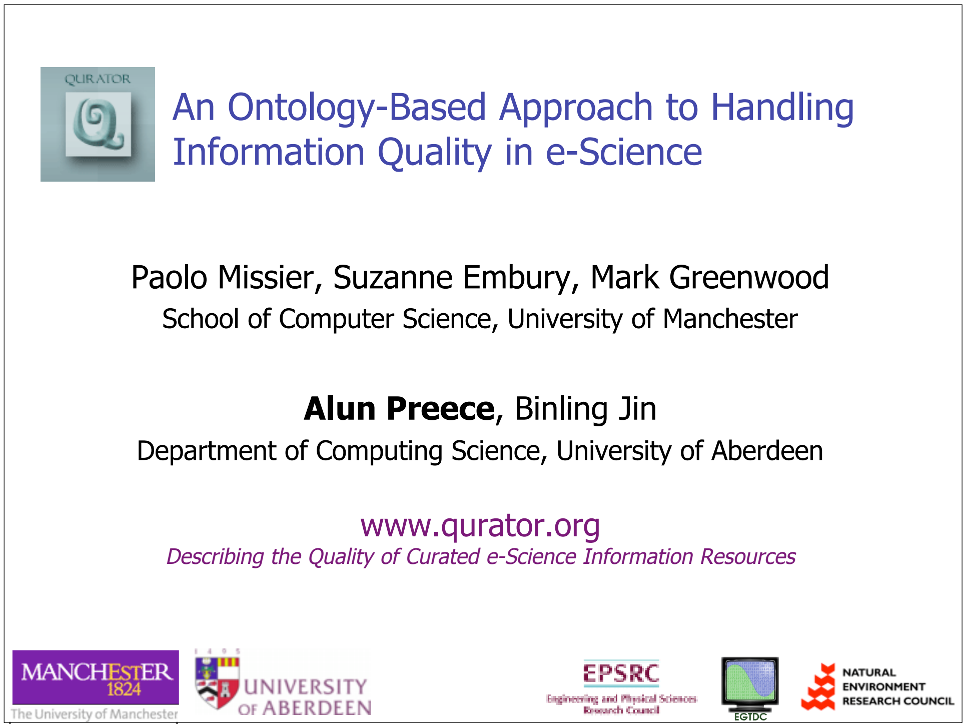


QURATOR

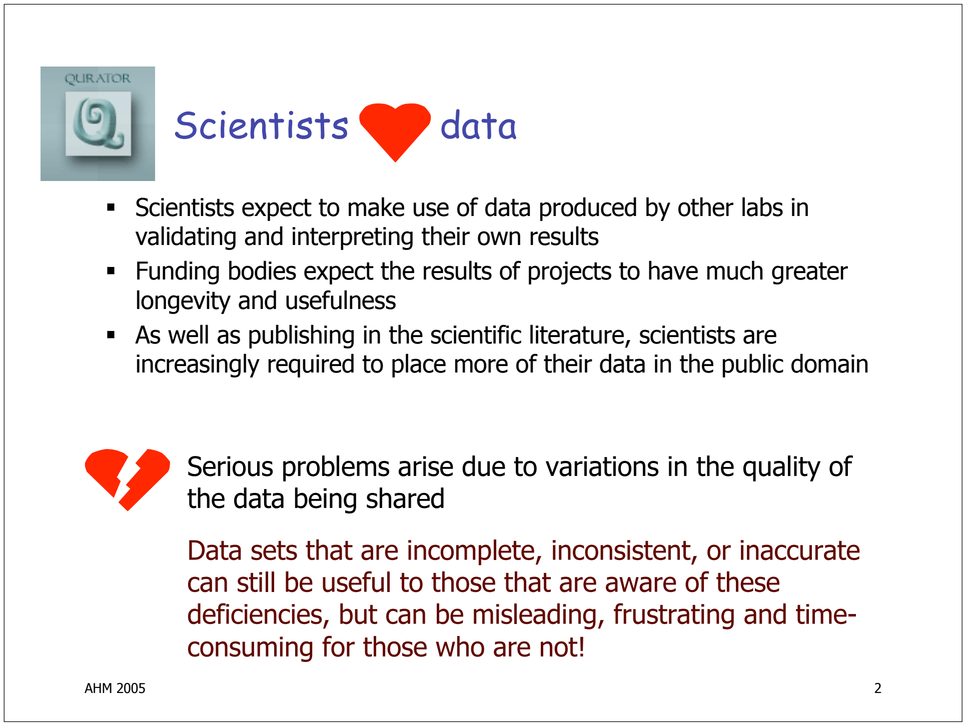
An Ontology-Based Approach to Handling Information Quality in e-Science

Paolo Missier, Suzanne Embury, Mark Greenwood
School of Computer Science, University of Manchester

Alun Preece, Binling Jin
Department of Computing Science, University of Aberdeen

www.qurator.org
Describing the Quality of Curated e-Science Information Resources

MANCHESTER 1824 The University of Manchester
UNIVERSITY OF ABERDEEN
EPSRC Engineering and Physical Sciences Research Council
EGTDC
NATURAL ENVIRONMENT RESEARCH COUNCIL



QURATOR

Scientists ❤ data

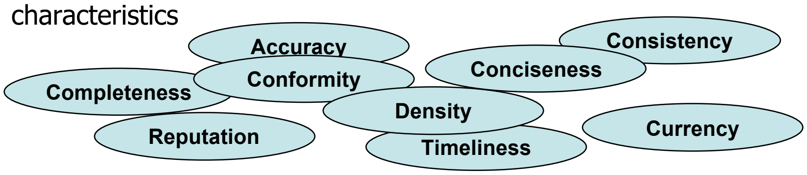
- Scientists expect to make use of data produced by other labs in validating and interpreting their own results
- Funding bodies expect the results of projects to have much greater longevity and usefulness
- As well as publishing in the scientific literature, scientists are increasingly required to place more of their data in the public domain

💔 Serious problems arise due to variations in the quality of the data being shared

Data sets that are incomplete, inconsistent, or inaccurate can still be useful to those that are aware of these deficiencies, but can be misleading, frustrating and time-consuming for those who are not!

Research in information quality (IQ)

Focus has traditionally been on the identification of generic quality characteristics



These "one-size-fits-all" quality characteristics are so broad in their meaning that they don't fit scientists' IQ requirements

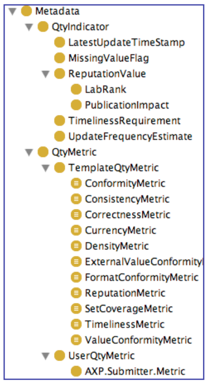
Alternative approach: identify the quality characteristics that are of importance in a particular domain. Example:

- one group of scientists may record "accuracy" in terms of some calculated experimental error,
- others might define it as a function of the type of equipment that captured the data…

Qurator manifesto I

It is possible to elicit detailed specifications of the IQ requirements of individual scientists or communities of scientists, preferably in a formal language so that the definitions are machine-manipulable

It must be possible for scientists to **use** the definitions, by creating executable metrics based on them, and also to **reuse** definitions created by others, e.g. by browsing and querying an organised collection of definitions

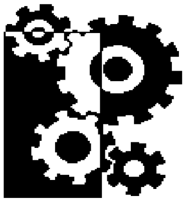


```
Metadata
  QtyIndicator
    LatestUpdateTimeStamp
    MissingValueFlag
    ReputationValue
      LabRank
      PublicationImpact
    TimelinessRequirement
    UpdateFrequencyEstimate
  QtyMetric
    TemplateQtyMetric
      ConformityMetric
      ConsistencyMetric
      CorrectnessMetric
      CurrencyMetric
      DensityMetric
      ExternalValueConformity
      FormatConformityMetric
      ReputationMetric
      SetCoverageMetric
      TimelinessMetric
      ValueConformityMetric
    UserQtyMetric
      AXP.Submitter.Metric
```

Qurator manifesto II

The annotation of information resources with detailed descriptions of their quality can be performed in a cost-effective manner

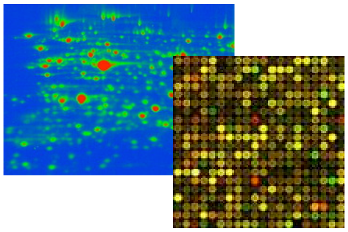
This means that the overhead of creating and managing the definition of a new IQ characteristic and its associated metrics should not be too high, and also that it should be possible to operationalise the computation of IQ measurements over sizeable datasets



Approach

Test the two statements by making a detailed study of IQ management in two "omic" biology domains:

- proteomics
- transcriptomics



Today we…

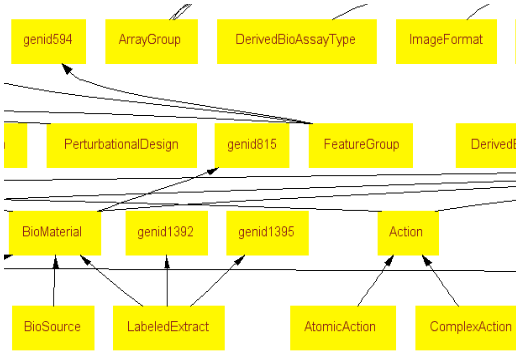
- present the initial version of our IQ framework for capturing scientists' IQ requirements
- show how a domain-specific IQ characteristic can be defined as part of our overall framework
- introduce a Web service that automates one kind of IQ annotation of datasets

… using a motivating example from transcriptomics

Transcriptomics example I

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*

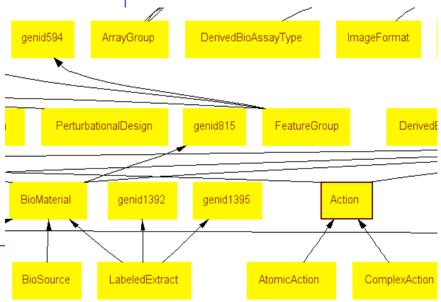
In searching for microarray experiment data to use for their own purposes, a particular biologist may specify a quality requirement on the extent to which particular elements of the dataset – called ontology entries – conform to the MGED Ontology



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

Transcriptomics example II

```
<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>
…
```



Core IQ concepts

A **Test Process** computes one or more **Quality Indicators** on some data

- e.g. **OntValidator** computes **OE Consistency** on MAGE-ML data

A **Quality Indicator** is an objectively-measurable value either computable from data or obtainable from a user

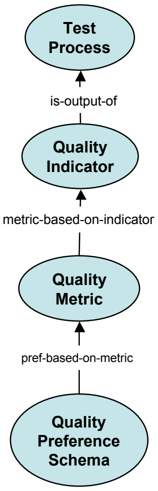
- e.g. **OE Consistency** indicates if an OE conforms to its ontology

A **Quality Metric** is derived from one or more **Quality Indicators**

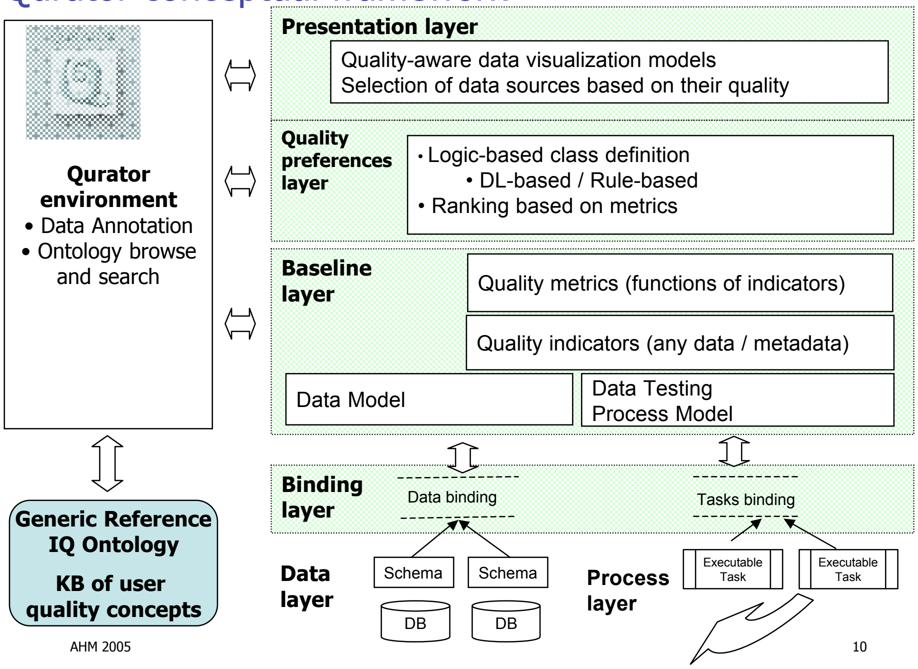
- e.g. **MGED-term-consistency** is the fraction of conforming OEs across an entire experiment

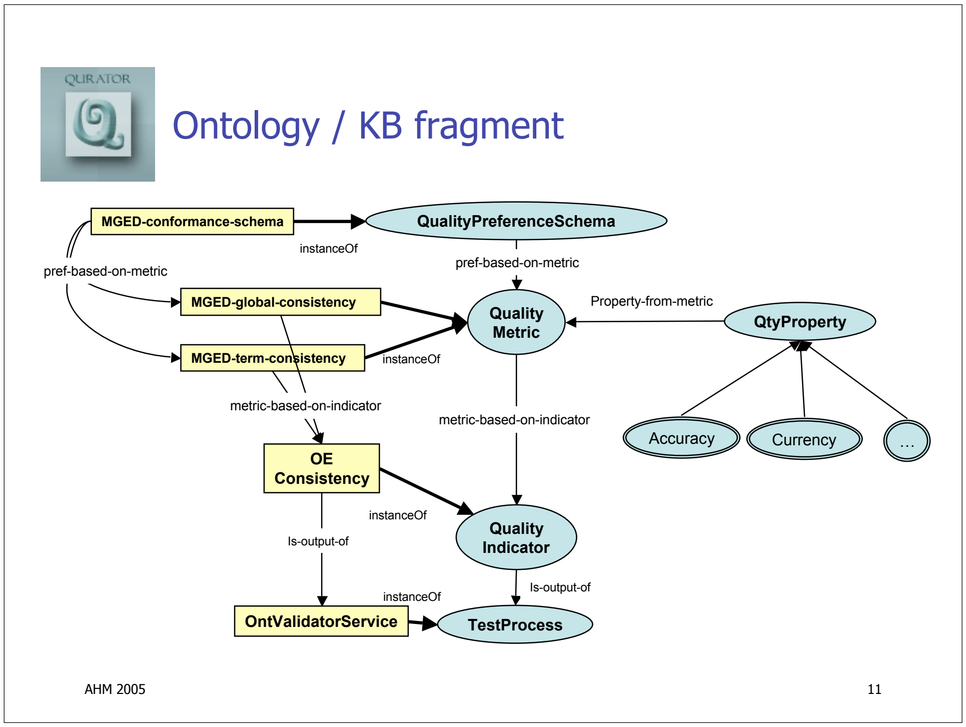
A **Quality Preference Schema** is based on one or more **Quality Metrics** and indicates how to produce a quality-based view of the data

- e.g. an "acceptable" MAGE-ML datafile may be defined as one in which all OEs must conform



Qurator conceptual framework


Qurator environment
• Data Annotation
• Ontology browse and search
Generic Reference IQ Ontology
KB of user quality concepts
Presentation layer
Quality-aware data visualization models
Selection of data sources based on their quality
Quality preferences layer
• Logic-based class definition
• DL-based / Rule-based
• Ranking based on metrics
Baseline layer
Quality metrics (functions of indicators)
Quality indicators (any data / metadata)
Data Model
Data Testing Process Model
Binding layer
Data binding
Tasks binding
Data layer
Schema
Schema
DB
DB
Process layer
Executable Task
Executable Task



QURATOR

Ontology / KB fragment

MGED-conformance-schema

QualityPreferenceSchema

instanceOf

pref-based-on-metric

pref-based-on-metric

MGED-global-consistency

MGED-term-consistency

Quality Metric

Property-from-metric

QtyProperty

instanceOf

metric-based-on-indicator

metric-based-on-indicator

Accuracy

Currency

…

OE Consistency

instanceOf

Is-output-of

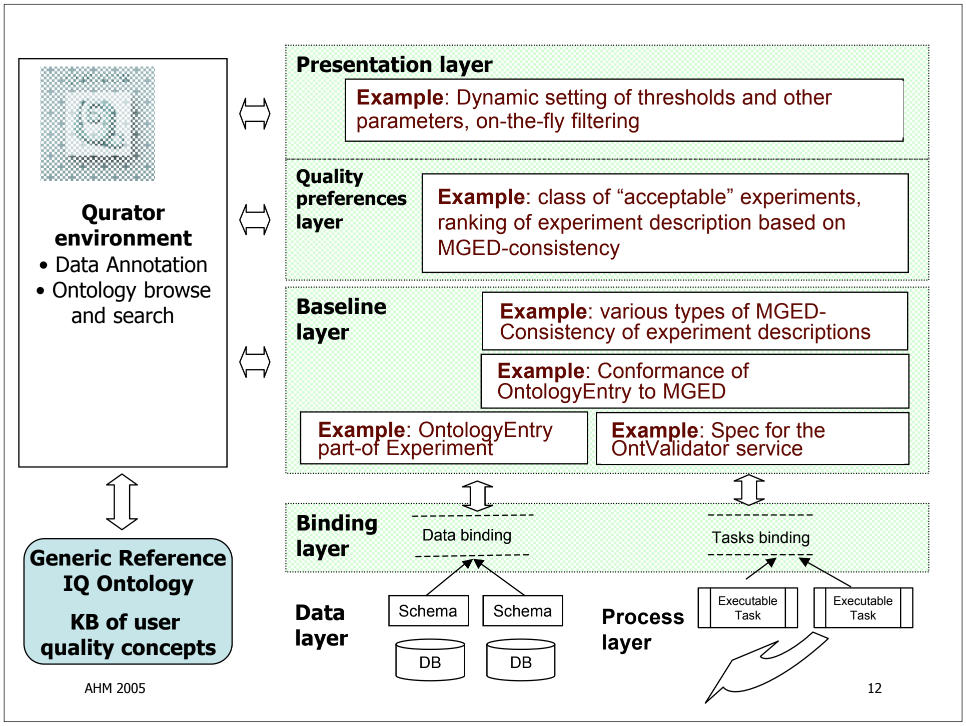
Quality Indicator

Is-output-of

instanceOf

OntValidatorService

TestProcess



Qurator environment

- Data Annotation
- Ontology browse and search

Generic Reference IQ Ontology

KB of user quality concepts

Presentation layer

Example: Dynamic setting of thresholds and other parameters, on-the-fly filtering

Quality preferences layer

Example: class of “acceptable” experiments, ranking of experiment description based on MGED-consistency

Baseline layer

Example: various types of MGED-Consistency of experiment descriptions

Example: Conformance of OntologyEntry to MGED

Example: OntologyEntry part-of Experiment

Example: Spec for the OntValidator service

Binding layer

Data binding

Tasks binding

Data layer

Schema

Schema

DB

DB

Process layer

Executable Task

Executable Task

IQ ontology FAQ

Why use an ontology at all?

- The formal ontology (expressed in OWL DL) explicitly specifies our IQ conceptualisation
- We can align it with related ontologies e.g. ^{my}Grid data ontology
- We can use a reasoner to check consistency/integrity
- In certain cases we can classify domain-specific IQ elements automatically (e.g. **OE Consistency** is related to a kind of **Accuracy**…)

Why are the domain-specific concepts instances (not classes)?

- Easier to maintain - the core ontology doesn't change when new bits of domain-specific apparatus are added

Why are the "generic" IQ properties included?

- Users have the option to browse/query the ontology/KB both "bottom-up" and "top-down"…

Sample IQ service: OntValidator

The OntValidatorService implementation is a Web service that

- takes a URI (LSID) to an experiment data file (XML doc) and a set of data bindings
- returns a set of annotations for the OEs in that file

Data bindings for OntValidatorService inputs are to OntologyEntry elements in MAGE-ML documents, via XPath expressions

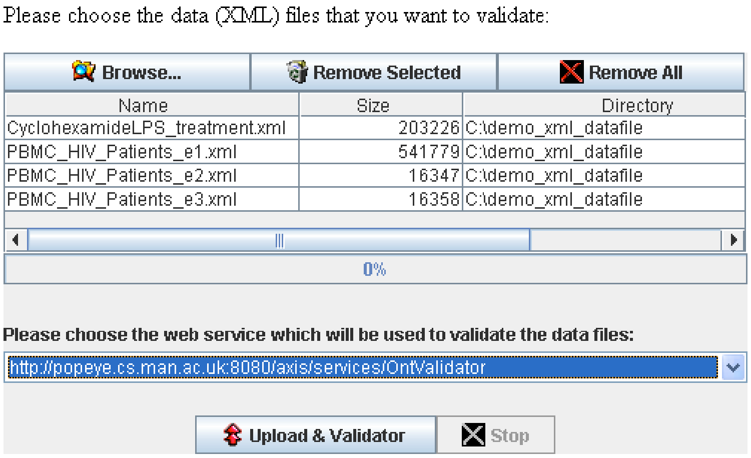
Annotations are RDF statements about the original experiment data file (resource)

For each OntologyEntry, three annotation values are possible

- **OK** - class/individual combination conforms to the ontology
- **BAD_IND** - individual is not defined for this class
- **BAD_CLASS** - class is not defined

Currently, we have simple preferences written as RuleML rules

OntValidator service Web client



Please choose the data (XML) files that you want to validate:

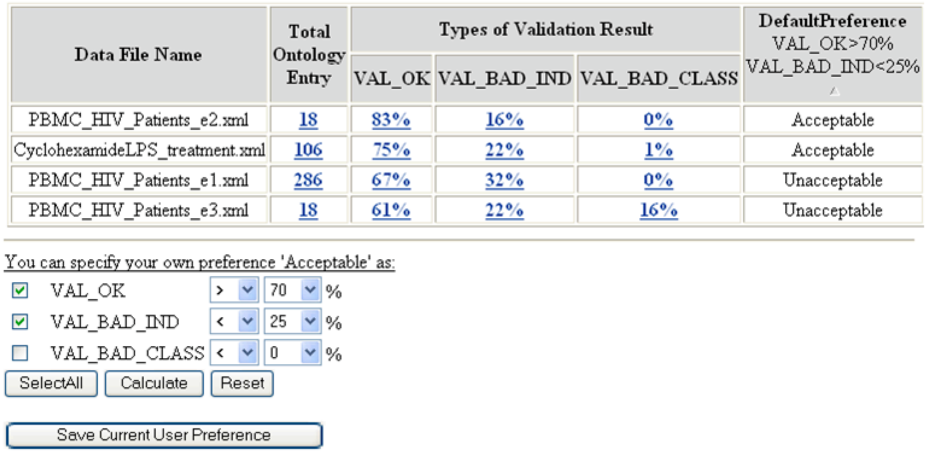
Name	Size	Directory
CyclohexamideLPS_treatment.xml	203226	C:\demo_xml_datafile
PBMC_HIV_Patients_e1.xml	541779	C:\demo_xml_datafile
PBMC_HIV_Patients_e2.xml	16347	C:\demo_xml_datafile
PBMC_HIV_Patients_e3.xml	16358	C:\demo_xml_datafile

0%

Please choose the web service which will be used to validate the data files:

http://popeye.cs.man.ac.uk:8080/axis/services/OntValidator

OntValidator results page



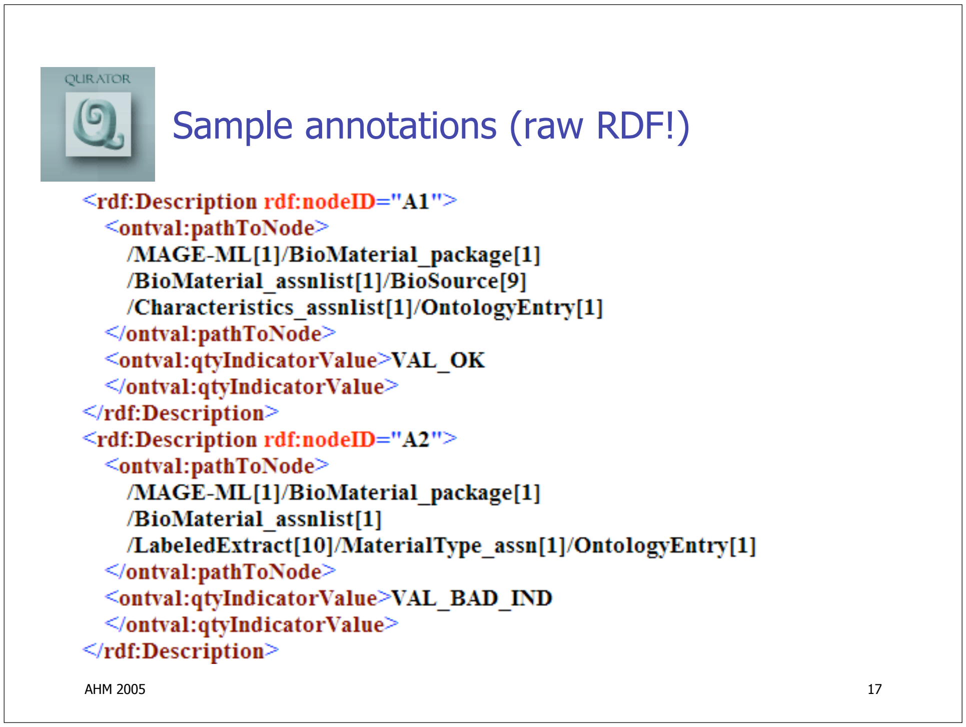
Data File Name	Total Ontology Entry	Types of Validation Result			DefaultPreference VAL_OK>70% VAL_BAD_IND<25%
		VAL_OK	VAL_BAD_IND	VAL_BAD_CLASS	
PBMC_HIV_Patients_e2.xml	18	83%	16%	0%	Acceptable
CyclohexamideLPS_treatment.xml	106	75%	22%	1%	Acceptable
PBMC_HIV_Patients_e1.xml	286	67%	32%	0%	Unacceptable
PBMC_HIV_Patients_e3.xml	18	61%	22%	16%	Unacceptable

You can specify your own preference 'Acceptable' as:

☑ VAL_OK > 70 %

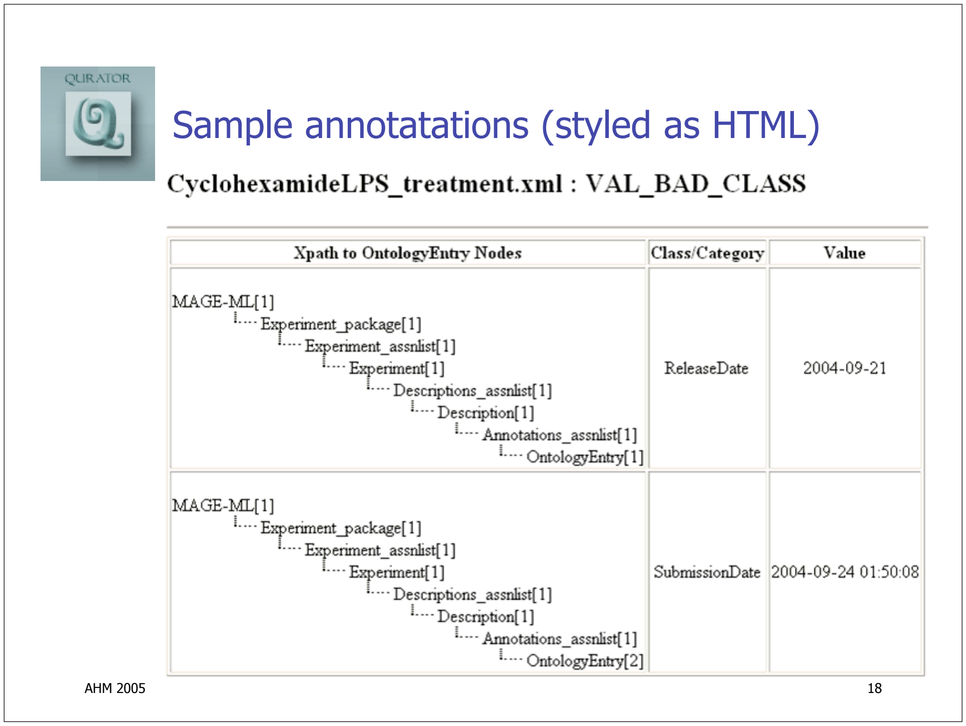
☑ VAL_BAD_IND < 25 %

□ VAL_BAD_CLASS < 0 %



Sample annotations (raw RDF!)

```
<rdf:Description rdf:nodeID="A1">
  <ontval:pathToNode>
    /MAGE-ML[1]/BioMaterial_package[1]
    /BioMaterial_assnlist[1]/BioSource[9]
    /Characteristics_assnlist[1]/OntologyEntry[1]
  </ontval:pathToNode>
  <ontval:qtyIndicatorValue>VAL_OK
  </ontval:qtyIndicatorValue>
</rdf:Description>
<rdf:Description rdf:nodeID="A2">
  <ontval:pathToNode>
    /MAGE-ML[1]/BioMaterial_package[1]
    /BioMaterial_assnlist[1]
    /LabeledExtract[10]/MaterialType_assn[1]/OntologyEntry[1]
  </ontval:pathToNode>
  <ontval:qtyIndicatorValue>VAL_BAD_IND
  </ontval:qtyIndicatorValue>
</rdf:Description>
```

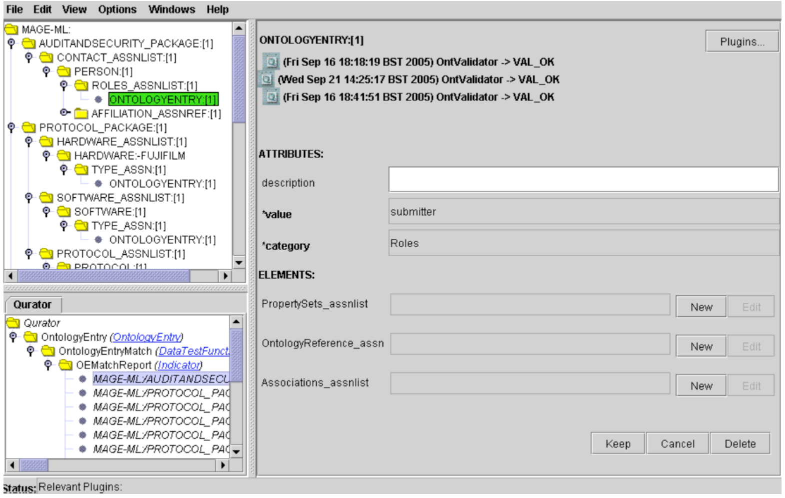


Sample annotatations (styled as HTML)

CyclohexamideLPS_treatment.xml : VAL_BAD_CLASS

Xpath to OntologyEntry Nodes	Class/Category	Value
MAGE-ML[1] └··· Experiment_package[1] └··· Experiment_assnlist[1] └··· Experiment[1] └··· Descriptions_assnlist[1] └··· Description[1] └··· Annotations_assnlist[1] └··· OntologyEntry[1]	ReleaseDate	2004-09-21
MAGE-ML[1] └··· Experiment_package[1] └··· Experiment_assnlist[1] └··· Experiment[1] └··· Descriptions_assnlist[1] └··· Description[1] └··· Annotations_assnlist[1] └··· OntologyEntry[2]	SubmissionDate	2004-09-24 01:50:08

Getting Qurator closer to biologists: a Pedro plugin client



```
MAGE-ML:
  AUDITANDSECURITY_PACKAGE:[1]
    CONTACT_ASSNLIST:[1]
      PERSON:[1]
        ROLES_ASSNLIST:[1]
          ONTOLOGYENTRY:[1]
        AFFILIATION_ASSNREF:[1]
  PROTOCOL_PACKAGE:[1]
    HARDWARE_ASSNLIST:[1]
      HARDWARE:-FUJIFILM
        TYPE_ASSN:[1]
          ONTOLOGYENTRY:[1]
    SOFTWARE_ASSNLIST:[1]
      SOFTWARE:[1]
        TYPE_ASSN:[1]
          ONTOLOGYENTRY:[1]
    PROTOCOL_ASSNLIST:[1]
      PROTOCOL:[1]
```

```
Qurator
  OntologyEntry (OntologyEntry)
    OntologyEntryMatch (DataTestFunct
      OEMatchReport (Indicator)
        MAGE-ML:/AUDITANDSECU
        MAGE-ML:/PROTOCOL_PAC
        MAGE-ML:/PROTOCOL_PAC
        MAGE-ML:/PROTOCOL_PAC
        MAGE-ML:/PROTOCOL_PAC
        MAGE-ML:/PROTOCOL_PAC
```

ONTOLOGYENTRY:[1]

(Fri Sep 16 18:18:19 BST 2005) OntValidator -> VAL_OK
(Wed Sep 21 14:25:17 BST 2005) OntValidator -> VAL_OK
(Fri Sep 16 18:41:51 BST 2005) OntValidator -> VAL_OK

ATTRIBUTES:

description:
*value: submitter
*category: Roles

ELEMENTS:

PropertySets_assnlist
OntologyReference_assn
Associations_assnlist

Status: Relevant Plugins:

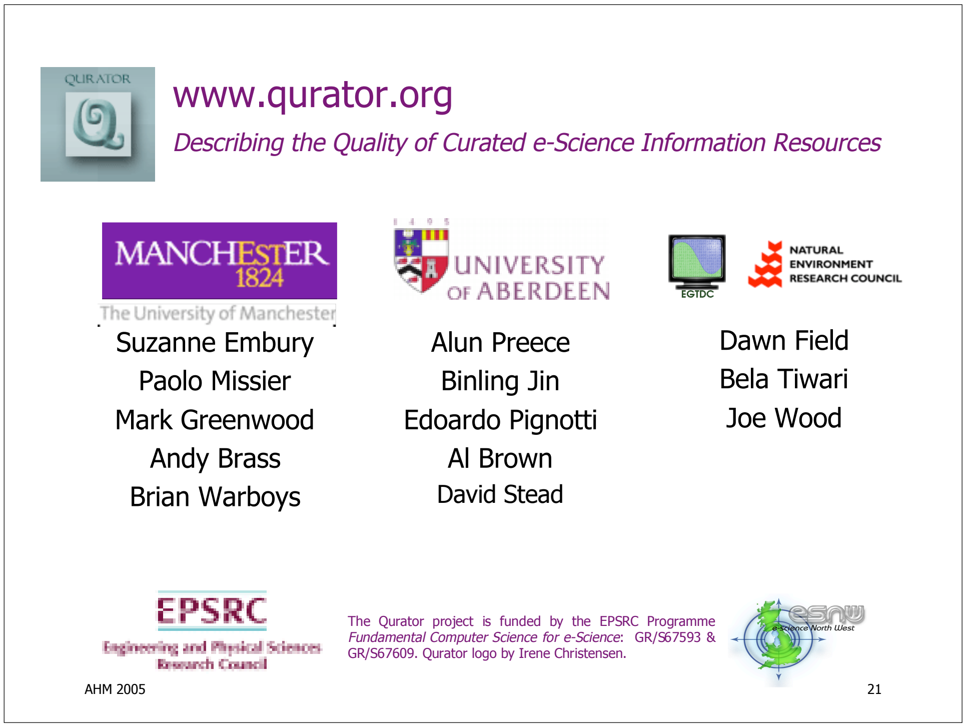
Conclusion

Core IQ framework and ontology is in place:

- Ontology scope extends
  - ✓ "up" to generic IQ concepts
  - ✓ "down" to domain-specific IQ concepts
- Bindings map things in the IQ-space to scientific data resources
- Test processes assign IQ annotations to data resources
- Preferences give users quality-based views on data
- We have a simple vertical demo in transcriptomics

We are in the process of

- Extending the framework at all levels, initially in proteomics
- Using the framework and demo to elicit user feedback and revised requirements
- Designing experiments to establish cost/benefits of the approach



www.qurator.org

Describing the Quality of Curated e-Science Information Resources

MANCHESTER 1824 The University of Manchester	UNIVERSITY OF ABERDEEN	EGTDC NATURAL ENVIRONMENT RESEARCH COUNCIL
Suzanne Embury	Alun Preece	Dawn Field
Paolo Missier	Binling Jin	Bela Tiwari
Mark Greenwood	Edoardo Pignotti	Joe Wood
Andy Brass	Al Brown	
Brian Warboys	David Stead	

EPSRC Engineering and Physical Sciences Research Council

The Qurator project is funded by the EPSRC Programme *Fundamental Computer Science for e-Science*: GR/S67593 & GR/S67609. Qurator logo by Irene Christensen.

esnw e-Science North West

AHM 2005

21