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

Suzanne M. Embury, Univ. of Manchester
Alun D. Preece, Univ. of Aberdeen

## E-Science: Implications for Data

Increase in amount of data generated
- High-throughput science (e.g. transcriptomics)

Internet + Web allow global collaboration and sharing
- Increased visibility of data
- Use of data by scientists not in original team

Increase in perceived/expected lifetime of data
- Funding bodies aim to maximise RoI
- Integration of data produced by multiple teams
- Extension/annotation of data by third parties
- Use of data for wholly unforeseen applications

## Biological IQ: Early Indications

There's already a lot of poor data out there

- Spelling errors: "Rabit" in SWISS-PROT
- Naming errors: Gene IDs not unique, change over time
- Inconsistencies: GenBank vs SWISS-PROT (Karp *et al.* Bioinformatics 17(6):2001)
- Out-of-date data: microarray annotations
- Data pollution: annotation errors (Gilks *et al.* Bioinformatics 18(12):2002)


THE UNIVERSITY of MANCHESTER UNIVERSITY OF ABERDEEN EPSRC Engineering and Physical Sciences Research Council



## Solutions?

Data Centres/Data Curators (e.g. NERC)

Community-driven data standards

- Data formats (e.g. MAGE-ML, PEDRo)
- Controlled vocabularies (e.g. Gene Ontology)
- Provenance and quality control (e.g. MIAME)

Good practice in experiment design and analysis

Is this enough?


THE UNIVERSITY of MANCHESTER UNIVERSITY OF ABERDEEN EPSRC Engineering and Physical Sciences Research Council

## "Fit for Purpose"

Definition of high IQ changes with:

- Domain/Application
- Passage of time

Therefore, need tools to help us live with poor quality data

Existing work in IQ:

- technology-led, not domain-led
- minimal impact on real applications


THE UNIVERSITY of MANCHESTER UNIVERSITY OF ABERDEEN EPSRC Engineering and Physical Sciences Research Council



## Qurator – approach to IQ

DON'T try to impose a common set of generic IQ priorities on all users of a resource

DO 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
- the specific application at hand


THE UNIVERSITY of MANCHESTER UNIVERSITY OF ABERDEEN EPSRC Engineering and Physical Sciences Research Council

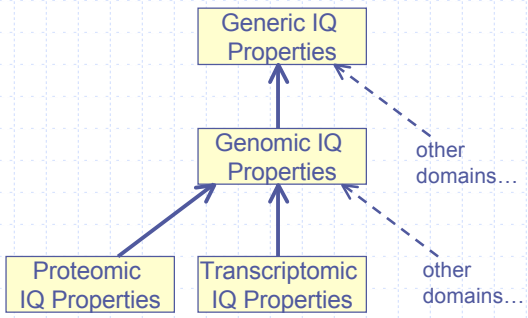
# Qurator – key hypotheses

1. It is possible to elicit detailed specifications of the IQ priorities of specific scientific domains

2. The annotation of sources (relative to the IQ priorities) can be performed in a cost-effective manner



# Towards an IQ framework




The two target domains will be used to instantiate and evaluate a generic IQ framework, which should be applicable across a range of e-science areas.

# Towards an IQ toolbox

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.



# Towards an IQ representation

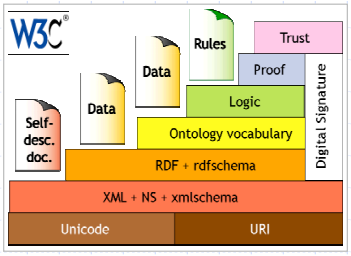
Qurator will explore a combination of representations to capture IQ semantics in a machine-manipulable form.

The project requires a language that is

- open-ended (in that IQ attributes may be only partly defined)
- evolvable
- that integrates a variety of definition styles

# Qurator & the Semantic Grid

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

THE UNIVERSITY of MANCHESTER UNIVERSITY OF ABERDEEN EPSRC



# Qurator team

**Suzanne Embury**
Manchester

**Brian Warboys**
Manchester

**Andy Brass**
Manchester

**Alun Preece**
Aberdeen

**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 UNIVERSITY of MANCHESTER UNIVERSITY OF ABERDEEN EPSRC

# Contacts

**www.qurator.org**

**info@qurator.org**

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.