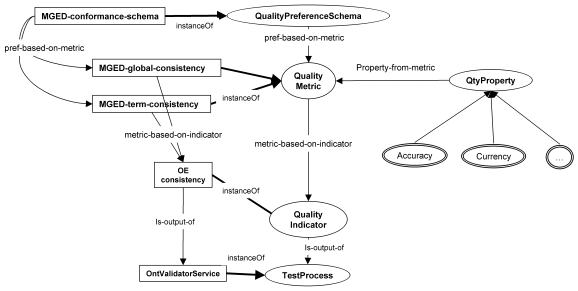

Figure 1: Fragment of the Qurator quality ontology

IQ indicators for specific resources are computed and associated with those resources as semantic annotations in RDF (linked to the IQ ontology). User-scientists can define rules that specify their IQ preferences, which are used to determine whether specific resources meet the users' IQ criteria. We are currently developing case studies of the use of this approach within two domains of post-genomic biology: proteomics and transcriptomics.

2 An IQ Ontology

At the core of the Qurator approach is an ontology for generic as well as domain-specific data quality concepts and terms, represented using OWL DL. A fragment is shown in Figure 1. Here, QualityMetric is a generic ontology concept whose semantic relationships to QualityIndicator, represented by the property metric-based-on-indicator, means that a metric is computed as a function of zero or more indicators. This root concept can be extended to include many different domain-specific, user-defined metrics, for instance MGED-global-consistency, described below.

In transcriptomics, microarray experiment data is routinely captured using the MAGE Object Model, and encoded using a standard XML syntax (MAGE-ML). The MGED Ontology provides common terminology for describing all aspects of the experiment design and of its execution. Let us suppose that, in searching for suitable microarray experiment data within a database, a biologist decides to adopt the consistency of use of MGED terms as one indicator for

*Funded by the EPSRC Programme *Fundamental Computer Science for e-Science*. See also: www.qurator.org

```
<BioSample
 identifier="S:Sample:MEXP:167278"
 name="CH131_1">
 <MaterialType_assn>
  <OntologyEntry
   category="MaterialType"
   value="whole_organism" />
 </MaterialType_assn>
 <Treatments_assnlist>
  <Treatment order="1"
  identifier="T:Sample:MEXP:167278">
   <Action_assn>
    <OntologyEntry
     category="Action"
     value="specified_biomaterial_action" />
   </Action_assn>
```

Figure 2: Fragment of a MAGE-ML data file

the overall quality of the experiment.

A *quality indicator* for a piece of data is an objectively measurable quantity whose value can be either computed from the data using an automated procedure, or be obtained interactively from the user. In our example, the MAGE standard prescribes which MAGE-OM entities, called OntologyEntry (OE), may refer to MGED Ontology entries. Figure 2 shows an XML fragment of a microarray experiment data file. The consistency status of each OE is computable and becomes an elementary quality indicator that annotates the corresponding XML element. From this fine-grain collection of indicators, useful *quality metrics* can then be computed by aggregation, such as the fraction of consistent values over the entire collection, or the consistency of use of particular MGED terms across the entire experiment (MGED-global-consistency and MGED-term-consistency in Figure 1).

This information is captured in the ontology as instances of existing concepts (the square elements in Figure 1). The test process model describes the process used to compute the quality indicators, in our case an "OntValidator service" that produces the OE consistency annotations. Data bindings map portions of the underlying data whose quality we are characterizing — that is, the experiment description, the OE elements — to the inputs of TestProcess instances. The ontology is aligned with the *my*Grid data ontology [Wroe *et al.*, 2003] (e.g. QualityMetric is a subclass of mygrid:data).

The framework allows for the definition of highly domain-specific IQ preferences, and supports the classification of these preferences under a generic IQ categorisation drawn from the earlier literature [English, 1999; Wang and Strong, 1996]. For example, the specific notion of MGED Ontology-conformance may classify as a special case of the generic notion of Accuracy. A biologist could use our ontology to browse for specialisations of Accuracy pertinent to their own domain, and reuse preferences defined by others.

The scientist may then use quality metrics to formulate *preference schemas* that indicate how a quality-based view of the data can be produced using the metrics. We are experimenting with both description-based and rule-based representations of preference schemas. Using a rule language (e.g. SWRL) one can define classes by giving necessary and sufficient conditions, like "an acceptable experiment is one in which for at least 75% of ontology entries the references are consistent, and the experiment was submitted within the past 2 years".

3 An Example IQ Annotation Service

As a concrete example of the Qurator approach to IQ management, we have implemented an ontology-conformance testing Web service. The service requires the URI of an XML document containing experiment data, and an XML control file specifying the elements to check in the experiment data (as XPath expressions). It returns a report detailing the conformance of each specified element. This conformance report can then be used to generate preference classifications and results for presentation. The conformance reports constitute quality annotations on the submitted datasets, and are represented in RDF which provides a natural way of making statements about Web resources, and integrates well with the OWL ontology.

The Web service is designed to handle a variety of different kinds of ontology entries. The MGED Ontology handler for the MAGE-ML experiment data uses both the OWL and (older) DAML versions of the ontology, and is able to check conformance of both ontology classes and individuals.

A key aim of Qurator is to embed the IQ-management tools within the scientists' working environment. To this end we are currently creating alternative clients for the ontology conformance Web service, including a general-purpose Web-based interface, and a client plugin for the Pedro data entry tool used by biologists (sourceforge.net/projects/pedro).

4 Conclusion

The Qurator project offers a framework for managing information quality in an e-Science context, allowing user-scientists to specify their IQ requirements against a formal ontology, so that the definitions are machine-manipulable. We have implemented an example Web service that computes RDF annotations for experiment datasets in transcriptomics: MGED Ontology conformance information. Following feedback from our collaborating users, we aim to further develop the IQ framework and associated toolset.

References

[English, 1999] L. English. *Improving Data Warehouse and Business Information Quality*. Wiley, 1999.

[Wang and Strong, 1996] R. Wang and D. Strong. Beyond accuracy: what data quality means to data consumers. *Journal of Management Information Systems*, 12(4):5–34, 1996.

[Wroe *et al.*, 2003] C. Wroe, R. Stevens, C. Goble, A. Roberts, and M. Greenwood. A suite of DAML+OIL ontologies to describe bioinformatics web services and data. *International Journal of Cooperative Information Systems*, 12(2):197–224, 2003.