Qurator: Describing the Quality of Curated E-Science Information Resources

The importance of information quality (IQ)

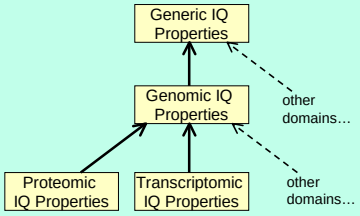
As more e-science information resources become available online, it is increasingly difficult to ignore the fact that much of the data they contain is of extremely variable quality. Very few tools exist by which providers, curators and consumers of e-science resources can discover and document their quality. The Qurator project aims to develop and test such tools in close collaboration with user-scientists, with the long term goal of providing generic information quality (IQ) support in e-science.

Qurator goals

The project 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 annotating their information with explicit descriptions of its quality in terms that are relevant to the domain of interest and the specific application in hand.

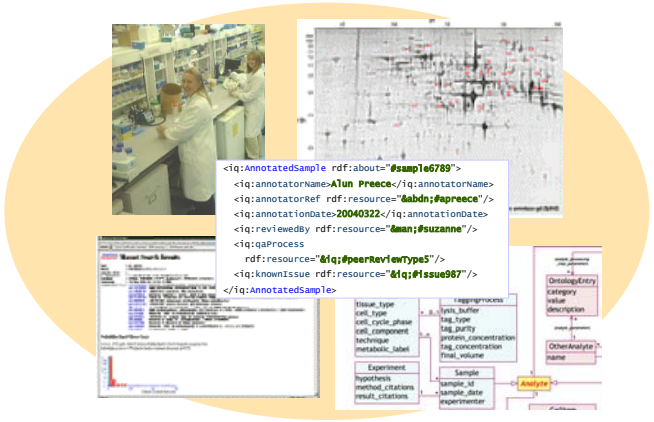
Working closely with user-scientists in two post-genomics domains - proteomics and transcriptomics - two hypotheses will be tested:

1. that it is possible to elicit detailed specifications of the IQ priorities of specific scientific domains;
2. that the annotation of sources (relative to the identified priorities) can be performed in a cost-effective manner.



The two domains were selected because scientists in these areas have identified genuine IQ concerns. They will be used to instantiate and evaluate a generic IQ framework, which should be applicable across a range of e-science areas.

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

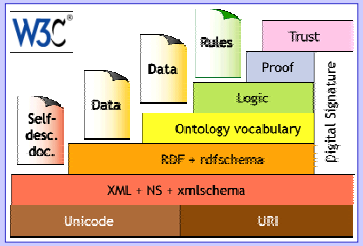


Qurator's Web-deployed tools will allow user-scientists to generate and interpret IQ representations as conveniently as possible. The tools will embed within the scientists' working environment, and be compatible with Grid infrastructure.

Qurator & the Semantic Grid

Qurator will explore a combination of representations to capture IQ semantics in a machine-manipulable form. The project requires a language that is open-ended (in that IQ attributes may be only partly defined), evolvable, and that integrates a variety of definition styles.

In developing the necessary IQ formalisms, Qurator will investigate the use of Semantic Web and emerging Semantic Grid work:



RDF(S) as a basic IQ annotation format

OWL for IQ concept taxonomies

RuleML/SWRL for IQ rules

DAML-S/OWL-S + myGrid service ontology for IQ processes

Example IQ issues in the target domains

- The quality of results from a microarray experiment can be partially described by stating the post-processing steps that have been applied. For example, several different forms of normalisation are now common and error models are being developed that can help to identify problems with the data before it is distributed.
- Functional annotations of gene sequences can be considered to be more reliable if the gene in question is homologous to many other genes than if it has no or few known homologues.
- If significant amounts of data have been added to the input sources since an *in silico* experiment experiment was last performed then the results may be less reliable than those from a more recently undertaken experiment, since they will not have been influenced by the newer data.
- The quality of proteomics data can be partially described by indicating the number of biological replicates that have been compared as well as by clarifying the post-processing quantification procedures that have been used.

Qurator investigators

Suzanne Embury
PI, Manchester

Alun Preece
PI, Aberdeen

Brian Warboys
Manchester

Andy Brass
Manchester

Collaborating groups

Molecular Evolution and Bioinformatics, CEH Oxford
led by Dr Dawn Field

Molecular and Cell Biology, University of Aberdeen
laboratory led by Prof Al Brown

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.