Position Paper for W3C Workshop on Semantic Web for Life Sciences

MGED Ontology

Chris Stoeckert, Helen Parkinson, Trish Whetzel, Paul Spellman, Catherine A. Ball, Joseph White, John Matese, Liju Fan*, Gilberto Fragoso, Mervi Heiskanen, Susanna Sansone, Helen Causton, Laurence Game, Chris Taylor

*Presenter: Liju Fan, KEVRIC - An IMC Company, Silver Spring, MD. lfan@kevric.com

Introduction

The Microarray Gene Expression Data (MGED) Society is an international organization that aims to facilitate the sharing of microarray data generated by functional genomics and proteomics experiments. The current focus is on establishing standards for microarray data annotation and exchange, facilitating the creation of microarray databases and related software implementing these standards, and promoting the sharing of high quality, well annotated data within the life sciences community. A long-term goal for the future is to extend the mission to other functional genomics and proteomics high throughput technologies.

The Microarray Gene Expression Database Group and OMG Gene Expression Standard (MAGE) is the available specification for gene expression at the Object Management Group (OMG). A number of implementations have already been developed, including those by Affymetrix, the EBI, TIGR, The University of Pennsylvania etc. MAGE aims to provide a standard for the representation of microarray expression data that would facilitate the exchange of microarray information between different data systems. Currently, this is done through the OMG by the establishment of a data exchange model (MAGE-OM: Microarray Gene Expression - Object Model) and data exchange format (MAGE-ML: Microarray Gene Expression - Markup Language) for microarray expression experiments.

The Minimum Information About a Microarray Experiment (MIAME) is a standard developed by the MGED Society to enable the interpretation of the results of the experiment unambiguously and potentially to reproduce the experiment. The MGED Ontology (MO) (http://mged.sourceforge.net/ontologies/MGEDontology.php), developed by the MGED Ontology Working Group, is closely linked to MIAME and compliance to the MO enables partly automatic validation of MIAME compliance for documents in MAGE-ML format. The MO provides standard terms for the annotation of microarray experiments to facilitate unambiguous interpretation of microarray annotations, as well as queries and searches in microarray databases.

Future development of the MGED Ontology

The MO has matured to the point of a code-stable version (v1.1.8) that is tied to (a code-stable version of) the MAGE-OM (v1.1). In both the near-term and long-term future plans for the MO, we envision a number of changes that fall into 5 categories: "ongoing maintenance", "ontology language", "non-array technologies", "biological domain extensions", and "MO v2 development".

1. Ongoing Maintenance

This category encompasses additions of new instance terms to existing classes, fixing typographical errors, and adding missing associations. In general, this category represents minor changes that should largely not affect software applications that are based on the MO.

2. Ontology Language

This category encompasses planned changes in the primary language format (from DAML to OWL) and ontology editing tool (from OilEd to Protégé). As with the first category, these should represent fairly minor differences as far as applications based on the MO are concerned. Some minor name changes will be needed to adjust for differences in allowed characters. New functionalities such as the availability of synonyms may be used to enrich the MO further.

3. Non-Array Technologies

The initial scope of MO and the MAGE-OM that MO is tied to was to cover microarray experiments. Standards efforts in "non-array technologies" such as proteomics would like to use the MO and add necessary terms for their specific needs.

The MO has two major branches: a core ontology (MGEDCoreOntology) that is stable and tied to MAGE and an extended ontology (MGEDExtendedOntology) that is free from these restrictions. Classes that are needed for new technologies can be placed under the MGEDExtendedOntology and linked to MGEDCoreOntology classes through relationships (i.e., MGEDExtendedOntologyClass has_property (MGEDCoreOntologyClass). Such development would not impact the MGEDCoreOntology and therefore allow addition of non-array technology classes at a point during the MO development. Instances that are needed for new technologies may be most appropriate for existing classes in the MGEDCoreOntology. The policy for adding and defining instances regarding technology-related terms is to provide a generic name and definition but to supply technology-specific examples (in the definition). Thus, for terms needed for new technologies, when possible, these can either be generalizations of existing terms or addition of new general terms. In those cases where technology-specific non-array terms are needed and cannot be generalized, these will be deemed

inappropriate for MGEDCoreOntology classes and new MGEDExtendedOntology class created for it.

4. Biological Domain Extensions

There are "biological domain extensions" needed for areas such as toxicogenomics where the current specification of experimental design and sample description is not sufficient to fully capture descriptions of experiments. Extensions to Experiment and Biomaterial necessitated by specific biological domains such as toxicogenomics should fit within the MAGE-OM v1.1 and so ultimately could go into the MGEDCoreOntology. However, as the new classes, subclasses, properties, and instances are under development (and therefore not stable), they would not be initially appropriate for direct inclusion in the MGEDCoreOntology. Therefore, the needed extensions should be placed in the MGEDExtendedOntology until mature enough to be migrated over to the MGEDCoreOntology. A complication to consider is when the most appropriate way to extend the MO is to add a subclass to a MGEDCoreOntology class. The subclass will therefore by inheritance also be in the MGEDCoreOntology. This may be acceptable if software applications based on the MO can easily recognize these and ignore them if desired. One mechanism might be to place the biological domain extension subclasses under the MGEDExtendedOntology and therefore it will belong to both MO branches (Core and Extended). Alternatively, such subclasses can remain wholly in the MGEDCoreOntology but named to indicate that it is under development.

5. MO v2 Development

MAGE v2 will have major structural changes from MAGE v1,1 and is likely to require major changes in the MO. With a MO v2 developed in parallel, this should not conflict with the stated plans of the MO to be consistent with MAGE as it will be tied to the new version.

Summary

The future development of the MO will strive to maintain the stated policy to be consistent with a version of the MAGE-OM and to be stable. Incremental changes will be allowed to flesh out and fix the MO. These changes will include those needed to move development of the MO to Protégé and OWL. In order to facilitate a wide dissemination of the MO and its potential utilization by non-array technologies and biological domain extensions, the MO will begin publishing in OWL. The changes will also include generalizations of terms to allow applicability to these non-array technologies where possible. Major changes to the MO are anticipated for MAGE-OM v2, but these will be restricted to the MO v2 and tied to the new MAGE version. With the above changes the Life Sciences Identifiers Specification (LSID) that was approved by the OMG will be explored in the MO. (http://www.omg.org/cgi-bin/doc?lifesci/2003-12-02)