Position paper
W3C Workshop on Semantic Web for Life Sciences, October 2004 (Cambridge, MA, USA).

# Ontology based data integration and context-based mining for life sciences

### Authors

Peddinti V. Gopalacharyulu <ext-peddinti.gopal@vtt.fi>
Erno Lindfors <erno.lindfors@vtt.fi>
Catherine Bounsaythip <catherine.bounsaythip@vtt.fi>
Winnie Wefelmeyer <wefelmeyer@web.de>
Matej Orešič* <matej.oresic@vtt.fi>

*Corresponding author
VTT Biotechnology, Tietotie 2, P.O. Box 1500, Espoo, FIN-02044 VTT, Finland

### Abstract

Semantic web is a useful framework for accessing heterogeneous data sources, and its supporting technologies have already been utilized for data integration in life science domain. In this context, the primary challenge is to tackle the problem of evolving concepts, which is a similar challenge encountered by cognitive scientists using rule-based formalisms. In our data integration effort, we aim at implementing a restricted semantic web framework, complementing it by conceptual modeling methods.

### Contents

## 1. Introduction

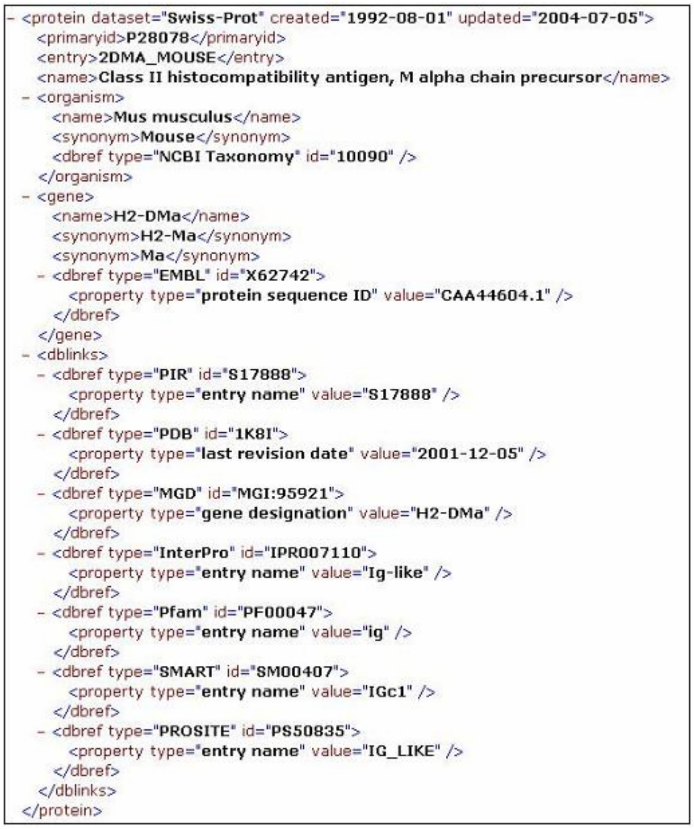
Overload of information is an increasing problem in life sciences. The emergence of novel technologies, with ability to generate large amounts of data, has not been matched with our ability to represent and exploit this data within the context of the system under investigation. Due to this mismatch the accumulation of knowledge and decision making processes based on this knowledge are actually increasingly difficult, and new solutions are needed to resolve this problem.

The emergence of semantic web (SW) and the supporting technologies such as XML and OWL/RDF offers a promise to facilitate organization of biological knowledge. Due to diversity of questions and applications relevant to life scientists, it is unlikely that all of the biological knowledge can be represented with the SW framework. We are interested in knowing to what extent the SW approach can be used for knowledge representation, and how to utilize this framework in answering biologist's questions.

## 2. Ontology based data integration

One of the on-going frustrations and practical challenges of life scientists is the diversity of identifiers for the same biological or chemical entities across different data sources. Any attempt at data integration should therefore start with identifying the "atoms of information" and creating solutions to resolve the names. XML is a useful technology for creating such identity-mappings across multiple data sources. Figure 1 shows an example of XML document we use to map protein names used across multiple pathway databases. The similar files, constructed according to corresponding XML Schema, are then stored in a native XML database, in our case Tamino XML Server. When describing more complex entities, such as diseases, or phenotypes in general, the vocabularies such as UMLS can facilitate the name resolution.

The next step in data integration is establishment of relationships between the entities. The relationships range from physical interactions at molecular scale to complex observations such as effect of a drug on specific phenotype. The SW framework is a powerful approach to organize different types of relationships between the biological entities, which also benefits from existing XML technologies. For example, the OWL/RDF definitions are represented in XML and then easily queried using XML databases. In our approach we map the names across multiple databases using XML, and relationships between them using existing or in house developed ontologies. The corresponding documents are stored in the native XML database. The underlying data, such as gene sequences or chemical compound information, can be stored elsewhere. The querying is therefore performed on the XML database.



```
- <protein dataset="Swiss-Prot" created="1992-08-01" updated="2004-07-05">
    <primaryid>P28078</primaryid>
    <entry>2DMA_MOUSE</entry>
    <name>Class II histocompatibility antigen, M alpha chain precursor</name>
  - <organism>
      <name>Mus musculus</name>
      <synonym>Mouse</synonym>
      <dbref type="NCBI Taxonomy" id="10090" />
    </organism>
  - <gene>
      <name>H2-DMa</name>
      <synonym>H2-Ma</synonym>
      <synonym>Ma</synonym>
    - <dbref type="EMBL" id="X62742">
        <property type="protein sequence ID" value="CAA44604.1" />
      </dbref>
    </gene>
  - <dblinks>
    - <dbref type="PIR" id="S17888">
        <property type="entry name" value="S17888" />
      </dbref>
    - <dbref type="PDB" id="1K8I">
        <property type="last revision date" value="2001-12-05" />
      </dbref>
    - <dbref type="MGD" id="MGI:95921">
        <property type="gene designation" value="H2-DMa" />
      </dbref>
    - <dbref type="InterPro" id="IPR007110">
        <property type="entry name" value="Ig-like" />
      </dbref>
    - <dbref type="Pfam" id="PF00047">
        <property type="entry name" value="ig" />
      </dbref>
    - <dbref type="SMART" id="SM00407">
        <property type="entry name" value="IGc1" />
      </dbref>
    - <dbref type="PROSITE" id="PS50835">
        <property type="entry name" value="IG_LIKE" />
      </dbref>
    </dblinks>
</protein>
```

Figure 1: An example translation of names and IDs

As soon as we start defining atoms of biological information and establishing relationships between them, we are faced with fundamental challenge of how to describe the biological phenomena. The SW approach is therefore inevitably limited to what we agree about in this context. For example, the level of agreement, and so the extent of the SW approach, is much lower within the whole life science community than within a specific organization, for example a drug discovery company, which is aimed toward a specific goal.

Our approach towards utilization of semantic web for biological data integration is to limit its scope to broadly agreed entities and relationships. This leads to low-level network representation of biological entities and their relationships, which serves as basis for deeper mining.

## 3. Context-based data mining

Formally, the challenges of studying biological phenomena are similar to those of cognitive science. In this domain two distinct approaches have been developed: (1) Symbolic and (2) Connectionist approach. The merits and disadvantages of both approaches have been thoroughly studied. The semantic web approach is similar to the symbolic model, which suffers from its lack of flexibility. In contrast, the limitation of the connectionist approach is its "black box" nature, i.e. the model is not easily interpretable in terms of known entities. In order to bridge the gap between the two methodologies, P. Gärdenfors introduced the conceptual modeling approach [1], in which not every concept has to maintain an absolute relationship with each other; they only need to be consistent when they both appear in the same context. In order to effectively mine the information using this approach, a context-dependent distance metric needs to be established across different types of relationships between the entities.

One research goal for us is to determine the semantic distance between different biological entities segregated across different databases. More specifically, given a concept in a particular source, one should be able to determine equivalent or semantically closest concept(s) in other sources. In general, the cases where there is a 1-1 correspondence between a term in one vocabulary with a term in an other are few. Terms might be related by synonyms, hyponyms and other hierarchical relationships. The issue is to estimate the semantic distance between a term and its multiple translations and to minimize it. In practice, once the metric is established, entities and relationships can be mapped using nonlinear mapping such as Sammon's mapping or self-organizing maps [2]. We created a generic Java user interface for visualizing the biological data retrieved from the databases. The queries are implemented using Tamino XML Server API. Our initial goal is to study topology of integrated biological networks, such as metabolic pathways, regulatory networks, and signaling networks [3].

## 4. Conclusions

Semantic web, with its supporting technologies, is a useful framework for accessing heterogeneous data sources. In life sciences, where concepts are rapidly evolving, SW has its limitations. In order to enhance its flexibility, it needs to be complemented by approaches such as conceptual modeling.

## References

1. Gärdenfors P: **Conceptual spaces: The geometry of thought**. Cambridge, MA: MIT Press; 2000.
2. Kohonen T: **Self organizing maps**, 3rd edn: Springer Verlag; 2001.
3. Gopalacharyulu PV, Lindfors E, Wefelmeyer W, Hollmen J, Oresic M: **Investigating the structure of integrated metabolic, protein-protein interaction, and regulatory networks**. 2004, in preparation.