Qurator: An Ontology-Based Approach to Handling Information Quality in e-Science

Paolo Missier, Suzanne Embury, Mark Greenwood (University of Manchester) Alun Preece, Binling Jin (University of Aberdeen)

Qurator approach to Information Quality (IQ)

Qurator seeks to assist scientists and data curators in managing the quality of their information. Rather than trying to impose a common set of generic IQ priorities on all users of a resource, an alternative is to provide scientists with the means of expressing explicit descriptions of quality that are relevant to their domain of interest and specific to their current task

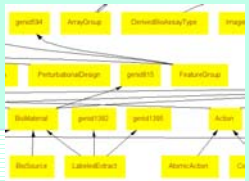
Working closely with user-scientists in two post-genomics domains - proteomics and transcriptomics - two hypotheses will be tested:

1. **that it is possible to elicit detailed specifications of the IQ requirements of user-scientists, preferably in a formal language so that the definitions are machine-manipulable**
2. **that the annotation of information resources with detailed descriptions of their quality can be performed in a cost-effective manner**

An example scenario in transcriptomics

In transcriptomics, microarray experiment data is routinely captured in MAGE-ML format. Elements of an experiment should be described in a standard way using terms from the **MGED Ontology***.

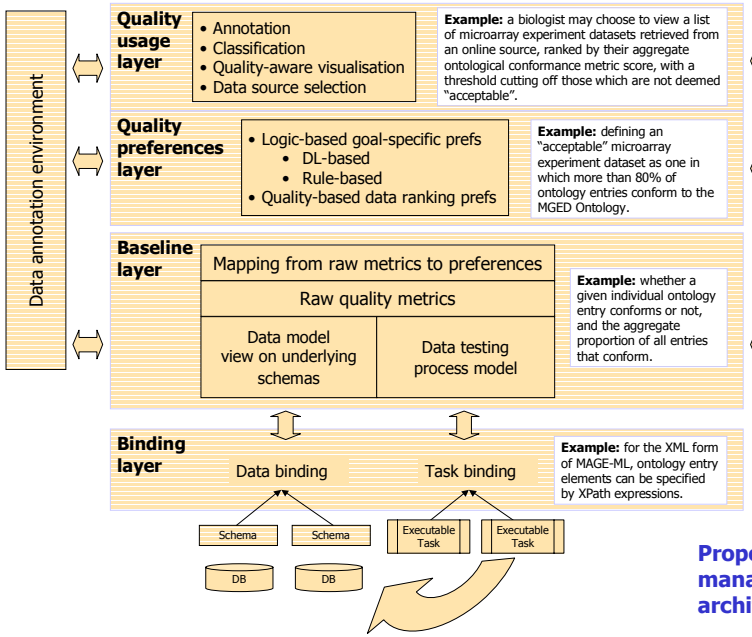
In searching for microarray experiment data to use for their own purposes, a particular biologist may specify a quality requirement on the extent to which particular elements of the dataset – called **ontology entries** – conform to the MGED Ontology.



```
<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>
```

*http://mged.sourceforge.net/ontologies/MGEDontology.php

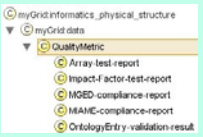
Visit **www.qurator.org** Contact **info@qurator.org**



Proposed IQ management architecture

An ontology-based approach to IQ

Ontologies play several roles in defining the Qurator IQ framework. The main concepts of the framework are defined in our **IQ Ontology** (represented using the Web Ontology Language, OWL) which is aligned with the *my*Grid project data ontology.



While the framework allows for the definition of highly domain-specific (and scientist-specific) IQ preferences, it allows for the classification of these preferences under a generic categorisation drawn from the IQ literature.

We are experimenting with the various possibilities afforded by the OWL-based representation for defining preferences in a machine-manipulable way.

Description logic-style IQ preference:

"an acceptable dataset is defined as one in which all ontology entries are MGED-conformant"

Rule-style IQ preference:

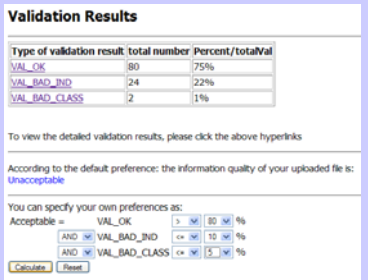
"a dataset is acceptable if more than 80% of ontology entries fully comply and the experiment was performed in the past 2 years"

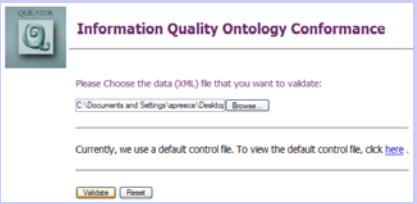
Example IQ service: testing ontology conformance

As a concrete example of the Qurator approach, we have implemented an ontology-conformance testing Web service. The service requires two pieces of information:

- the URI of an XML document containing **experiment data**
- an XML **control file** specifying the elements to check in the experiment data (as XPath expressions)

The service 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 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 the DAML version of the ontology, and is able to check conformance of both ontology **classes** and **individuals**. Other forms of ontology and controlled vocabulary in common use can also be checked with alternative handlers, including simple textual/lexical lists of terms, RDF Schemas, and the various forms of OWL.

Part of the aims of Qurator are 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 (shown here), and a client plugin for the Pedro data entry tool widely used by biologists.

The Qurator project is funded by the EPSRC Programme Fundamental Computer Science for e-Science: GR/S67593 & GR/S67609 - Describing the Quality of Curated e-Science Information Resources.

Qurator logo by Irene Christensen.