Metabolomics Society - Metabolomics Standards Initiative (MSI)

Ontology Working Group (OWG) road map

http://www.metabolomicssociety.org/mstandards.html

This document describes the purpose and the working strategy of the Metabolomics Society Ontology Working Group (OWG) in an effort to reach a broad consensus in the community on the semantics required to report metabolomics experiments.

The Metabolomics Society standards initiative

The Metabolomics Society has appointed an Oversight Committee to monitor, coordinate and review the efforts of working groups (WGs) in specialist areas that will examine standardization and make recommendations. The five WGs, some of which are divided into further subgroups are listed here:

- Biological context metadata WG
- Chemical analysis WG
- Data processing WG
- Ontology WG
- Data exchange WG

The structure of the WGs thus follows the general "workflow" model in metabolomics: from a description of the study design to sample workup, data acquisition, processing and export, bound together by controlled vocabularies and relationships between the terms used.

OWG statement of purpose

The Ontology Working Group (OWG) seeks to facilitate the consistent annotation of metabolomics experiments by developing an ontology to enable the broader scientific community to understand, interpret and integrate data. This will be valuable resource not only for the groups involved in the WG but also for the metabolomics-user community at large, allowing for the consistent semantic understanding of and data across disparate sources (software and databases, private and public).

Operating plan

The OWG will tackle the semantics issue by:

1. Reaching a consensus on a core set of controlled vocabularies (CVs) and
2. Developing a corresponding ontology.

Specifically the CVs and ontology will aim to

- Provide a consensus set of descriptors for the consistent semantic representation of and data across disparate resources (software and databases, both private and public).
- Assist to model the design of an investigation, the protocols and instrumentation used, the data generated and the types of analyses performed on it.

The developmental process will require the following groups of people to provide input:

- The OWG members as developers of the CVs and ontology;
- Ontology experts/knowledge engineers to provide advice about the engineering of the ontology;
- Metabolomics practitioners to provide use cases, validate the CVs and ontology produced and advise on additional terms to be included into the ontology.

Operating principles

The OWG will seek to represent the diverse community of metabolomics users in an unbiased and open fashion. The group will integrate and harmonize with other WGs within the standardization initiative. Communications will be frequent, respectful but candid and widely distributed. Every effort will be made to meet group goals in a timely fashion, although no central fund exits for this initiative and the members participate on a volunteer base.

To achieve these goals the OWG will:

- Work cooperatively, maintain a mailing list and a website with the names of participating members to remain approachable, inclusive and transparent while the size of the group and the complexity of the tasks increase.
- Produce and maintain a set of documents - which are either common practice descriptions, or recommendations- to ensure that the statements from this group are clear, accurate and accessible.
- Leverage on previous and relevant work in other omics studies, and recent metabolomics standardization efforts.

- Represent the metabolomics domain within a larger, international effort developing an ontology for functional genomics experiments.
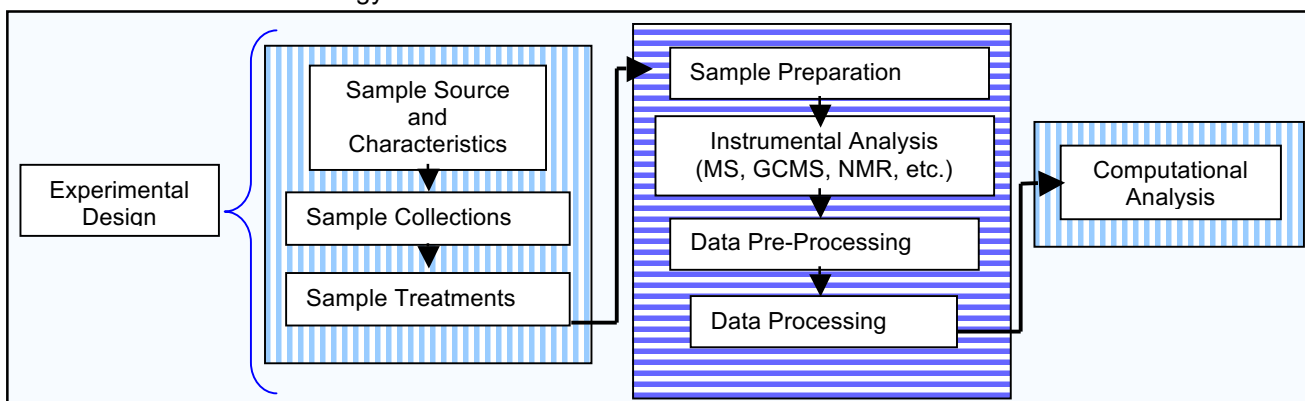
Phase 1 – Consensus on CVs

The first phase focuses on developing CVs master lists, representing the consensus set of descriptors for the consistent semantic representation of the experimental workflow and the data across disparate resources (software and databases, both private and public).

CVs coverage

The OWG has divided the CVs coverage into two main components. The figure below shows the technology- dependant components in the centre (horizontal lines) and the general experimental components on the sides (vertical lines). Conforming to a generally accepted view that duplication and incompatibility should be avoided, the development of CVs (and ontology) for the general experimental components should be coordinated with standardization initiatives in other omics domain, such as the HUPO Proteomics Standards Initiatives (PSI) and the Microarray Gene Expression Data (MGED) Society, as part of an ontology for functional genomics investigations (see Phase 2 section below).

Every effort will be made to cover as many components as possible. The CVs for the instrument-dependant components, however, will be the primary focus of this OWG, starting from NMR sub-component. For MS sub-component the OWG will leverage on the previous work by PSI MS Ontology WG. The chromatography, also shared by proteomics and metabolomics domains, will be developed in close collaboration with PSI Ontology WG.

Sources of terms

The OWG will reach out, evaluate and leverage previous and relevant work done, including:
- Collaborative Computing Project for the NMR Community: http://www.ccpn.ac.uk/datamodel/datamodel.html
- PSI-MS: Mass Spectrometry Standards WG: http://psidev.sourceforge.net/ms/
- NMR-STAR web page: http://www.bmrb.wisc.edu/
- ArMet model: http://www.armet.org/
- MeMo http://dbkweb.ch.umist.ac.uk/memo/
- HUSERMET project: http://www.metabolomics.co.uk/
- MeT-RO: http://www.metabolomics.bbsrc.ac.uk/MeT-RO.htm
- IUPAC terminology for analytical chemistry
- Human Metabolome Project (HMP): http://www.metabolomics.ca
- UMLS http://www.nlm.nih.gov/research/umls/
- KEGG http://www.genome.ad.jp/dbget-bin/www_bfind?compound
- ChEBI http://www.ebi.ac.uk/chebi/

Naming conventions

At present, neither unified naming conventions, nor common recommendations have been agreed upon by the ontology-oriented communities. This group will propose good practice for naming Knowledge Representation (KR), so that the lists of CVs collected are consistent locally at the syntactic level.. The naming conventions will be shared with other communities working towards an ontology for functional genomics investigations (see Phase 2 section below).

Use cases and CVs master list

CVs master list for each sub-component will be created. Compiling such master lists should be an iterative process and the proposed steps are:

1. Compile a list of use cases for the application with the help of practitioners in the relevant metabolomics area. Use cases will highlight usage scenarios, goals and requirements.
2. Take a list of terms for a sub-component from a certain resource and use it as a basis; make it compliant to the naming conventions, identify synonyms and definitions for each term. Keep track of the relationships between the terms (if provided) for the second phase (ontology development).
3. Discuss the CVs in the OWG, then circulate the CVs to the practitioners in the relevant metabolomics area. This will ensure that the lists are as complete as possible, that we get valid definitions and will aid the ontology construction later on.
4. Once a general agreement has been reached on the initial CVs, further vocabularies or ontologies will be processed in turn by deciding which of their terms should be incorporated into the initial CV. For each of these terms synonyms, definitions, relationships will be identified as before.
5. When all sources of relevant vocabularies and ontologies for a certain sub-component have been exhausted, it will be determined which concepts/domains remain to be covered. We will need to actively collaborate with metabolomics practitioners at this stage to ensure the quality and completeness of the proposed vocabulary.

Phase 2 – Ontology development

The OWG's ultimate goal is to combine the CVs master lists, add structure and create an ontology. Ontology-based knowledge representations have proved to be successful in providing the formal semantics for standardised annotation, integration and exchange of data.

The larger scientific community will best be served if the resulting ontology overcomes duplications across omics domains where commonality of the concepts exists, like in the general experimental component (see figure above). To achieve this goal, the OWG will engage in the Functional Genomics investigation Ontology (FuGO, http://fugo.sourceforge.net) project, a wider collaborative effort bringing together HUPO-PSI, MGED Society and other communities. The MSI OWG will represent the metabolomics domain in this collaborative effort.

FuGO will be developed using Protégé, a freely available tool, and in Ontology Web Language (OWL) format. Ultimately FuGO aims to be developed as a Foundry (core) Ontology under the Open BioMedical Ontology umbrella (OBO, http://obo.sourceforge.net).

Each participating community will develop its technology- dependant components by using the relevant FuGO "leaf nodes" (e.g. *Instrument*) as top-level classes. The MSI OWG will post and maintain its ontological components under the OBO umbrella, in anticipation of FuGO to be completed. A first draft of FuGO will be considered 'completed' when the general (common) experimental component and all the technology- dependant components have been developed and harmonized (redundancy removed).

Every effort will be made to develop FuGO in a in a timely fashion It should be noted, however, that no central fund exits for this project and the members of the community involved participate on a volunteer base.

MSI OWG members

Andy Tseng (Human Serum Metabolome Project, Un of Manchester, UK)

Bertram Ludaescher (UCDavis Genome Center, US)

Chris Taylor (EBI and HUPO-PSI)

Daniel Schober (EBI)

Helen Jenkins (Un of Wales, Aberystwyth, UK)

Irena Spasic (Center for Integrative Systems Biology, Manchester, UK)

Larissa Soldatova (University of Wales, Aberystwyth, UK)

Lori Querengasser (Un of Alberta)

Martin Scholz (UCDavis Genome Center, US)

Matej Oresic (VTT Technical Research Centre of Finland)

Oliver Fiehn (UCDavis Genome Center, US), MSI Oversight Committee Chair

Philippe Rocca-Serra (EBI and MGED Society)

Susanna-Assunta Sansone (EBI and MGED Society), MSI OWG Chair

The OWG's expertise covers the following area:

- Machine learning, knowledge representation and ontology
- Modelling, software development, data analysis and text mining
- Computational Biology, Systems Biology, Plant Metabolomics and Clinical Metabolomics