A proposed semantic framework for reporting omics investigations

Alexander Garcia[1,2,3], Jennifer Fostel[4], Norman Morrison[5], Philippe Rocca-Serra[1] & Susanna-Assunta Sansone*[1] (The MGED RSBI Working Group)

[1] *EMBL-EBI The European Bioinformatics Institute, The Wellcome Trust Genome Campus. Cambridge, UK*
[2] *The Institute for Molecular Bioscience, The University of Queenslan, Brisbane, Australia*
[3] *Australian Center for Functional Plant Genomics.*
[4] *National Institute of Environmental Health Sciences, National Center for Toxicogenomics, Research Triangle Park, USA.*
[5] *The University of Manchester, School of Computer Science, Oxford Road, Manchester, UK.* *

Corresponding author: sansone@ebi.ac.uk

Abstract. The current science landscape is rapidly evolving and it is increasingly driven by computational tasks. The deluge of data unleashed by omics-technologies, such as transcriptomics, proteomics and metabolomics, requires systematic approaches for reporting and storing the data and the experimental processes in a standard format, relating the biology information and the technology involved. Ontology-based knowledge representations have proved to be successful in providing the semantics for a standardised annotation, integration and exchange of data. The framework proposed by the MGED RSBI working group would provide semantics for upper level elements relevant to the representation and interpretation and of omics-based investigations.

1. Introduction

When the first microarray experiments were published, it became apparent that the lack of robust quality control procedures and capture of adequate biological metadata impeded the exchange and reporting of array-based transcriptomics experiments. The MIAME Checklist (Brazma *et al*., 2001) was written in response to this lack, by a group of biologists, computer scientists, and data analysts, and aims to define the minimum information required to interpret unambiguously and potentially reproduce and verify a microarray experiment. This group then went on to make its composition official and founded the Microarray Gene Expression Data (MGED) Society. The response from the scientific community has been extremely positive and currently most of the major scientific journals and funding agencies require publications describing microarray experiments to comply with MIAME standard. The adoption of these standard by public and community databases, Laboratory Information Management Systems (LIMS) and several microarray informatics tools has greatly improved the interpretation of microarray experiments described in a structured manner. The MIAME model has been adopted by other communities (reviewed by Quackenbush 2004) and as microarrays are incorporated into other complex biological investigations (including toxicogenomics, nutrigenomics and environmental genomics), it has become apparent that analogous minimal descriptors should be identified for these applications. There have been several extensions to MIAME. MIAME/Tox is an array based toxicogenomics standard developed by the EBI in collaboration with the ILSI Health and Environmental Sciences Institute (HESI), National Institute of Environmental Health Sciences (NIEHS), the National Center for Toxicogenomics, the FDA National Center for Toxicological Research (NCTR). MIAME/Env has been developed by the Natural Environmental Research Council (NERC) Data Centre to fulfill the diverse needs of those working in functional genomic of ecosystems, invertebrates and vertebrates which are not covered by the model organism community. MIAME/Tox and MIAME/Env have initiated several discussions in the academic settings as well as in the industrial and regulatory

arenas (OECD Toxicogenomics Guidelines, 2004) However it has become evident that when other –omics technologies will be used in combination with microarrays these MIAME-based checklists will soon be insufficient to serve the scope of experimenters' needs. The toxicogenomics, nutrigenomics and environmental genomics communities have soon recognized the need for a strategy that capitalizes on synergy, forming the Reporting Structure for Biological Investigations (RSBI) working group under the MGED umbrella. The RSBI working group feels that it is very important to agree on a single source of basic conceptual information relating to the reporting process of complex biological investigations, employing omics-technologies. This unified approach to describe the upper level elements relevant to the representation and interpretation and of these investigations should encompass any specific application. The possibility to enable 'semantic integration' of complex data, facilitating data mining, and information retrieval is the rationale for developing an ontologically grounded conceptual framework. Ultimately, the effort by the RSBI working group aims to constitute the foundation of standard reporting structure in publications and submission to public repositories and knowledge-bases. The need for information on which to base the evaluation and interpretation of the results underlies the objectives of presenting sufficient details to the readers and/or reviewers.

The information in complex biological investigations is highly nested and formalizing this knowledge to facilitate data representation is not a trivial task. To tackle this issue, the RSBI working group has established links with the several standardization efforts in their biological domains (as reviewed by Sansone *et al*., 2005) and is working closely with the MGED Ontology working group, the HUPO Proteomics Standards Initiative (PSI), the Standard Metabolic Reporting Structure (SMRS) group These groups can clearly draw in large numbers of experimentalists and developers and feed in the domain-specific knowledge of a wide range of biological and technical experts. This paper is organized as follows. In Section 2 we briefly describe the methodology we followed for developing an ontologically grounded conceptual framework; in section 3 we present the proposed upper level ontology, Section 4 includes conclusions and future directions.

2. Methodology

Our scenario involves communities distributed geographically and for the domain analysis and knowledge acquisition phases the group has used different independent technologies that were not always integrated into the Protégé suite (Noy *et al.*, 2003). From these experiences members of RSBI are also working with others on a collaborative and knowledge acquisition tool for the development of ontologies integrated in Protégé (Garcia *et al*., 2005).

Figure 1 schematizes the methodology we followed. We built different models throughout our analyses of available knowledge sources and information gathered in previous steps. Firstly, a "baseline ontology" was gathered, *i.e.* a draft version containing few but seminal elements of an ontology. Typically, the most important concepts and relations were identified somewhat informally. We could assimilate this "baseline ontology" into a taxonomy, in the sense of a structure of categories and classifications. We consider a taxonomy as "a controlled vocabulary which is arranged in a concept hierarchy", and ontology as "a taxonomy where the meaning of each concept is defined by specifying properties, relations to other concepts, and axioms narrowing down the interpretation." As the process of domain analysis and knowledge acquisition evolves, the taxonomy takes the shape of an ontology. During this step, the ontologist worked primarily with only very few of the domains experts; the others were involved in weekly meetings. In this phase the ontologist sought to provide the means by which the domain experts he or she was working with could express their knowledge. Some deficiencies in the available technology were identified, and for the most part were overcome by our use of conceptual maps (CMs).

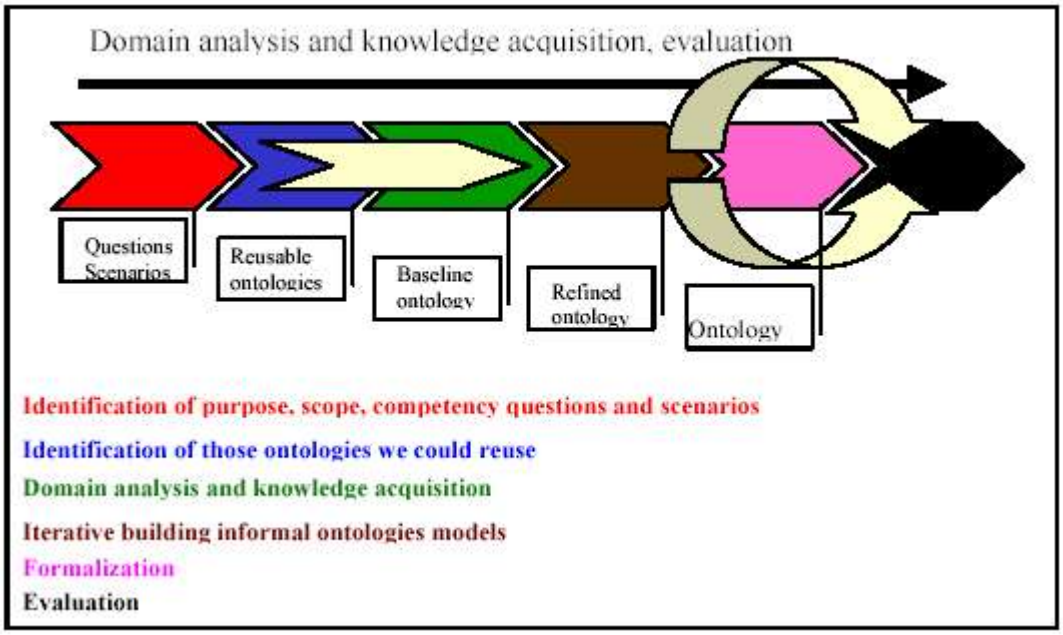

Fig. 1. Our methodology

3. The RSBI Semantic Framework

Our approach is one of an upper ontology that would provide high-level semantics for the representation of omics-based investigations that serves as a conceptual scaffold from which other ontologies may be hooked. An example for the latter could be an ontology specific for the microarray technology, such as the MGED Ontology, and/or specific for an applications, such as toxicology. In order to describe the interaction of different technologies during the course of a scientific endeavour we considered there was the need for a high-level container where to place the information relevant to the biology as well as that relevant to those different assays. Our high-level concept is an *Investigation*, a self-contained contained unit of scientific enquiry, containing information for *Study(s)* and *Assay(s)*. We consider a *Study* to be the set of steps and descriptions performed on the *Subject(s)*. In the cases where the *Subject* is a piece of tissue, and no steps have been performed but just an *Assay* has been carried out, then we the *Study* contains only the descriptors of the *Subject* (e.g. provenance, treatments, storage etc). We consider an *Assay* as the container for the test(s) performed and the data produced for computational purpose. There are different *AssayType(s)* and the different omics technologies fall within this category. A view of the RSBI upper ontology is shown in Figure 2 and the ontology is available from the RSBI webpage.

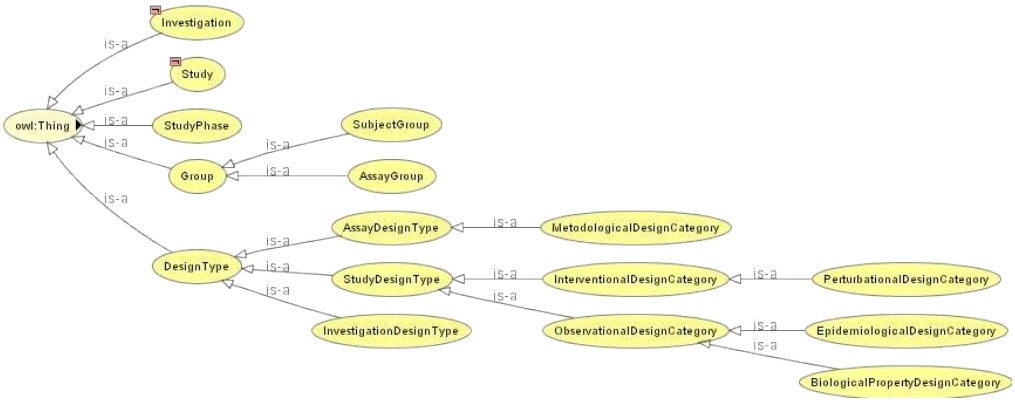

Fig. 2. A view of a section of the RSBI ontology.

4. Conclusions and Future Directions

Since our framework will allow the use of different ontologies the definition for

whole/part relationships should be consistent across those different ontologies. However, currently there are no standards of guidance for defining whole/part of relationships, adding another layer of complexity when developing an upper level ontology. Ultimately the RSBI upper level ontology should be able to answer a few questions and position almost anything approximately in the right place, even if the spot has a non-existent ontology. The relationship between *Study* an *Assay* defines an *Investigation*, different things participate in different processes and on the same token some things retain their form over time. *Study* and *Assay* contain information about those processes. It is particularly important to have minimal commitment when developing upper level ontologies, only those concepts providing a common scaffold should be considered.

Formalizing knowledge to facilitate data representation is not a trivial task and should be noted that this effort is work in progress. As next step, we plan to formalize our ontology, and validate it against more complex uses cases.

5. References

Brazma, A., Hingamp, P., Quackenbush *et al*. 2001. Minimum information about a microarray experiment (MIAME)- toward standards for microarray data. *Nat Genet.* 29 (4): 365-71.

Garcia Castro, A., Sansone S.A., Rocca-Serra, P., Taylor, C., Ragan, M.A. 2005. The use of conceptual maps for two ontology developments: nutrigenomics, and a management system for genealogies. *Proceedings of the 8th International Protege Conference*. (Accepted for Publication)

HUPO PSI: http://psidev.sourceforge.net

MGED Ontology: http://mged.sourceforge.net/ontologies/index.php MGED RSBI: http://www.mged.org/Workgroups/rsbi

Noy, N.F., Crubezy, M., Fergerson, R.W. *et al.* 2003. Protege-2000: an open-source ontology-development and knowledge- acquisition environment, *AMIA Annu Symp Proc*, 953.

OECD Toxicogenomics Guidelines:
http://www.oecd.org/document/29/0,2340,en_2649_34377_34704669_1_1_1_1,00.html

Quackenbush J. 2004. Data standards for 'omic' science.*Nat Biotechnol.* 22:613-614.

Sansone, S.A, Morrison, N., Rocca-Serra, P., Fostel, J. 2005. Standardization initiatives in the (eco) toxicogenomics domain: a review. *Comp. Funct. Genomics*. 8, 633-641. SMRS: http://www.smrsgroup.org