Journal of Biomedical Informatics 45 (2012) 782–794

Contents lists available at SciVerse ScienceDirect

Journal of Biomedical Informatics

journal homepage: www.elsevier.com/locate/yjbin

Translating standards into practice – One Semantic Web API for Gene Expression

Helena F. Deus [a,*,1], Eric Prud'hommeaux [b,1], Michael Miller [c], Jun Zhao [d], James Malone [e], Tomasz Adamusiak [e], Jim McCusker [f], Sudeshna Das [g], Philippe Rocca Serra [h], Ronan Fox [a], M. Scott Marshall [i,j]

[a] Digital Enterprise Research Institute (DERI), National University of Ireland, Galway, Ireland
[b] World Wide Web Consortium (W3C), MIT, Cambridge, MA, USA
[c] Institute for Systems Biology (ISB), Seattle, WA, USA
[d] The Image Bioinformatics Research Group, University of Oxford, Oxford, UK
[e] European Bioinformatics Institute (EBI), Cambridge, UK
[f] Tetherless World Constellation, Rensselaer Polytechnic Institute (RPI), NY, USA
[g] Massachusetts General Hospital & Harvard Medical School, Boston, MA, USA
[h] Oxford e-Research Centre, University of Oxford, Oxford, United Kingdom
[i] Leiden University Medical Center, University of Amsterdam, Netherlands
[j] Informatics Institute, University of Amsterdam, Netherlands

ARTICLE INFO

Article history:
Received 9 August 2011
Accepted 13 March 2012
Available online 24 March 2012

Keywords:
Semantic Web technologies
Genomics
Transcriptomics
Pharmacogenomics
Microarrays
Linked Data

ABSTRACT

Sharing and describing experimental results unambiguously with sufficient detail to enable replication of results is a fundamental tenet of scientific research. In today's cluttered world of "-omics" sciences, data standards and standardized use of terminologies and ontologies for biomedical informatics play an important role in reporting high-throughput experiment results in formats that can be interpreted by both researchers and analytical tools. Increasing adoption of Semantic Web and Linked Data technologies for the integration of heterogeneous and distributed health care and life sciences (HCLSs) datasets has made the reuse of standards even more pressing; dynamic semantic query federation can be used for integrative bioinformatics when ontologies and identifiers are reused across data instances. We present here a methodology to integrate the results and experimental context of three different representations of microarray-based transcriptomic experiments: the Gene Expression Atlas, the W3C BioRDF task force approach to reporting Provenance of Microarray Experiments, and the HSCI blood genomics project. Our approach does not attempt to improve the expressivity of existing standards for genomics but, instead, to enable integration of existing datasets published from microarray-based transcriptomic experiments. SPARQL Construct is used to create *a posteriori* mappings of concepts and properties and linking rules that match entities based on query constraints. We discuss how our integrative approach can encourage reuse of the Experimental Factor Ontology (EFO) and the Ontology for Biomedical Investigations (OBIs) for the reporting of experimental context and results of gene expression studies.

1. Introduction

Personalized genomic information is becoming increasingly relevant for targeted therapy strategies such as pharmacogenomics [1]. The US Food and Drug Administration (FDA), for example, is now requesting genetic testing to determine the applicability of certain drugs [2]. The falling prices of high-throughput technologies such as microarrays, next generation sequencing (NGS) and mass spectrometry are also enabling genetic testing on a much larger scale by enabling the simultaneous measurement of thousands of genes/proteins in a single high-throughput analysis. Making full use of this data for targeted therapy requires (1) standardization of data collection and analysis in order to ensure that results can be replicated; (2) shared and integrated experimental results across multiple institutions so that genetic cohorts of patients can be assembled; and (3) the ability to identify, among hundreds of differentially expressed genes, a small list of genes that can be targeted for genetic testing and targeted treatment.

* Corresponding author.
E-mail addresses: helena.deus@deri.org (H.F. Deus), eric@w3.org (E. Prud'hommeaux), mmiller@systemsbiology.org (M. Miller), jun.zhao@zoo.ox.ac.uk (J. Zhao), malone@ebi.ac.uk (J. Malone), tomasz@ebi.ac.uk (T. Adamusiak), james.mccusker@yale.edu (J. McCusker), sudeshna_das@harvard.edu (S. Das), proccaserra@gmail.com (P. Rocca Serra), ronan.fox@deri.org (R. Fox), marshall@science.uva.nl (M.S. Marshall).
[1] These authors contributed equally.

1532-0464/$ - see front matter © 2012 Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.jbi.2012.03.002

1.1. Standards for genomics: proposed and in practice

Standardization is tied directly to the legacy value of high-throughput experimental results as it ensures that the methodologies and experimental context used to produce a dataset are described unambiguously and with sufficient detail to enable other scientists to replicate its results [3,4]. It is only when experimental results can be accurately replicated and linked to phenotypical information that discoveries can be translated into a clinical scenario to improve decision making [5]. Furthermore, reuse of standards when linking high-throughput experimental results to phenotypes can be essential for making use of biological resources such as tissue banks [6] and for linking across several "-omic" sciences.

Standardization efforts and the creation of guidelines have been under way for recording and sharing many aspects of functional genomics experiments related to its experimental context. Once a standard for representing experimental biological data becomes widely used by the communities involved in generating and using the data, it can be reliably applied in tool development and data sharing. The Minimum Information about a Microarray Experiment (MIAME) [3], for example, is a widely used guideline, created by the microarray community for describing a microarray experiment. MIAME has gained momentum when journals began, in 2002, requesting gene expression experiments to be MIAME-compliant before publication. The Microarray and Gene Expression (MAGE) Object Model began development shortly after MIAME and a mapping to an XML format was created, MAGE-ML, to capture requirements of the MIAME guideline [7]. Several other guidelines and standard formats for "omics" results have been developed since: in specific contexts, such as RNAi screening experiments, the Minimum Information About an RNAi Experiment guidelines [8] have been issued to incorporate information about experimental context. In order to promote cross talk between individual guideline checklists and modular reuse, Minimum Information for Biological and Biomedical Investigations started compiling proposed guidelines for reporting experimental results in biosciences [4].

Standards can also become a powerful enabler for data integration: once a standardized model for representing experimental "omics" data becomes widely used and accepted by the communities involved in generating and sharing the data, the syntactic barrier that prevents multiple source datasets from being merged is eliminated. Studies and approaches to integrating experimental results from several other "-omics" sciences such as genomics and proteomics have also relied on reusing standards [9,10].

Experimental results from "-omics" sciences, however, are difficult to standardize. One of the reasons causing this has been that changes in technologies over the years have brought multiple methods for measuring altered gene expression – from different cDNA microarrays platforms to the different next generation sequencing platforms. In fact, Genomic Technologies and Biotechnology seems to be advancing more rapidly than our ability to standardize its methods and results.

1.2. The "sticking point" for standards: tools and databases implementing standards

To achieve wide acceptance, standards and guidelines must successfully describe the requirements in terms that the final users (i.e. the biology domain experts generating the data) can use and understand [11]. However, their usability is limited without the appropriate tools to produce and manipulate the metadata structure. For example, insufficient tool support, such as a lack of simple user interfaces, has made it impractical to apply MAGE-ML in the daily practice of most wet lab scientists. These conditions led to the creation of MAGE-TAB [12], a simpler spread-sheet based format for reporting microarray experimental context, developed by reusing standards-compliant tools (Perl packages [13], Annotare [14]). This has further boosted the popularity and acceptance of MIAME as a standard and three large microarray databases have since integrated the MIAME guideline in their repositories: ArrayExpress from the European Bioinformatics Institute (EBI) [15], which relies on MAGE-TAB format to help provide consistent syntactic support; the Gene Expression Omnibus (GEO) from the National Institutes of Health [16] which makes use of the MIAME notation in Markup Language [17] and the Simple Omnibus Format in Text [18] as alternatives to structure similar information; and the Center for Information Biology gene Expression database [19], also using MAGE compliant formats. Data integration across databases is somewhat enabled but comes at the cost of converting the native database syntax, e.g. conversion between GEO and ArrayExpress is a non-trivial exercise.

A second barrier to the adoption of standards in genomics for reporting experimental results has been that standards implemented in tools such as MAGE-TAB support only description of the experimental context but often do not include sufficient details for reporting experimental results nor the statistical algorithms and parameters used in its analysis. These are increasingly relevant for an accurate interpretation of the results as they enhance provenance information pertaining to the data transformation process and quality control. Even when this information is available in the models, its inclusion in tools and databases is not always straightforward. For example, in the light of recent additions to standard models, it is now possible to report '*p*-value' and '*q*-value', however, MIAME based tools do not capture this information since relevant statistical measurements to be reported must often be decided *ad hoc,* according to the experimental setup and the method used to analyse the data, making genomics experimental context and its results challenging to standardize and integrate. As a result, experimental results are typically reported in the format of MS excel spread-sheets or as journals supplementary material. Connecting experimental context to items of interest in the results, such as drugs, genes, proteins, can also be used to further explore the metabolic processes that explain the phenotype (e.g. multiple genes affected in the same pathway) and to link to results obtained in other "omics" sciences. Tools and software that produce/consume experimental results in standardized formats, such as [20] can provide incentives for the adoption of standards by the community.

1.3. The promises and challenges of Linked Data and Semantic Web technologies

The introduction of Linked Data and Semantic Web technologies to improve discovery and integrate datasets from many different types of biological domains and from multiples sources has renewed the standardization debate [21–23]. According to [24] the term Linked Data refers to a set of best practices for publishing and connecting structured data on the Web. Semantic Web, as defined in [25] refers to an extension of the current Web in which information is given well-defined meaning by relying on a stack of technologies, e.g. the resource description framework (RDF), the web ontology language (OWL) and the Sparql Protocol and RDF Query Language (SPARQL), for describing and linking data. By reusing concepts, properties and instance identifiers defined as ontologies when describing and publishing biological datasets as Linked Data, multiple genomics data sources can be more easily integrated [26].

Agreement on standards is particularly important in Semantic Web queries as they ensure that the effort of converting data to a semantic format such as the RDF is not lost. The urgency to share and reuse ontologies has led to the creation of community portals, such as Bioportal [27] from the National Center for Biomedical Ontology, where biomedical ontologies can be easily shared, anno-