Building Mouse Phenotype Ontologies

G.V. Gkoutos, E.C.J. Green, A.M. Mallon, J.M. Hancock, and D. Davidson

Pacific Symposium on Biocomputing 9:178-189(2004)

BUILDING MOUSE PHENOTYPE ONTOLOGIES

G. V. GKOUTOS, E. C. J. GREEN, A.M. MALLON, J.M. HANCOCK
MRC UK Mouse Genome Centre and Mammalian Genetics Unit, Harwell, Oxfordshire, England
E-mail: g.gkoutos@har.mrc.ac.uk

D. DAVIDSON
MRC Human Genetics Unit, Edinburgh, England
E-mail: ***Duncan.Davidson@hgu.mrc.ac.uk***

The structured description of mutant phenotypes presents a major conceptual and practical problem. A general model for generating mouse phenotype ontologies that involves combing a variety of different ontologies to better link and describe phenotypes is presented. This model is based on the Phenotype and Trait Ontology schema proposal and incorporates practical limitations and designing solutions in an attempt to model a testbed for the first phenotype ontology constructed in this manner, namely the mouse behavior phenotype ontology. We propose the application of such a model could provide curators with a powerful mechanism of annotation, mining and knowledge representation as well as achieving some level of free text disassociation.

1 Introduction

With the advent of functional genomics, the types and amounts of data that need to be stored in databases has changed both quantitatively and qualitatively. In particular, many types of information that were previously collected on an *ad hoc* basis now need to be stored in a more structured manner. Furthermore, as additional data sets (such as those for gene expression, proteomics and protein-protein interactions) grow in complexity and size, biologists and bioinformaticians are being faced with an increased demand for the construction of queries across these large, diverse datasets. For example, given a gene that was detected to be over-expressed in a microarray experiment it might be of interest to ask whether it was associated with an N-ethyl-N-nitrosourea (ENU) mutant, and whether that ENU mutant had a phenotype that resembled a human disease. It might also be useful to know if the function of the gene, or any homologues, was known, and whether a protein structure for any one of them had been determined.

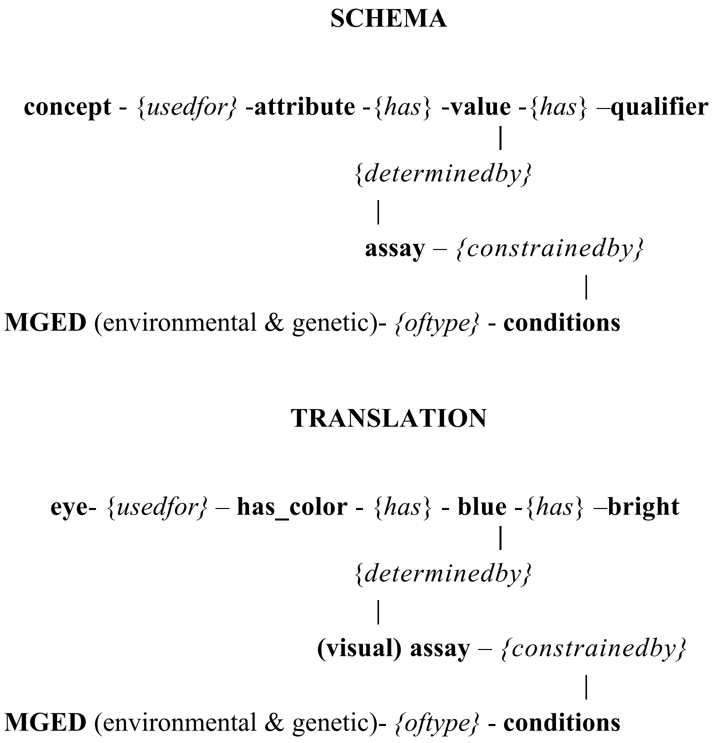
A number of laboratories worldwide are now carrying out detailed analysis of mouse phenotypes that have been generated from the large scale ENU mutagenesis of the mouse genome. Description of mouse phenotypes has not traditionally adhered to pre-defined rules or been recorded in databases but a number of requirements are now driving the developments of such databases, including the requirement to share data from high-throughput screens (such as ENU mutagenesis) and the need to record data in a paperless environment in modern experimental facilities. Here we describe attempts to develop ontologies to aid in the description and mining of mouse phenotype data and make some suggestions, which have more general applications, concerning the ways in which ontologies might be combined to facilitate reasoning with data representing complex domains of knowledge.

2 Mouse Phenotype Ontology

The description of (mutant) phenotypes presents a major conceptual and practical problem. Currently ontological description of phenotypes are mostly linked to individual species databases and have evolved in necessary ad hoc manner. We note here that similar efforts described later are currently being made to describe phenotypic instances in different species [1]. Conceptually, the description of phenotypes requires combinations of orthogonal ontologies with the ability to correlate factors depending on experimental values. Practically, if the data are to be efficiently analysed computationally, then there is a need for consistency between expressions in different phenotypic domains as well as different species. The term "phenotype" can adopt a variety of definitions depending on different fields in biology, and indeed on different researchers in those fields. It may be taken to mean anything from the complete set of phenotypic attributes (traits) that describe an individual to a single phenotypic attribute that distinguishes an individual from other, "normal" individuals. The details of the use of terminology can be divorced from the ontological structures used to describe phenotypic descriptions.

In February 2002, Ashburner proposed a schema (PATO) [1] that could provide a platform of consistent representation of phenotypic data. According to this schema, *"phenotypic data can be represented as qualifications of descriptive nouns or nounal phrases"* [1]. For each noun there will be a set of relative attributes defining a set of appropriate values. The use of these three semantic classes (namely nouns, attributes and values), plus the assays by means of which the phenotypes were determined and the conditions (MGED [2]), both environmental and genetic, under which these assays were performed, will form the basis for the systematic description of phenotype. Figure 1 presents an adaptation of the proposed schema [1], and an example of its application. Its simplicity could provide a common interface upon which to model all phenotype ontologies. Although, from an ontological point of view, its complexity could escalate, this schema provides a firm basis for generating consistent expression of phenotypic data.

Such a schema could provide not only a sophisticated representation of knowledge but also, and perhaps more importantly, an efficient means to annotate and analyse phenotypic data. One can envisage applications that would allow the generation of powerful and advanced ways of searching, retrieving and performing added value mining operations in a particular field and across different domains.



SCHEMA

concept - {*usedfor}* -**attribute** -{*has*} -**value** -{*has*} –**qualifier**

{*determinedby}*

assay – *{constrainedby}*

MGED (environmental & genetic)- *{oftype}* - **conditions**

TRANSLATION

eye- {*usedfor}* – **has_color** - {*has*} - **blue** -{*has*} –**bright**

{*determinedby}*

(visual) assay – *{constrainedby}*

MGED (environmental & genetic)- *{oftype}* - **conditions**

Figure 1. Schema adapted from PATO proposal [1]

The particular domain of our interest, the mouse phenotype ontology, should comprise of at least the following:

- Anatomy – The Anatomical Dictionary for the Adult Mouse [1] has been developed by Terry Hayamizu, Mary Mangan, John Corradi and Martin Ringwald, as part of the Gene Expression, Database (GXD) [3] Project, Mouse Genome Informatics (MGI), The Jackson, Laboratory, Bar Harbor, ME [4]
- Ontogeny – The Anatomical Dictionary for Mouse Development has been developed at the Department of Anatomy, University of Edinburgh, Scotland (Jonathan Bard) and the MRC Human Genetics Unit, Edinburgh (Duncan Davidson) as part of the Edinburgh Mouse Atlas project (EMAP), in collaboration with the Gene Expression (GXD) project at MGI, The Jackson Laboratory, Bar Harbor, ME. Copyright 1998-2002 University of Edinburgh (UK) and MRC (UK). [5]
- Behavior – Parts of Behavior have been expressed in a consistent manner [6,1]
- Pathology – The Pathbase mouse pathology ontology provides a description of mutant and transgenic mouse pathology phenotypes and incorporates 425

known mouse pathologies hierarchically organised as "instances of" pathological processes. [7]

- Gene Ontology – GO describes the roles of gene products and allows genomes to be annotated with a consistent terminology (The Gene Ontology Consortium 2002) [8]
- others ….

These orthogonal ontologies can be combined with PATO to provide phenotypic instances. Generated instances could then be linked to provide individual phenotypes.

Generation of such a combination of ontologies [10] needs to be done collaboratively within the community. Associating concepts with their attributes and values is not an easy task. More often than not, the distinction between these terms is difficult and subjective. Therefore, domain expert knowledge is essential.

We chose to model the behavioral phenotype ontology as a testbed for subsequent parts. We intend to use domain experts' knowledge available to us through EUMORPHIA [9], a European program that we are part of, and collaborate with the Jackson Laboratory [4]. Here, we present our methodology findings, our adaptation schema and raise some modeling issues.

3 Methodology

3.1 Tools Summary

Several tools exist for modeling and building ontologies. Below a small selection is listed, although comprehensive evaluations have been given elsewhere [11, 12, 13, 14].

- **DAG-Edit** [15] provides an interface to browse, query and edit GO or any other vocabulary that has a DAG data structure
- **GKB-Editor** [16] (Generic Knowledge Base Editor) is a tool for graphically browsing and editing knowledge bases across multiple Frame Representation Systems (FRSs) in a uniform manner
- **OilEd** [17] is an ontology editor allowing the user to build ontologies using DAML+OIL
- ***Open*KnoME** [18] is a complete GRAIL knowledge management and ontological engineering environment
- **Protégé-2000** [19, 20] is the most widely known and used tool for creating ontologies and knowledge bases.
- **WonderTools** [21] is an index with the objective of supporting a decision in selecting an ontology-building tool
- **WebOnto** [22] is a Java applet coupled with a customised web server which allows users to browse and edit knowledge models over the web

Since current versions of DAG-edit do not support slots (although a version supporting slots is very close to being released) we have chosen Protégé-2000, which was developed in the Musen Laboratory at Stanford Medical Informatics. Protégé incorporates modeling features such as multiple inheritance, relation hierarchies, meta-classes, constraint axioms and F-Logic. It is written in Java and is well supported with frequent updates and plug-ins for several options (consistency checks, graphical viewing ontology merging etc.). It supports several formats such as RDF(S), XML, RDB, DAML+OIL.

3.2 Knowledge representation languages

A variety of languages can be used for representation of conceptual models, each with different expressiveness, ease of use and computational complexity [23]. Extended comparisons and evaluations have been discussed in detail elsewhere [24], and although the complexity of our current models can be described with existing tools, it should be noted that in the future, upon dealing with more complex phenotype domains, requiring different levels of constraint and expression of relationships varying in complexity, a migration to a finer grained conceptualization will be necessary. Indeed, such approaches have been described in the Gene Ontology Consortium [25] and elsewhere [26].

3.3 Translation of existing ontologies into Protégé-2000

Since most of the ontologies we are planning to use were generated using DAG-edit [18] we had to convert them in the Protégé-2000 format, a frame based system, using the tools [27] written in Java and the methodology as described by Yeh *et al* [27] with minor modifications to the code. Yeh *et al* presented a method for knowledge acquisition, consistency checking and concurrency control for Gene Ontology based on Protégé-2000 [27].

3.4 Metaclasses and Metaslots

We have modeled the converted ontologies such as the anatomy ontology into Protégé-2000 metaclasses, including GO attributes such as name, database references, synonyms and IDs. Protégé allows only is-a relationships to form the class hierarchy so part-of relationships were modeled as slots, as discussed elsewhere [27]. Behavior phenotype ontology slots are described in metaclasses containing fields, such as *Term*, *Documentation*, *Definition*, *Definition Reference*, *ID*, *Associative Ids*, *Synonyms*, *Associative Annotations*, etc. Typical examples of metaclasses are given in reference 27.

The first version of PATO was converted to form the slots for the behavior ontology. As initially conceived, PATO will be updated with attributes required for individual ontologies as appropriate. Protégé allows slot hierarchy (mimicking the PATO hierarchy) with additional information attached such as *Documentation*, *Template values*, *Default*, *Value type*, *Cardinality*, *Minimum and Maximum Values*, *Inverse slots* etc.

It should be noted that care should be taken when new attributes are created. PATO should hold general attributes that can be applied through different phenotypic ontologies and attributes specific to classes should be assigned only when they cannot be modeled with existing options.

3.5 A typical example of implementation

PATO's main advantage is the ability to allow expressions of phenotypic ontologies based on concept relations rather than instances. Using PATO, the ontology can constrain relationships and values for expressing phenotypic instances without the need of assigning the latter. Below is an example of a Behavior class called Feeding Behavior, a subclass of a class named Feeding and Drinking Behavior, (present in both GO [8] and the MGI Mammalian Phenotype Ontology [6]). The example also shows how the Mammalian Phenotype Ontology could possibly be linked to PATO.

Based on this schema one can express a variety of phenotypic data such as preference of cookies versus sausage with a consumption of 40 gr. in a 24 hour period. The ability to interchange use of absolute and relative values combined with different attributes allows the ontology to model and express all possible combinations of phenotypic data for that particular class.

Table 1. A typical example of modeling the Behavior ontology with PATO

CONCEPT	ATTRIBUTE	ASSAY	VALUE
Feeding Behavior	1.attribute:food_type	a. Specialised Diets & Choice Tests	1. chocolate, cheese, cookies vs. sausages
	2.attribute:food_discrimination	b. 24 hour Consumption	2. preference, indifferent,

		aversion, cookies
	3.attribute:food_consumption 3a relative_food_consumption 3b absolute_food_consumption	 3. 3a. increased, aphagia, poluphagia, 3b. 40 gr
	4. attribute_time 4a attribute:relative_time 4b attribute:absolute_time	4. 4a latency 4b 24 hours
Adult feeding behavior	1. *Inherited attributes of class* ***Feeding behavior***	1. abnormal
Preweaning feeding behavior	1. *Inherited attributes of class* ***Feeding behavior*** 2. attribute:suckling_reflex 3. attribute:swallowing_reflex	1. decreased 2. present 3. absent

4 Proposed New Schema

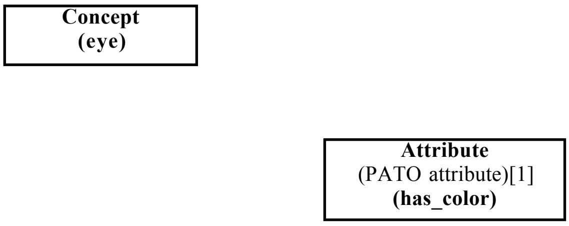
Upon implementation of the schema, we discovered certain modeling and practical limitations. In order to address these, we introduced an alternative version of the schema as presented in Figure 2.

As far as this schema is concerned, a phenotype can be described with the combination of two parts. The phenotypic attribute and the assay.

Phenotype = Phenotypic Attribute + Assay

The phenotypic attribute includes the core ontology concepts plus the associated attributes.

Phenotypic Attribute = Core Concept + Attribute (PATO)



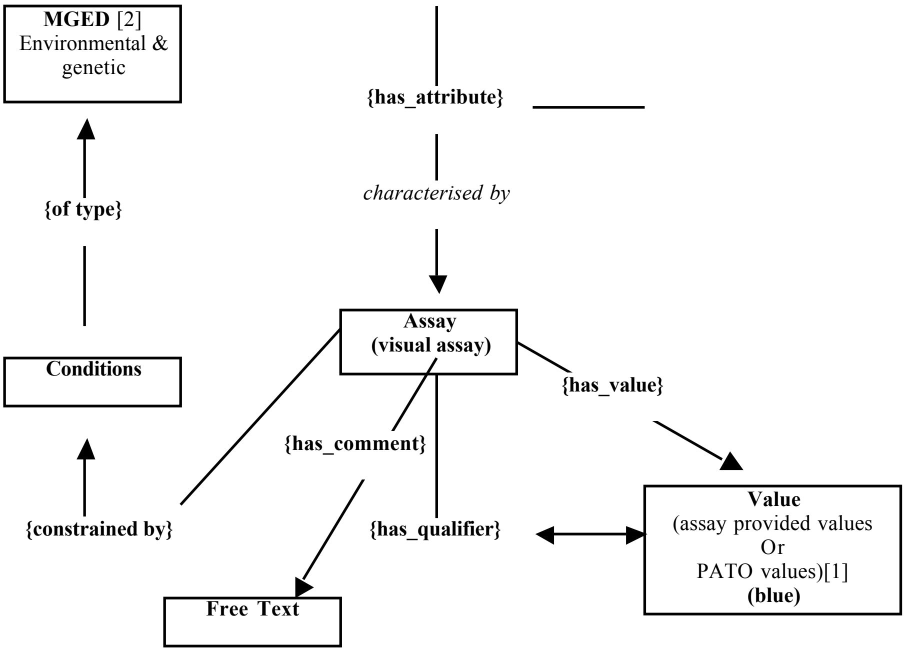


Figure 2. Alternative version of the schema

So in the example of Table 1, the phenotypic attribute would be the class Feeding Behavior plus any associated attributes such as attribute:food_discrimination or attribute:food_consumption etc. In order to reconstruct the phenotype one must take into account the *Assay* that will dictate, control and define any associated values, their units, their definition and the manner they could be assigned to that particular phenotype. In the case of lack of such assays, for example when a phenotype is assigned with visual inspection and no controlled assay is involved, the phenotypic attribute could take its value from a logical assay such as the common values that PATO provides.

In this schema, the *Assay* plays a very important role in controlling the relationship between the attribute, the concept and the values. Beside the practical implementation advantages, discussed in section 5, we note here that it is also conceptually valid, since PATO in itself is a form of a logical assay, as created by its curators. Since in phenotypic knowledge domains, such as the mouse behavior ontology, many values are only speculative interpretations of the assay (i.e. learning and memory assays) it is important for the values to be linked directly to the assays (that describe their interpretation) in order for them to have a sensible meaning. The slot termed *Free Text* is included to capture knowledge that cannot be expressed in the ontology, which is both practical and necessary. This will allow curators to express knowledge, which although it will not be available for advanced computational operations, can still be used via traditional operations, such as free text searching.

5 Discussion

The advantages of such schema are considerable. Firstly, by having the *Assay* ontology to constrain and define the values, these values can be constrained further through the class hierarchy rather than in individual instances. This, as the PATO ontology grows to cover individual domains, would become an important factor in maintenance, consistency, scalability etc. It will also allow us to restrict the values that an attribute can take without asking the data to be input and then referring back to information on the assay to check them.

Furthermore, if the *phenotypic attribute* is disassociated from individual (less common) values, it is not necessary for it to constrain their range, definition, units or general metadata, which is in itself an almost impossible task. This schema also implies that labs using different assays (and more often than not, different values and scoring systems) can associate their results with the preceding part of the hierarchy (phenotypic attribute) by implementing a particular assay ontology (which they might get off the self or develop themselves). The advantage of this is that such procedures are good for scientific autonomy and moreover will allow more stable versions of PATO and reduce maintenance costs for both parts (namely, PATO and individual phenotype ontology curators). It should also be noted, that if values are linked directly to attributes, it will be much more complicated (requiring the use of instances) to assign what assays are allowed for common attributes. For example, abnormal would be a common value for most, if not all, attributes. Linking this value to assays used to determine it would require the generation of a new instance (and ids) that would hold the phenotypic attribute plus the value.

Finally, from a data collection and electronic recording perspective, it would be much easier for institutes to work from the assay, which does not require a general comprehension of the domain, in order to populate their knowledgebase. In the case of the EUMORPHIA project [9] (taken as an example), whose aim is to produce a standardized set of phenotyping protocols, it would be possible to develop a free-standing database based on the standard protocols (along with their values) it produces and associate these with related phenotypic attributes to produce description of phenotypic data.

6 Conclusions

We have proposed, presented and analyzed a general model for building mouse phenotype ontologies. We have highlighted some technical aspects and given general modeling directions. We believe that the idea of creating universal attributes applicable across domains using common application models, will present a powerful and meaningful way of achieving consistency in phenotypic data expression. This has the potential to solve current problems, faced by most databases, of expressing mutant phenotypes, currently described by free text. We are currently in the process of assessing the scalability and versatility of this approach to cope with complex phenotypic data.

In order to take advantage of the large amount of data that is continuously increasing and the particularities of each database and format, there is a need for facilities instigating human and machine-understandable data accessible and processable by humans and automated tools. The vision of a Semantic Web [28], as proposed by Tim Berners-Lee [29], will be realised by data used not only for display purposes but for automation, integration and reuse across various applications [30]. Achieving even partial disassociation from free text generates enormous computational and conceptual potential.

7 Acknowledgements

This project is funded by the European Commission under contract number QLG2-CT-2002-00930.

References

1. Open Global Ontologies (OBO). Available on line: http://obo.sourceforge.net/ . For PATO see: ftp://ftp.geneontology.org/pub/go/gobo/phenotype.ontology/phenotype.txt
2. Microarray Gene Expression Data Society (MGED). Available on line: http://www.mged.org/
3. Ringwald M, Eppig J.T., Begley D.A., Corradi J.P., McCright I.J., Hayamizu T.F., Hill D.P., Kadin J.A., Richardson J.E. The mouse gene expression database. *Nucleic Acids Res*, **29** (2001) pp. 98-101.
4. Mouse Genome Informatics (MGI), The Jackson, Laboratory, Bar Harbor, ME. Available on line: http://www.informatics.jax.org/
5. Davidson D., Bard J., Kaufman M. and Baldock R., The MouseAtlas Database: a community resource for mouse development, *Trend Genetics*, **17** (2001) pp. 49-51
6. Mammalian Phenotype Ontology, Mouse Genome Informatics Web Site, The Jackson Laboratory, Bar Harbor, Maine. Available on line: http://www.informatics.jax.org/searches/MP_form.shtml
7. European mutant mouse pathology database (Pathbase), University of Cambridge. Available online: http://www.pathbase.net)
8. The Gene Ontology Consortium. Gene Ontology: tool for the unification of biology. *Nature Genetics* **25** (2000) pp. 25-29.
9. EUMORPHIA. Understanding Human Discease through Mouse Genetics. Available on line: http://www.eumorphia.org/

10. Holloway E., Meeting Review: From Genotype to Phenotype: Linking Bioinformatics and Medical Informatics Ontologies, *Comparative and Functional Genomics* (2002) pp. 447-450
11. Gangemi A., Some tools and methodologies for domain ontology building, *Comp. Funct. Genom.*, **4** (2003) pp. 104-110
12. Duineveld A.J., Stoter R., Weiden M.R., Kenepa B., Benjamins V.R., WonderTools? A comparative study of ontological engineering tools, *Int. J. Hum-Comp. St.* 52 (2000) pp. 1111-1133
13. Stevens R., Bio-ontology Page. Available on line: http://www.cs.man.ac.uk/~stevensr/ontology.html
14. Denny M., Ontology Building: A Survey of Editing Tools, 2002, Available on line: http://www.xml.com/pub/a/2002/11/06/ontologies.html
15. Richter J. and Lewis S., DAG-Edit. Available on line: http://www.geneontology.org/doc/GO.tools.html#dagedit
16. GKB-Editor (Generic Knowledge Base Editor). Available on line: http://www.ai.sri.com/~gkb/
17. Bechhofer S., Horrocks I., Goble C. and Stevens R., OilEd: a Reason-able Ontology Editor for the Semantic Web**,** *Proceedings of KI2001*, **2174** (2001) pp 396--408
18. OpenKnoME. Available on line: http://www.topthing.com/
19. Grosso E. W, Eriksson H., Fergerson R. W., Gennari J. H., Tu S. W., and Musen M. A., Knowledge Modeling at the Millennium (The Design and Evolution of Protege-2000), 1999. Available on line: http://smi-web.stanford.edu/pubs/SMI_Abstracts/SMI-1999-0801.html
20. Protégé-2000. Available on line: http://protege.stanford.edu/
21. Wondertools. Available on line: http://www.swi.psy.uva.nl/wondertools/
22. WebOnto. Available on line: http://kmi.open.ac.uk/projects/webonto/
23. Stevens R., Goble C. A. and Bechhofer S., Ontology-based knowledge representation for bioinformatics, *Briefings In Bioinformatics*, **4** (2000), pp. 398-414
24. Stevens R., Knowledge Represenation Languages. Available on line: http://www.cs.man.ac.uk/~stevensr/onto/node14.html
25. Wroe C.J., Stevens R., Goble C.A., Ashburner M.. A Methodology To Migrate The Gene Ontology To A Description Logic Environment Using DAML+OIL. *Proceedings of the 8th Pacific Symposium on Biocomputing (PSB)*, Hawaii. 2003.
26. Stevens R., Wroe C., Bechhofer S., Lord P., Rector A., and Goble C., Building ontologies in DAML plus OIL, *Comp. Funct. Genom.* **4** (2003) pp. 133-141
27. Yeh I., Karp P.D., Noy N.F. and Altman R.B. Knowledge acquisition, consistency checking and concurrency control for Gene Ontology (GO), *Bioinformatics*, **19** (2003) pp. 241-248
28. T. Berners-Lee. Reflections on Web Architecture, Available on line http://www.w3.org/DesignIssues/CG.html
29. T. Berners-Lee. Available on line: http://www.w3.org/People/Berners-Lee/

30. Gkoutos G.V., Leach C. and Rzepa H.S., ChemDig: new approaches to chemically significant indexing and searching of distributed web collections, *New J. Chem.* **26** (2002) pp. 656-666