



Sixth Framework Programme for Quality of Life  
and Management of Living Resources

Project no. LSHG-CT-2006-037686

## **EMERALD**

Empowering the Microarray-Based European  
Research Area to Take a Lead in Development and  
Exploitation

EU Deliverable:

D3.6, D3.7

Due Date:

D3.6 October 2007

D3.7 December 2007

Delivery Date:

D3.6 October 2007

D3.7 December 2007

Version 1.0

Partner responsible: EBI

Author: James Malone, Helen Parkinson

## Contents

Work Package 3 .....	3
3.6.3 Workshop Agenda .....	3
D3.7 Data Transformation Ontology Workshop Report .....	4
3.7.1. Introduction .....	4
3.7.2 Background.....	4
3.7.3 Workshop Report.....	5
3.7.3.2 Cooperation with OBI Consortium.....	6
3.7.3.3 Reuse of existing ontologies .....	6
3.7.3.4 Defining and placing concepts .....	7
3.7.3.6 Evaluating the ontology.....	10
3.7.3.7 Future Plans.....	10
References .....	11

## **Work Package 3**

### **Deliverable D3.6**

#### **Data Transformation Workshop**

**Due Date: October 2007**

**Delivery Date: November 2007**

#### **3.6.1 Introduction**

The workshop took place at the EBI in Hinxton between the 5<sup>th</sup> and 9<sup>th</sup> November, 2007. Delegates were invited from a diverse range of backgrounds in order to gather a comprehensive range of perspectives and input from I users of a NTO and those with skills and experience in ontology development. These included biologists, biochemists, statisticians, computer scientists, ontologists and experts in microarray data analysis, data management and ontology application. The attendees, their affiliations and the agenda are provided below, the outcome and summary of the workshop are provided in the workshop report: deliverable 3.7. Complete proceedings of the workshop are available at: [https://wiki.cbil.upenn.edu/obiwiki/index.php/Data\\_transformation\\_workshop](https://wiki.cbil.upenn.edu/obiwiki/index.php/Data_transformation_workshop)

#### **3.6.2 Attendees**

- Helen Parkinson (chair), EBI, UK
- James Malone (co-chair), EBI, UK
- Tina Boussard, Stanford University, USA
- Christian Cocos, University of Saarland, Germany
- Melanie Courtot, BCCRC, Canada
- Elisabetta Manduchi, University of Pennsylvania, USA
- Monnie McGee, Southern Methodist University (SMU, Dallas), USA
- Richard Scheuerman, Texas Southwestern University, USA
- Daniel Schober, EBI, UK
- Robert Stevens, University of Manchester, UK
- Daniel Rubin, Stanford University, USA – by teleconference

#### **3.6.3 Workshop Agenda**

- Determining scope of ontology – determining areas of focus for the NTO developed as part of the EMERALD grant
- NTO context – identification of other relevant ontology development efforts and strategies for community development
- Identification of areas of commonality with other ontologies – e. g reuse of existing ontologies such as the newly developing Software Ontology
- Developing the ontology content - defining concepts and placing them into an ontological hierarchy
- Developing design principles for the NTO
- Evaluating the ontology

## **D3.7 Data Transformation Ontology Workshop Report**

### **Deliverable D3.7**

#### **Data Transformation Workshop**

**Due Date: December 2007**

**Delivery Date: December 2007**

### **3.7.1. Introduction**

The aim of the Quality Metric and Ontologies work package (WP1) is to develop and disseminate quality metrics and tools for determining data quality and communicating data transformations. Specifically D3.7 specifies the development of a 'normalisation and transformation ontology' (NTO) required for describing microarray experiments and analysis. The data transformation ontology invitation only workshop was organized to facilitate the development of the ontology during a week-long face-to-face meeting with community experts and members of the Emerald consortium. This report details the outcome of this workshop and highlights the next step towards the development of the NTO.

### **3.7.2 Background**

Fundamental to the successful analysis and reproducibility of microarray experiments is the quality of the documentation and descriptions that are used to report microarray experiments. It is because of the value of standardised reporting that initiatives such as MIAME (Minimum Information About a Microarray Experiment) (Brazma *et al*, 2001) and MAGE-TAB (MicroArray Gene Expression Tabular) (Rayner *et al*, 2005) emerged and were rapidly adopted by the microarray community. It is essential that it is possible to report the transformations that are performed upon data in an unambiguous and consistent manner so that others can reproduce these experiments and analyses reported in a publication. This form of representation would go some way towards increasing the relevance and utility of public microarray data archives. The emergence of workflow technologies in bioinformatics such as Taverna (Hull *et al*, 2006) whereby workflows are shared and can be re-used also requires the use of ontologies to describe the processes of data manipulation and transformation.

There is a movement towards formal knowledge representation in the biomedical domain. The work of the OBO (Open Biomedical Ontologies) Foundry has helped to enable those wishing to use formal knowledge representation techniques to model their data using ontology standard representations such as OBO and OWL (Web Ontology Language). Ontologies offer the advantages of shared understanding, reduced ambiguity and richer representations of data. They also offer an increased degree of machine readability, allowing computation to be performed over models which use the ontology. This can enable, for instance, consistency checking across the model (i.e. ensuring that classes and instances have attributes which conform to the knowledge model (Yeh *et al*, 2003)), querying of complex relationships within large data sets and can facilitate automatic inference of new knowledge (Bada and Altman, 2000).

The diversity in microarray experiment designs and applications requires that a large number of pre-processing approaches are available. In order to facilitate unambiguous and consistent descriptions of experimental data transformation the development of a 'normalisation and transformation ontology' (NTO) is required. This will provide a means to conceptualise and classify the approaches used, describe relationships between these concepts and store these in a machine readable form.

### 3.7.3 Workshop Report

#### 3.7.3.1 Setting scope of the ontology

Standard ontology development methodology to determine scope and requirements of an ontology is through the use of competency questions and use cases. After identifying the intuitive main scenarios envisaged for an ontology, a set of natural language questions, called competency questions are used to extract the main concepts and their properties, relations and axioms of the ontology (Gómez-Pérez *et al*, 2004). This process is characterized by the following:

1. Identification of motivating scenarios relating to the applications that will use the ontology. These scenarios describe a set of requirements the ontology should be able to satisfy.
2. Identify a set of elaborate informal competency questions written in natural language which are required to be answered by the completed ontology. Such questions form a requirement specification and can be used to evaluate whether or not the ontology is complete.
3. Specify terminology using the competency questions. This concerns taking the informal competency questions and formally representing the concepts, attributes and relationships in a language (such as OWL).
4. Write competency questions in a formal manner using formal terminology.
5. Specify definitions of terms using axioms and constraints.

Competency questions and use cases for data transformation were requested from members of the biomedical community. An example of the types of competency questions submitted is shown below. They are worded as questions which we would expect the ontology to be able to answer.

- Which genes have a 2 fold change in expression where MAS5 has been applied as a data transformation methodology?
- Which pre-processed microarray data expresses values as log ratios (of two conditions) for a specified logarithmic base?

An example of a textual use case is shown below. These are generally more detailed and are given from an actors point of view in terms of how they would interact with the ontology:

- An experimenter has conducted an expression microarray experiment involving two conditions with replicate assays per condition, where they have both biological and technical replicates. They are running two kinds of differential expression analyses: (a) one at the gene level and (b) one at the gene set level. In (a) the aim is to identify differential expressed genes (e.g. via algorithms like PaGE and SAM). In (b) the aim is to identify, from an a priori given collection of gene sets (e.g. user provided,

or based upon GO annotation), which of these sets are differentially expressed as a whole (e.g. via algorithms like GSEA or SAM-GSA). Before running the analyses the data is preprocessed with the following data transformation series: (i) filter out flagged reporters, (ii) normalize the individual assays, (iii) average across technical replicates (but not across biological replicates). The above steps all requires annotation using the ontology.

Further use cases and competency questions are being sought from the community and will be included in future NTO development.

### **3.7.3.2 Cooperation with OBI Consortium**

The initial strategy of this project was for the project consortium to work closely with the MGED consortium to 'develop a component of the MGED Ontology that can be used to describe data transformations.' However, the MGED Ontology has since been subsumed into the Ontology for Biomedical Investigations (OBI) (<http://obi.sourceforge.net/>) which has become the focus of efforts for many previously working on the MGED Ontology. The OBI project has the much wider scope of being able to represent any (not just microarray) biological and medical experiment and investigation. The ontology aims to model the design of an investigation, the protocols and instrumentation used, the material used, the data generated and the type analysis performed on it.

Although the focus of OBI is broader than the MGED Ontology, domain specific concepts still play a major role in the gathering of terms from the different biomedical communities, of which transcriptomics is an important part. For this reason, coordination and cooperation with OBI Consortium is inline with the NTO aims. OBI is an OBO foundry ontology and has data transformation terms within its scope. This means that any separate effort would likely not be accepted as an OBO foundry ontology as efforts should be orthogonal. Moreover, the OBI Consortium benefits from a diversity of membership as well as an existing infrastructure framework which would be compatible with the NTO. Specifically, the 'data transformation' branch (a branch in OBI being the equivalent to a sub-section of the overall ontology with related content) of the OBI development structure fits well with the NTO scope (<https://wiki.cbil.upenn.edu/obiwiki/index.php/DataTransformation>). The representation format of OBI is OWL and is therefore compatible with the skills of the personnel employed on EMERALD and is our preferred format, therefore no additional overhead is required to participate in the OBI community. A decision has been taken to develop the NTO as part of OBI, specifically the Data Transformation branch, rather than the now defunct MGED ontology as previously described.

### **3.7.3.3 Reuse of existing ontologies**

There are several ontologies currently in existence which may be reused, either by incorporation into the NTO or by a mapping to the ontology as an external resource. One such example is the Software Ontology (SO) current under development by Daniel Rubin (<http://bioontology.org/projects/ontologies/SoftwareOntology/>). Following a discussion during the workshop and subsequent evaluation the ontology is currently insufficiently mature in terms of microarray specific terms for mapping. However, it is clear that the various microarray software types should be

described either in the NTO or the software ontology and we will cooperate with the software ontology effort to ensure that this happens.

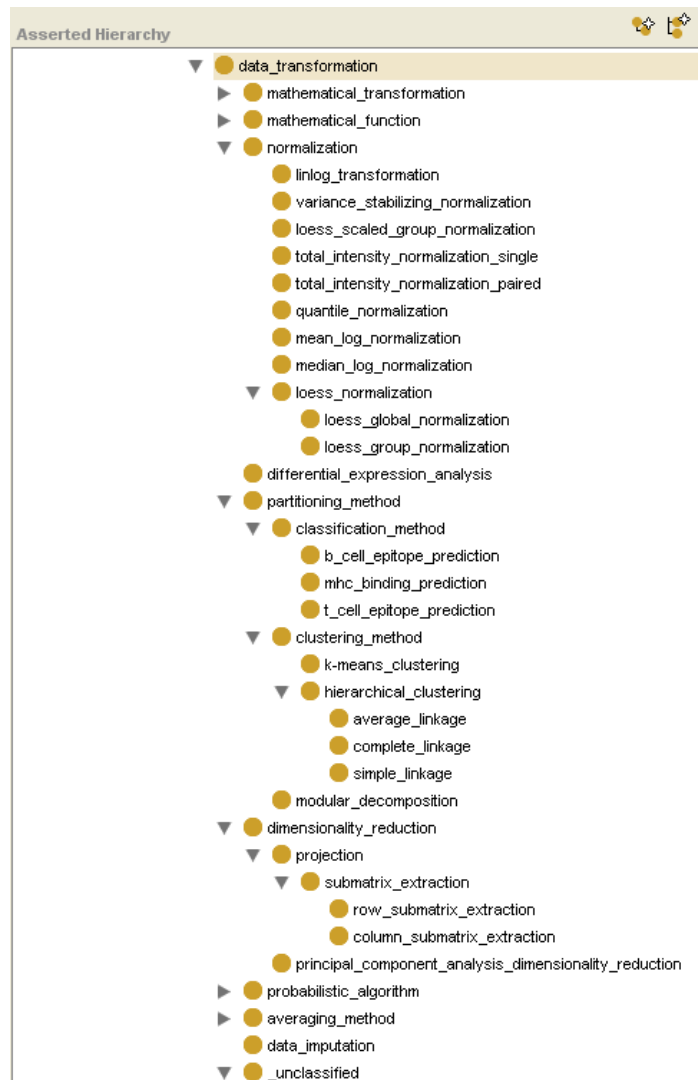
An upper level ontology is needed to map between existing ontologies and to provide high level structure. OBI has evaluated existing upper level ontologies including BFO (Smith *et al*, 2005) and SUMO (Niles and Pease, 2001). BFO is in use by several OBO foundry ontologies and BFO has been selected for use by OBI and is compatible with the NTO.

#### **3.7.3.4 Defining and placing concepts**

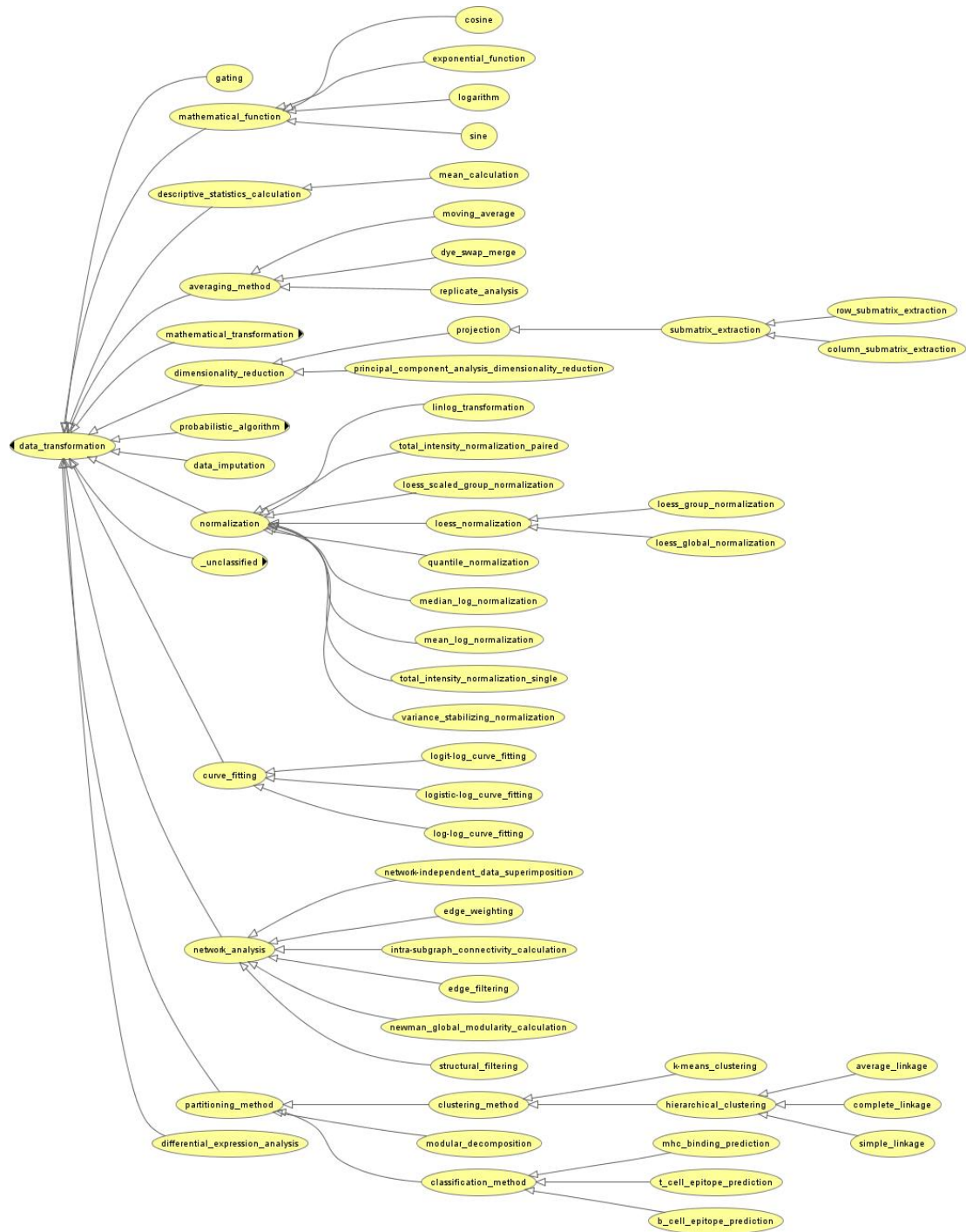
A draft version of the NTO was available prior to the workshop consisting of terms submitted by the community, gleaned from other ontologies e.g. MGED ontology and obtained from literature searching of key papers. The terms, their definitions and the existing structure of the ontology was examined and a new hierarchy developed according to design principles developed during the workshop.

New terms were identified using a card sorting technique for statistical methods, data transformation software, mathematical methods and these were added to the ontology.

Below is the hierarchy after progress was made and an implementation was produced in an owl ontology file. The view is from the ontology editor Protégé.



The view below gives an alternative view of the hierarchy using the OWL Viz visualization tool plugin for Protégé.



The latest SVN version of the ontology can be found at the following link  
<http://obi.svn.sourceforge.net/viewvc/obi/branchDevelopment/trunk/DataTransformation.owl?view=log>.

### 3.7.3.5 Design Principles

There were several design principles that emerged as a result of the discussion and development undertaken during the workshop.

- Many of the data transformation terms describe processes that are application specific. We have tried to identify the generic equivalent of the application specific term, to use this term as an upper level node in the hierarchy and then allow the use of the application-specific term as a child/sub-concept. This allows the re-use of the generic term in other applications while supporting dataset annotation using the application-specific terms.
- Avoiding multiple parentage using OBI concept of **roles**. Multiple inheritance can cause complications in terms of maintainability and conflicts within the model (e.g. altering one parent which causes a conflict with the other parent of a child concept). For this reason it is useful to avoid this practice wherever possible. Roles are an OBI concept used to give a concept, via a relationship (e.g. has\_role), a particular usage depending upon the context they are being used in. For example, concept of dye-swap merge which can play role of normalization and averaging (i.e. dye-swap merge for normalization and dye swap merge for averaging).
- Use of **qualities**, another OBI concept, to describe inherent properties/characteristics in other concepts, i.e. a test having the quality of being 'parametric' or 'non-parametric'.
- Policy related to other OBI branches of **plan** and **application of plan**. If we create classes in only one of these, do it in plan, not application, as we have qualities that need to inhere in something. According to BFO they inhere in independent continuant, i.e. plan.

### 3.7.3.6 Evaluating the ontology

The following strategy was developed during the workshop and will be used to evaluate the NTO:

1. Competency Questions – Using the competency questions initially created to assess coverage of the ontology
2. Use Cases – More complex evaluation which evaluates usage from a particular user (actors) point of view and a complete case of how they may interact with NTO in the application environment.
3. Consistency Checking – Using a reasoner to computationally evaluate the ontology for consistency and to infer hierarchies via the use of restrictions.

This type of strategy is one which is often adopted when evaluating ontologies and is documented in ontology methodologies. The On-To-Knowledge project methodology is one such example (Staab *et al*, 2001). It is suggested in this method, that the evaluation process should consist of checking requirements through competency questions and testing the ontology in the target application environment, similar to that proposed above. Examples of competency questions and use cases can be seen previously in section **3.7.3.1**.

### 3.7.3.7 Future Plans

Following the workshop the ontology will be developed iteratively and the following tasks will be performed

1. Continued collection of core concepts from communities

2. Creation definitions for each concept by experts in respective domains (e.g. Statistics) and placement in hierarchy or definition using restrictions.
3. Coding of Core Concepts into an ontology (OWL) using Protégé ensuring upper ontology compliance
4. Integration with current OBI structure
5. Evaluation using strategy above
6. Iteration using evaluation feedback

## References

Bada MA and Altman RB (2000) Computational modeling of structural experimental data. *Meth. Enzymol.* 317, pp. 470–49.

Brazma A, Hingamp P, Quackenbush J, Sherlock G, Spellman P, Stoeckert C, Aach J, Ansorge W, Ball CA, Causton HC, et al. (2001) Minimum information about a microarray experiment (MIAME) - toward standards for microarray data. *Nat. Genet.* 29, pp.365-371.

Gómez-Pérez A, Fernández-López M, Corcho O (2004) *Ontological Engineering*. Springer-Verlag, London.

Hull D, Wolstencroft K, Stevens R, Goble C, Pocock MR, Li P and Oinn T (2006) Taverna: a tool for building and running workflows of services. *Nucleic Acids Res.*, 34, pp. W729–W732.

Niles I and Pease A (2001) Towards a standard upper ontology. In 2nd International Conference on Formal Ontology in Information Systems (FOIS-2001), Ogunquit, Maine.

Rayner T, Rocca-Serra P, Spellman PT, Causton HC, Farne A, Holloway E, Liu J, Maier DS, Miller M, Petersen K, et al. (2006) A simple spreadsheet-based, MIAME-supportive format for microarray data: MAGE-TAB. *BMC Bioinformatics.*7(489).

Smith B, Kumar A and Bittner T. (2005) Basic formal ontology for bioinformatics. *Journal of Information Systems.*

Staab S, Schnurr HP, Studer R and Sure Y (2001) Knowledge processes and ontologies. *IEEE Intelligent Systems* 16(1), pp. 26-34.

Yeh I, Karp PD, Noy NF, Altman RB. (2003) Knowledge acquisition, consistency checking and concurrency control for Gene Ontology (GO). *Bioinformatics.* 19(2), pp. 241-8.