Towards the Management of Information Quality in Proteomics

Alun Preece, Binling Jin
University of Aberdeen
Computing Science
Aberdeen, UK
{apreece, bjin}@csd.abdn.ac.uk

Paolo Missier, Suzanne Embury
University of Manchester
School of Computer Science
Manchester, UK
{pmisser, embury}@cs.man.ac.uk

David Stead, Al Brown
University of Aberdeen
Molecular and Cell Biology
Aberdeen, UK
{d.stead, al.brown}@abdn.ac.uk

Abstract

We outline the application of a framework for managing information quality (IQ) in proteomics. The approach allows scientists to define the quality characteristics that are of importance in their particular domain, by extending a generic ontology of IQ concepts. Two quality indicators are defined for proteomic experiments: hit ratio and mass coverage. We describe how our framework allows experiments marked-up in a standard format (e.g. PEDRo) to be annotated with these computed indicators, and how the annotations can be viewed using a convenient plugin to the commonly-used Pedro data entry tool.

1. Introduction

There is a debate in progress that seeks to define what information is required when reporting the results of protein identifications by mass spectrometry. For peptide mass fingerprinting, the journal *Molecular and Cellular Proteomics* has proposed publication guidelines that include the number of peptides matched to the identified protein, the number that were not matched in the mass spectrum, and the sequence coverage observed [1].

It would be useful for biologists seeking to interpret the results of proteomic experiments to have a tool that can apply certain quality preferences to a list of protein matches, for the purposes of accepting or questioning a protein identification result. Such functionality would be particularly useful to scientists wishing to compare protein identification results generated by other labs with those produced within their own.

The Qurator project[1] [3, 4] is developing techniques for managing information quality (IQ) [2] in e-science domains including proteomics and transcriptomics. 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) [6], we allow scientists to define the quality characteristics that are of importance in their particular domain. These domain-specific IQ descriptions are defined by extending a core *IQ ontology*, which classifies new descriptors within an 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.

In the Qurator approach to quality-aware information management, IQ descriptions for specific resources (e.g. experiment data files) are computed and associated with the resources as *annotations* related to concepts in the IQ ontology. IQ annotations on a resource are essentially quality metadata, and can be used to derive higher-order IQ metrics or rankings over sets of resources. Annotations are generated by data checking services, and are used by quality preference services to display and rank resources.

Data resources are modelled by concepts in the IQ ontology, so that the ontology can express which kinds of IQ descriptor make sense for which kinds of resource. The relationship between actual types of resource (for example a particular data model expressed as an XML Schema) and the abstract model of this resource in the IQ ontology is

[1]Funded by the EPSRC Programme Fundamental Computer Science for e-Science: GR/S67593 & GR/S67609 — *Describing the Quality of Curated e-Science Information Resources, http://www.qurator.org.*

captured by a *binding*. In practice, much of this framework is hidden from the user. Our aim is to embed the framework components in quality-aware versions of familiar desktop tools.

In effect, the Qurator approach supports management of knowledge about the quality of various data sets. Our framework in intended to allow a lab to build up a set of standard and trusted tools for assessing the quality of new data, and recording these assessments for future use. The framework facilitates reuse of knowledge of quality between projects, but also allows specific scientists or projects within a lab to add their own project-specific notions of quality.

This paper illustrates how the elements of our framework can be used in practice, by applying them to IQ management in proteomics. Section 2 introduces two IQ indicators for protein identification experiments, Section 3 gives further detail of the Qurator framework in this context, and Section 4 shows how the various components have been implemented within the Pedro data entry tool used by biologists. The two proteomics indicators should be viewed as lab-specific formalised IQ definitions that improve on the generic indicators accepted by the community, and that — through our framework and its embedding in the Pedro desktop tool — now become resources for the wider community to use.

2. IQ Indicators in Protein Identification

The most important output from many proteomics experiments are the identities of the proteins of interest. The accuracy of these identifications is of crucial importance for the correct biological interpretation of such experiments. Almost all protein identifications are made using mass spectrometry data and concerns have been raised about the quality of some of this data and the potential for false positive identifications [1]. At present, there are no clear acceptance criteria for mass spectrometry-based protein identifications, nor agreement on how to measure the quality of these data.

In "classical" proteomics experiments, proteins are extracted from the biological samples under study and separated by 2-dimensional gel electrophoresis as a prelude to their quantification and identification. Protein identifications in such experiments are routinely obtained by peptide mass fingerprinting (PMF). In this technique, the protein within the gel spot is first digested with an enzyme that cleaves the protein sequence at certain predictable sites. The fragments of protein that result (called peptides) are extracted and their masses are measured in a mass spectrometer. The experimental list of peptide masses (the "fingerprint") is then compared against theoretical peptide mass lists, derived by simulating the process of digestion on sequences extracted from a protein database (e.g. NCBInr[2]). Since, for various reasons, it is unlikely that an exact match will be found, the protein identification search engines (e.g. Mascot[3]), that perform this task typically return a list of potential protein matches, ranked in order of search score. Different search engines calculate these scores in different ways, so their results are not directly comparable. It may therefore be difficult for the experimenter and subsequent users of the data to decide whether a particular protein identification is acceptable or not.

Two readily accessible indicators that are independent of the particular search engine used can be used to rank protein identification data (these definitions are based on guidelines from the proteomics literature, e.g. [1]):

- *Hit ratio*: the number of peptide masses matched, divided by the number of peptide masses submitted to the search. This indicator effectively combines the number of matched peptides and the number of unmatched peptides mentioned above. Ideally, most of the peaks in the spectrum should be accountable for by the protein identified, but because of the presence of other components and unpredicted modifications to the matched peptides the hit ratio is unlikely to reach unity.

- *Mass coverage*: the number of amino acids contained within the set of matched peptides, expressed as a fraction of the total number of amino acids making up the sequence of the identified protein and multiplied by the total mass (in kDa) of the protein. We consider mass coverage superior to sequence coverage, because peptide mass fingerprints of equal quality give low (percent) sequence coverage for large proteins and high (percent) coverage for small proteins.

These two indicators can be combined in a logical expression that allows us to classify protein matches as acceptable or unacceptable. A software tool could then allow the user-scientist to set threshold values (that is, acceptance criteria) for each metric independently and to see the effect in real time of altering any or all of the threshold values on the acceptability of the data set. This is an example of the kind of quality-aware data analysis that Qurator aims to support.

3. Overview of the Qurator IQ Framework

The embedding of domain-specific quality indicators within standard tools requires a means by which the (abstract) quality requirements of users can be associated with the (concrete) resources available to them. For example,

[2] ftp://ftp.ncbi.nlm.nih.gov/blast/db/blastdb.html
[3] http://www.matrixscience.com/

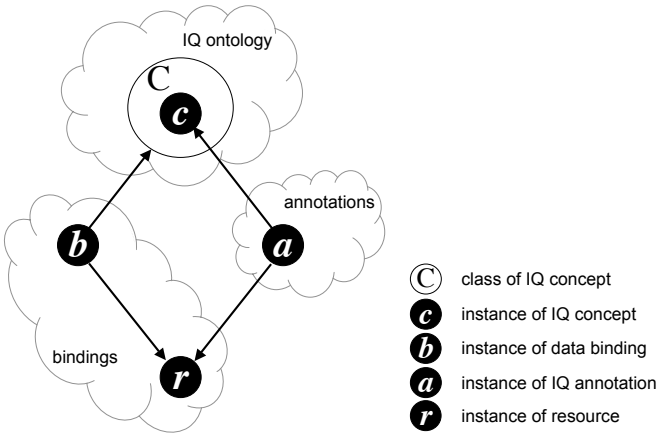

Figure 1. Overview of the elements of the Qurator IQ framework.

a user scientist may be concerned about the accuracy of a particular data collection. This abstract concern must be translated into a concrete query in terms of available quality indicators on the data set in question. In other words, we require a means to navigate from the conceptual view of the world described by the quality ontology to the resources that make up the computation environment in which the scientist is working, and *vice versa*.

Figure 1 sets out the components of the Qurator framework that support this navigation, and the relationships between them. At the top we have the IQ ontology itself, which includes definitions of domain-independent IQ concepts such as Accuracy and also classes of domain-specific indicator such as Hit Ratio and Mass Coverage. The IQ ontology also models the various kinds of abstract data entities to which we might wish to apply IQ indicators, such as a Protein Hit obtained from a PMF database search. The ontology then captures the fact that the Hit Ratio indicator applies to a Protein Hit. Finally, the ontology defines the various kinds of data checking function available.

At the bottom of Figure 1 we have instances of specific resources (r), for example a particular protein hit derived from a database search. These are often represented in XML; in the proteomics case the PEDRo data model [5] is widely used for this purpose, by means of the PEDRo XML Schema[4].

The bindings and annotations shown in the middle of the diagram are the links that allow navigation between the conceptual and the technical worlds. An instance of a binding (b) relates a resource instance (r) to the corresponding class in the ontology (C); e.g. a specific PEDRo protein hit list to the model class Protein Hit. One of the main uses of bindings is to determine which parts of the IQ conceptualisation are relevant to a particular concrete data model. So, for example, the binding from a PEDRo protein hit structure to the ontology Protein Hit class also lets us identify relevant indicators (such as Hit Ratio) and associated checking functions.

Finally, an instance of an annotation (a) relates a specific resource instance (r) to the instance of a quality concept (c). For example, an instance of Hit Ratio with a specific value (e.g. 0.45) might be associated with an individual concrete PEDRo protein hit via an annotation. The IQ instance is said to annotate the associated resource.

To illustrate the roles of the components further, Figure 2 shows some fragments of the components from the proteomics domain. At the top, a fragment of the IQ ontology is shown, which includes the concept of a *ProteinHit* data entity and a *HitRatio* quality indicator. These are bound to the resources that implement them in the real world: the protein hits in the data set of interest and the program code that evaluates hit ratios, respectively. The annotation of the data set using the quality indicator is also recorded.

4. Embedding IQ into Proteomics Tools

The Pedro[5] data entry tool is commonly used in proteomics — and several other e-Science domains — to enter and manage XML-based data. To make our approach convenient to user-scientists we have therefore embedded elements of the Qurator framework in the Pedro desktop software. Figure 3 shows a screenshot of the augmented Pedro tool. The top-left area of the screen is the XML document tree and the right-hand panel is the data entry area. When the user starts up the tool, they are prompted to select the data model on which they will work, for example the PEDRo model for proteomics data. Choice of the data model then drives the content of the top-left and right-hand panels in the standard Pedro environment: users may enter and edit data, and export it to various formats.

Our augmented version of Pedro introduces the lower-left panel, which contains a tree view of the portions of the IQ ontology relevant to the loaded data model. These elements are obtained by querying the ontology dynamically. For the PEDRo data model, they include domain-specific elements such as HitRatio and MassCoverage as well as associated generic concepts such as Accuracy. This panel allows users to discover available indicators for the data model at hand, and follow hyperlinks to explore the ontology.

The augmented tool also uses Pedro's plugin model to invoke any available test functions for the model at hand. If the user clicks on the *Plugins* button at the top-right of Figure 3 they are offered two services, to annotate the data with

[4] http://pedro.man.ac.uk/files/PEDRoSchema.xsd

[5] http://pedrodownload.man.ac.uk/

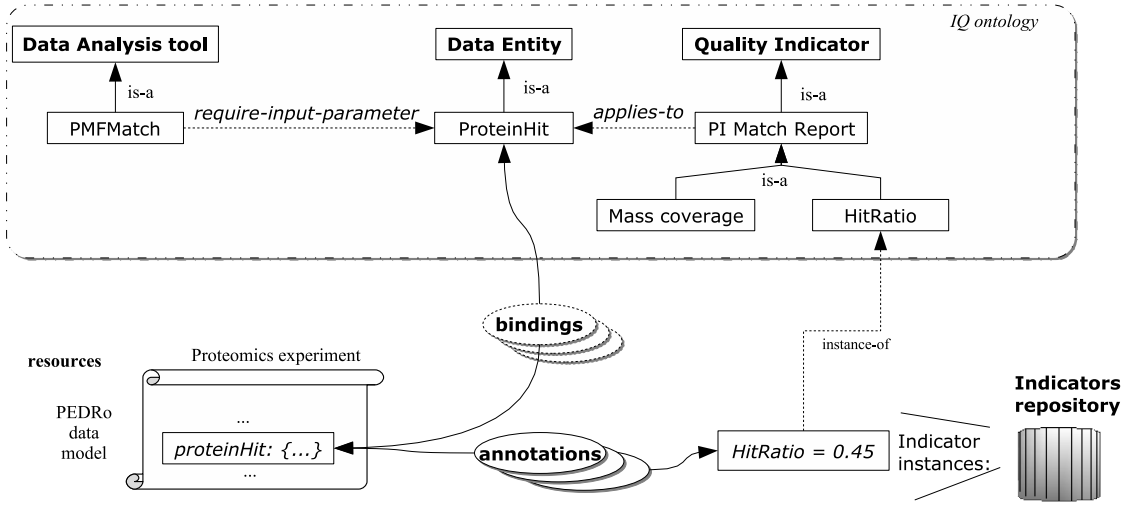

Figure 2. Example elements of the Qurator framework in the proteomics domain.

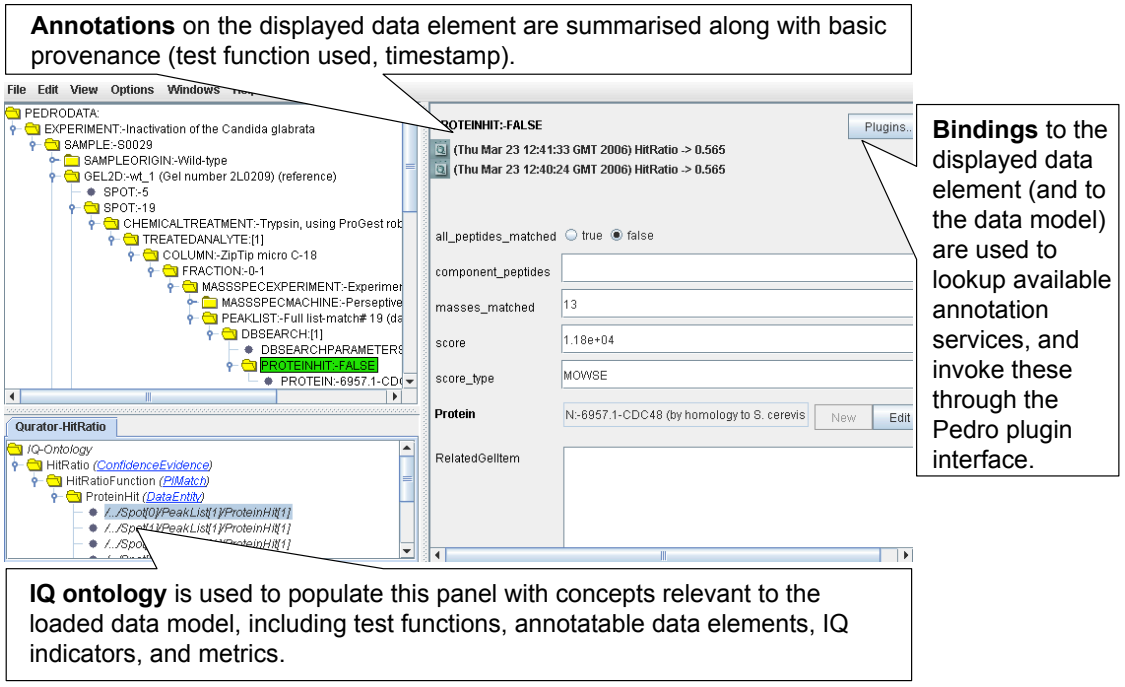

Figure 3. Augmented "quality-aware" version of the Pedro data entry tool.

respect to the HitRatio and MassCoverage indicators which are important to biologists (see Section 2). The choice of available service is determined dynamically, using available bindings obtained from an online *binding repository*. Invoking these services results in annotations being added to an online *annotation repository*. The augmented Pedro desktop tool is configured to act as a client to these two repositories, both of which are central components of the Qurator infrastructure and shared with various other IQ services [3]. By querying the annotation repository, Pedro can retrieve any annotations associated with the displayed data elements shown in the right-hand panel.

It is worth emphasising that the augmented Pedro tool is intended to be a natural and convenient way for user-scientists to access the facilities of the Qurator framework; however, there is nothing in the framework specific to its use in the Pedro tool. In fact, the various data-checking services and repositories can also be accessed via a Web browser through HTML and JavaScript front-ends [3], and we are currently developing interfaces that allow the components of the framework to be invoked as part of e-Science workflows using the Taverna environment[6].

5. 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. To the best of our knowledge, this ontology is the first systematic attempt to capture generic and domain-dependent quality descriptors in a semantic model. In this paper we discuss the application of the Qurator framework in proteomics; again, to the best of our knowledge, this is a novel approach to scientific information management in this domain. We are currently gathering feedback from our collaborating users, after which we aim to further develop the IQ framework and associated toolset.

[6] http://taverna.sourceforge.net

References

[1] S. Carr, R. Aebersold, M. Baldwin, A. Burlingame, K. Clauser, and A. Nesvizhskii. Editorial: The need for guidelines in publication of peptide and protein identification data. *Molecular and Cellular Proteomics*, 3:531–533, 2004.

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

[3] P. Missier, S. Embury, M. Greenwood, A. Preece, and B. Jin. An ontology-based approach to handling information quality in e-science. In *Proc 4th e-Science All Hands Meeting*, 2005.

[4] P. Missier, A. Preece, S. Embury, B. Jin, M. Greenwood, D. Stead, and A. Brown. Managing information quality in e-science: A case study in proteomics. In *Proc 1st Workshop on Quality of Information Systems (QoIS 2005), Lecture Notes in Computer Science, Volume 3770*, pages 423–432. Springer, 2005.

[5] C. F. Taylor, N. W. Paton, K. L. Garwood, P. D. Kirby, D. A. Stead, Z. Yin, E. W. Deutsch, L. Selway, J. Walker, I. Riba-Garcia, S. Mohammed, M. J. Deery, J. A. Howard, T. Dunkley, R. Aebersold, D. B. Kell, K. S. Lilley, P. Roepstorff, J. R. YatesIII, A. M. Brass, A. J. Brown, P. Cash, S. J. Gaskell, S. J. Hubbard, and S. G. Oliver. A systematic approach to modeling, capturing, and disseminating proteomics experimental data. *Nature Biotechnology*, 21(3):247–254, Mar. 2003.

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