From an ontology-based search engine towards a more flexible integration for medical and biological information

G. Marquet, C. Golbreich, A. Burgun
Laboratoire d'Informatique Médicale, Faculté de Médecine, 35033 Rennes, France

Christine.Golbreich@uhb.fr {Gwenaelle.Marquet, Anita.Burgun}@univ-rennes1.fr

Abstract. *Better understanding pathologies-genes relationships requires semantic integration of heterogeneous information distributed in multiple 'medical' and 'biological' sources. This paper presents an ongoing project that aims at developing an information integration system providing a unified access to biomedical resources. The basic idea is to use for semantic integration the existing knowledge available in standard domain terminologies e.g. GeneOntology™, UMLS® and databanks e.g. HUGO, GOA. A first tool, BioMeKe, has been achieved in that perspective. BioMeKe is an ontology-based search engine designed to facilitate the extraction and connection of biological and medical information, accessible from multiple public resources and biologists local repositories, for a system devoted to liver transcriptome analysis. The paper presents existing resources, describes BioMeke. Then, general lessons learnt from this practical experience are discussed.*

1 Introduction

The Word-Wide Web has made available a tremendous amount of biomedical information, but it remains tedious and time-consuming for biologists and physicians to access the information relevant to their queries. Multiple public resources are available in genomics including databanks such as SWISS-PROT[1], OMIM[2], LocusLink[3], GenBank[4], as well as many others, and some systems e.g. TAMBIS [27] are being developed to provide transparent access to bioinformatics sources. But a step further is needed for better understanding of the pathological processes that are involved in human diseases. To develop research, suggest new hypotheses about molecular mechanisms of human diseases and take advantage of recent research for patient care (e.g. [7]), biologists and physicians need to access and to relate numerous information from both genomics *and* medicine,. Therefore, tools to acquire and connect relevant data from existing resources are required. The problem is that there is considerable semantic heterogeneity between the sources, both intra and inter-domain. Different resources use different conceptualizations or different terms for the same concept or the same individual, although standard terminologies have been defined for each domainsuch as GeneOntology™ (GO) for molecular biology and genomics, and the Unified Medical Language System® UMLS® for the biomedical domain. The goal of the project is to develop a semantic integration system that offers a uniform interface for querying multiple heterogeneous sources *both* in genomics-molecular biology and in medecine, together with services for combining pieces of medical and molecular biology information relevant to answer queries. The basic idea is to use for semantic integration the knowledge available in the existing standard domain terminologies, namely GO and UMLS® and in databanks. Section 2 presents the main existing terminologies and databanks, section 3 describes BioMeKe (Biological and Medical Knowledge Extraction) [21], an ontology-based tool achieved to facilitate the access and association of knowledge from Web or local resources, for liver transcriptome analysis. Then, general lessons learnt from this practical experience are discussed.

2 Molecular biology and medicine resources

Information in the biomedical domain is scattered through multiple public databanks and bibliographic systems. But for each domain, « ontologies » have been defined to provide a unified and controlled vocabulary.

2.1 Ontologies

- **GeneOntology™** (GO)[5] is an ontology for molecular biology and genomics. GO is organized with three top categories Molecular Function, Biological Process, and Cellular Component. In May 2003 GO contained 7172 processes, 5386 molecular functions and 1265 component concepts. GO itself is not populated with gene products. It provides a controlled vocabulary for annotating sequences and gene products. GO concepts are broadly used as attributes in many public databases e.g. SWISS-PROT, as well as in

[1] http://us.expasy.org/sprot/
[2] http://www.ncbi.nlm.nih.gov/omim/
[3] http://www.ncbi.nlm.nih.gov/LocusLink/
[4] http://www.ncbi.nlm.nih.gov/Genbank/
[5] http://www.geneontology.org

specific applications. In the context of microarray experiments, biologists use GO for annotating the genes they are studying (Table 3).

- The **UMLS®** is a medical ontology intended to help health professionals and researchers use biomedical information from different sources [19]. It has two major components, the Metathesaurus®, a large repository of concepts (900,551 concepts in the 2003AA release), built by merging more than 100 families of vocabularies (including MeSH), and in grouping synonymous terms under a same concept and the Semantic Network, a limited network of 135 semantic types. The Metathesaurus concepts are assigned to one or more semantic types. The Metathesaurus is. In addition to the standard MeSH, the US National Library of Medicine created and maintains the **MeSH-ST**, ST standing forSupplementary Terms, which contains records that cover the fields of chemicals and molecular biology. (134,749 records in the 2003 release). The MeSH-ST files are updated continuously. MeSH-ST terms are integrated in the UMLS, making most of the terms, but not all the information provided by MeSH-ST accessible through the UMLS.

2.2 Multiple heterogeneous public databanks

Multiple public databanks provide information on genes, sequences and proteins, discovered upon a published experiment e.g. **SWISS-PROT** (SW), **GenBank**, **LocusLink, HUGO, G0 Annotation @EBI :**

- **LocusLink**[6] is a genes database to unify kowledge about genes. It provides official nomenclature, aliases, sequence accessions, cross-references to other banks via identifiers (EC Id, MIM Id, etc.).
- **HUGO**[7] (Human Gene Nomenclature Database) provides official gene names e.g. *ferritin, heavy polypeptide 1*, their synonyms, official symbol e.g. *FTH1*, and various links to other databases LocusLink, SWISS-PROT, OMIM, etc. via identifiers (e.g. ID: P02794 , LocusLink ID: 2495, OMIM ID: 134770).
- **GO Annotation @EBI**[8] (GOA) objective is to assign GO terms to *gene products*. GOA provides a file of human proteins assigned with GO terms, and a specific file of SWISS-PROT-TrEMBL data with their GO assignments. For each entry, GOA gives links towards GO molecular function, biological process, and cellular component (Table 1) and many cross-references towards public databanks, GenBank, LocusLink, MedLine, IPI, Ensembl, HUGO and RefSeq via an accession number, which is a means to get for a protein all the information and bibliography stored other databanks.

Molecular Function	Binding activity, Ferric iron binding activity, iron ion binding activity, iron ion homeostasis
Biological Process	Intracellular iron ion storage, Iron ion transport, Cell proliferation
Cellular Component	Ferritin complex

Table 1 : Assignments of GO terms to the protein ferritin heavy chain in GOA

There are also many public databanks in medicine. Among them **OMIM**[9] a database relating human genes and genetic disorders, and **MedLine**[10], which contains 12,000,000 biomedical journal citations accessed through the PubMed service of the National Library of Medicine.

2.3 Existing mappings and links

Many mappings and relations between standard ontologies and databanks are stored in these online resources.

2.3.1 Mappings and links databanks ↔ standard ontologies

- **Databanks ® GO**. For many biological databases, mappings to GO[11] ontology concepts are explicitely defined e.g. mappings of SW keywords to G0 terms (Table 2), mapping of Enzyme Commission Numbers entries to GO function ontology enzymes etc. Moreover, there are also implicit mappings since many public banks e.g. SWISS-PROT-TrEMB data are indexed with GO concepts thanks to GOA (§2.2)

```
!date: 2003/07/14 21:07:05
! Evelyn Camon, SWISS-PROT.
!Mapping of SWISS-PROT KEYWORDS to GO terms
SP_KW:Metal-thiolate cluster > GO:metal ion
binding ; GO:0046872
SP_KW:Metalloenzyme inhibitor > GO:enzyme
inhibitor activity ; GO:0004857 …
```

Table 2 Mappings of SW keywords to G0 terms

- **GO ® Databanks**. Reversely, GO terms are connected to various databanks (Prosite, InterPro, SW etc.) :

'External References' (Figure 1) defines links

[6] http://www.ncbi.nlm.nih.gov/LocusLink/
[7] http://www.gene.ucl.ac.uk/nomenclature/
[8] http://www.geneontology.org/#annotations
[9] http://www.ncbi.nlm.nih.gov/omim/
[10] http://www.ncbi.nlm.nih.gov/entrez/query.fcgi
[11] http://www.geneontology.org/doc/GO.indices.html

from GO terms to entries or indexes of external databanks, e.g. *iron ion homeostasis* is mapped to the SW keyword *Iron storage*.

'Associated Genes' associates GO terms to a list of Gene Products e.g. *iron ion homeostasis* is associated with *PF14_0518*, *CERU_HUMAN* etc. Reversely, for a Gene Product e.g. *CERU_HUMAN*, the field '**Associated to Terms**' provides its GO annotations e.g. *copper ion homeostasis* , *extracellular space*, *ferroxidase activity*, *iron ion homeostasis*

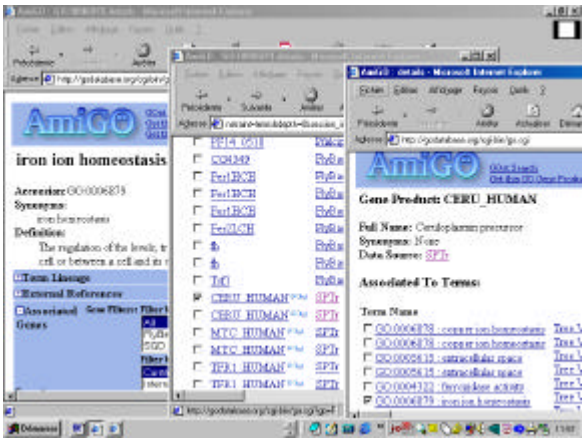

Figure 1 Browsing by chaining links: from a biological process, to a gene, and its function

- **Databanks ↔ UMLS®** There are also links from medical databases to the UMLS®, since the UMLS® is built by integration of dozens of existing terminologies that are used to code data in medical databanks, e.g. MeSH, which is used for indexing the biomedical literature in MedLine.

2.3.2 Links between databanks

Most databanks provides cross-references to other databases via accession numbers (§2.2). HUGO and GOA provide links particular useful for gene annotation systems:

HUGO relates gene to gene products, providing for a given gene the SWidentifier of its associated gene products. For instance, the gene. *ferritin, heavy polypeptide* (FTH1), is related to the SWID: P02794 of its corresponding protein. Accessing it then enables to get its stored information eg. its name *ferritin heavy chain* (FRIH_HUMAN), and synonym *Ferritin H*.

GOA relates gene products and GO terms. It provides for SWentries their relations with GO molecular function, biological process, and cellular component term. From these associations, it is possible to get for a protein, e.g. *Ferritin heavy chain* its GO assignments e.g. its molecular function "*iron ion binding activity*", biological process "*intracellular iron ion storage*".

2.4 Needs of information integration

It is really tedious for biologists and physicians looking for pathologies-genes relationships to browse the relevant information along such mappings and links (Figure 1). The problem is that the knowledge about the sources, their content, links to standard ontologies and between them, is not explicitely represented. An intelligent information integration system is needed providing them with a uniform access to sources both in genomics and medicine. A first tool has been achieved to meet urgent needs of researchers at INSERM U522, which study molecular mechanisms involved in human liver diseases (§3). The more long term objective is to build a more flexible system providing a unified access and services to combine information from various resources accessible on the Web or from local repositories, and to answer complex queries such as find « all the *metalloproteins* involved in *iron homeostasis* that have a *copper ion binding activity* and possible relationships to *liver* diseases » or « all *gene products* involved in proccesses such as *cell proliferation* and *ferric iron binding* with possible relationships to diseases *hemochromatosis and cataract* ».

3 BioMeKe

Biologists and physicians of INSERM U522 and LIM at Rennes study molecular mechanisms involved in human liver diseases, by means of transcriptome analysis. The objective is to find out the genes that are expressed in liver, to correlate them with patient data, in order to better understand pathological processes in liver. But for example, more than 3,000 SWentries are isolated from the tissue « Liver ». BioMeKe (Biological and Medical Knowledge Extraction), has been achieved to help them to extract and to associate medical and biological information accessible from multiple public sources, GenBank, Swissprot, LocusLink, MedLine, etc, and to correlate it to the biologists data laying in their local repository (Gedaw [10]).

3.1 Components and functionalities

BioMeKe, is an ontology-based tool composed of two parts: a core ontology and a query processor :

- **BioMeKe Core Ontology** (BCO) includes the main standard of the biomedical domain: for the medical domain, the UMLS® plus MeSH-ST, for genomics, GO plus GOA which has been added. since GO itself is not populated with gene products nor genes. Different synonyms may be used for a single gene in different databases and all synonymous are not necessarly found in a given database. Therefore, HUGO which adresses such issues is integrated into BCO. All terminologies are separatly stored in a MySQL relational database.

Links between items are dynamically created during the search for a given term or an annotation request.

- **BioMeKe Query Processor** uses BCO knowledge to search information in the external sources. It has three components. The *heterogeneity manager* (HM) uses HUGO and the UMLS for semantic unification of the different names and cross-references, HM returns for a gene its official name and symbol, and SWidentifiers. The *biological search module* (BS) is in charge of searching for biological information in GO, and to provide access to information of several public databanks. For a given term, BS searches for it in GO, GOA. If it is not directly found, it calls HM. If the term is matched and some synonyms provided, the search is done again for those new terms. If it is still unsuccessful, the SWor LocusLink ID provided by HM is then used to access GOA. Since GOA provides cross-references to other databanks, they can then be browsed to pick up relevant information. This unified access to the external banks is possible in the interactive mode, but not in the automatic mode (see 21 for details). The *medical search module* (MS) is in charge of searching for medical information in UMLS. For a given term, MS searches for it in the UMLS. If the term is found its context is displayed, including co-occurrences in MedLine, thus MedLine abstracts can be accessed through MeSH.

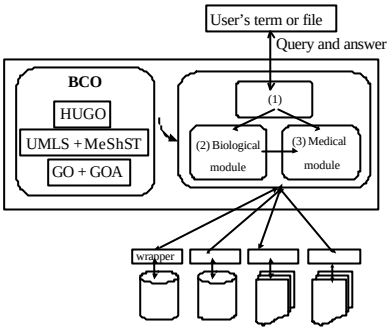

Figure 2 BioMeke

Implementation of the BioMeKe system relies on a MySQL relational database and JAVA A set of JAVA functions (wrappers) have been implemented to access to the content of several public databases, the BCO databases content is accessed thanks SQL queries.

BioMeKe prototype can be used either in an interactive or automatic mode. The automatic mode allows biological and medical annotation for a gene. The interactive mode offers a unified interface that enables, for a term entered by the user, to get biomedical information from the UMLS and GO and to browse across several public databanks the information related to a gene product.

Example. A user may search for biological and medical annotations for the gene *ferritin, heavy polypeptide 1*. BS searches for it in GO, GOA but does not find it, so it calls HM who returns the SWand LocusLink IDs of the corresponding protein *Ferritin heavy chain* (found in HUGO), from which the wanted biological information is obtained thanks GOA (Table 1). These accession numbers can also serve for browsing relevant information in other public databases. The user can search for the item in the UMLS but, the query *Ferritin heavy chain* is unsuccessful in UMLS. Indeed the term that is broadly used in medecine for this gene is *Ferritin H.* Reformulaing the query for the synonym *Ferritin H* provides its context, i.e. here the table MRCOC, from which concepts that co-occur in MedLine (e.g. liver, hemochromatosis, cataract) can be extracted and abstracts accessed (Table 1).

3.2 Application

Gene Name	*Ceruloplasmin (ferroxidase)*	*Ferritin*
Molecular Function	Oxidoreductase activity, Copper ion binding, Multicopper ferroxidase iron transport mediator activity	Binding activity, Ferric iron binding activity
Biological Process	Iron ion homeostasis, Copper ion homeostasis	Iron ion transport, Intracelluar iron ion storage, Cell proliferation, Iron ion homeostasis
Cellular Component	Extracellular space	Ferritin complex
Co-occurences Disease or Syndrom	Nervous system diseases, Iron overload[12], s etc.	Hemochromatosis, Cataract, etc.

Table 3: BioMeKe automatic annotation (extracts)

BioMeKe is being evaluated for the automatic annotation of genes for transcriptome analysis in the domain of liver diseases [10]. The process has three main steps:

Step1: Synonyms management. In order to reconciliate all the identifiers stored in the datawarehouse, and to solve gene synonymy problems, Locuslink identifiers are extracted from the GenBank file, then the HQ module provides the official names and symbols, and SW identifiers.

Step2: GO annotation. From SW identifiers, the BS module returns GO biological information via GOA

Step 3: UMLS annotation. The MS module uses the names and symbols provided at step 1 as inputs to search information in the UMLS. The UMLS annotations are filtered by semantic types to keep the 25 most relevant types to relate genotypes to

[12] the generated report contains 80 associated diseases

phenotypes (e.g 'Disease or Syndrom') (Table 3).

3.3 Limitations

BioMeKe main innovation is to be an ontology-based tool. However, the ontologies are non formal. Second, it is a procedural tool, and it provides semantic integration, but it is still limited.

- **Limits of non formal ontologies.**

GO and UMLS are not structured according formal principles, and exhibit many inconsistencies. For example (Fig. 3) in GO *Multicopper ferroxidase iron transport mediator activity* is child of *metal ion transporter activity*, which is sibling of *cation transporter activity*, while in another subtree, *metal ion homeostasis* is defined as child of *cation homeostasis*. Hierarchies for the *copper ion binding*, *copper ion transporter* functions, the *copper ion transport* process, all have a different pattern (resp. iron) etc. !

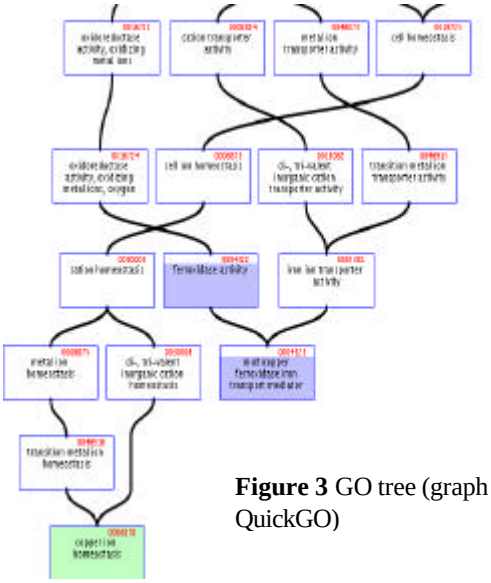

Figure 3 GO tree (graph from QuickGO)

Since GO 'is-a' hierarchy is not rigorous, it entails that gene annotation, may exhibit inconsistencies, redundancies, lacking, or heterogeneity., e.g. BioMeKe relates *Ceruloplasmin* to *Multicopper ferroxidase iron transport mediator activity* and to one arbitrary subsumer *Oxidoreductase activity*, but not to the others, e.g. *ion transporter activity*. Informal ontologies are clearly not appropriate in a context of integration.

- **Limits of the query engine**

BioMeKe is mainly grounded on various "mappings" and relations between the standard ontologies and databanks, or between databanks (by cross-references). However, since this knowledge remains implicit, many tasks are still grounded on user's skill and own responsibility: reformulation, selection of databanks to browse etc. Even assisted by BioMeKe, it remains difficult for researchers looking for pathologies-genes relationships, to navigate along such mappings and links across databanks to get the relevant information, and useful relations may easily be missed. BioMeKe is a procedural system, based on a fixed process. As the number of online databanks always increases, more automatization and more flexibility are required, providing extensibility and dynamic sources selection possibilities

- **Limits in semantic integration.**

BioMeKe management of heterogeneity is limited. First, it is mainly based on the synonyms found either in HUGO or the UMLS, but it does not exploit other information available in external databanks e.g. the synonymy of *ferritin heavy chain* and *Ferritin H* is asserted in SWISS-PROT. Second, GO, UMLS, HUGO must be frequently updated. Third, BCO has to be customized for specific use, e.g. a lexical database associating official gene names with complementary simplified names has to be added for liver transcriptome analysis. Moreover, heterogeneity concerns not only the data, but also at a more generic level, ontology concepts and relations. Although GO and UMLS have been recently merged on a lexical basis [25], generalizing mappings between ontologies is difficult. even with recent interactive tools such as PROMPT.

4 Lessons learnt

Some improvements are possible in BioMeKe. But, addressing all above problems clearly requires a *declarative* (knowledge-based or database) approach, allowing an explicit representation of the knowledge (ontology, mappings, queries) and an inference (query) engine with powerful services in particular for ontology automatic classification, consistency checking, and dynamic chaining of mappings. There is clear needs of formal ontology web languages, and of more flexible integration.

4.1 Needs of formal ontologies

Most people now agree about the limits of non formal ontologies and benefits of a formal language ontologies, for the Web in general [26] and in the biomedical domain [23] [27] [8]. First, "multiple viewpoints" is an old problem in biomedicine. For example, in GO functions, processes hierarchies are organized from a biochemical viewpoint derived from the EC Enzyme Commission classification, or from the chimical substances they act on*metal ion*, *cation*, *transition metal ion*, *iron*, *copper*. Multiple viewpoints are source of inconsistencies, when the ontology structuration is not automatized (§ 3.3). Moreover biologists and physicians are interested in clustering diseases, genes according to different dimensions, e.g. genes according to their functions or related pathologies, also in identifying all the gene products that share a same feature. Description Logics (DL) provide powerful services

for that, and the next W3C standard Ontology Web Language OWL[13] comes with useful tools e.g. the FaCT automatic classifier [14], the OilEd editor [2].

Example. The following example shows how constructing a global formal ontology for genomics in OWL will prevent from many inconsistencies. GO concepts below (in DL syntax) are based on MeSH definitions expressing that a cation is a postively charged atom, a cation divalent has valence of plus 2, an ion metal is a cation and a metal etc.

Cation:= Ion $\wedge$ ($\geq$ charged PositiveCharge)

CationDivalent:=Cation $\wedge$ ($\leq$ 2 charged PositiveCharge)

IonMetal: = Cation $\wedge$ Metal

TransitionMetalIon:= CationDivalent $\wedge$ ($\forall$ belongsto PeriodicGroup 3-12)

The "root" concepts Transport, Binding are using explicit roles "transported" "bound" relating them to the Chemical ontology concepts.

Transport:= Activity $\wedge$ ($\forall$ transported Chimical)

TransitionMetalIonTransportActivity:= Transport $\wedge$ ($\forall$ transported TransitionMetalIon)

Binding:= Activity $\wedge$ ($\forall$ bound Chimical)

TransitionMetalIonBinding:= Binding $\wedge$ ($\forall$ bound TransitionMetalIon)

Then all the sub-ontologies stemming from these concepts are globally consistent (and more generally so built ontologies, provided the related ontologies consistence e.g. Chemical). Such a formal ontology, also enables defining rigourous rules for gene annotation, for example "annotation must be done with the most specific function (resp. process, etc.)" since the others can be infered.

4.2 Needs of a more flexible information integration

Extensibility and real-time data are crucial requirements. Bioinformatics is a very fast-moving field. Web sources are multiple, with huge and constantly evolving contents. New online ontologies and specialized databanks often appear. Datawarehouses are not well appropriate and more flexible integration, such as mediator-based *centralized* systems, or new approaches proposing *distributed* integration are quite attractive [4]. *Local as view* (LAV) mediators defining the content of sources in terms of views over the global ontology, might be preferred to *global as view* (GAV), defining the global ontology in terms of views over the sources e.g. Tambis [27]. But although mediators are a significant progress, they may be not even flexible enough for scaling up the Web, and distributed systems are perhaps more appropriate. As described, databanks are not only data "sources" but also include precious links and mappings, through their cross-references to ontologies and other databanks. Such local relations between sources should be explicitly represented and directly exploited to infer new information. Peer-based integration where "every participant should be able to contribute new data and relate it to existing concepts and schemas, define new schemas that others can use as frames or reference for their queries or define new relationships between existing schema or data providers" is therefore a challenging approach to meet the extensibility and distribution encountered in biomedical information integration. But, whatever mediator or peer-based integration systems, rich formal languages are required for representing ontologies, queries, and mappings,. [9]

[13] http://www.w3.org/TR/owl-ref/

[14] http://www.cs.man.ac.uk/~horrocks/FaCT/

5 Discussion

Other systems have been achieved for gene annotations e.g. Source [5], or MatchMiner [14]. BioMeKe and Source annotation results have been compared on a sample of 364 genes : among the 250 gnes annotated by both systems, Source provide a more complete annotation for 15%, while BioMeKe for 38%. BioMeKe is based on GO and the UMLS, but several other ontologies exist like GALEN [23], TaO [1] for molecular biology and bioinformatics OMB (Ontology for Molecular Biology). The next perspective is to developp either a LAV mediator, opposed to TAMBIS GAV approach [27] or a distributed system. A LAV mediator requires a *global* ontology for genomics *and* medicine. Building such a formal ontology joins recent projects aiming at migrating GO to DL [30] or at merging the UMLS and GO [24]. Another perspective is to build an hybrid tool combining a search based on the formal ontology together with a classical search based on GO and UMLS.

6 Conclusion

BioMeKe is a first ontology-based tool facilitating the access and search of biological and medical information related to gene or gene products. An automatic mode allows annotation of gene files. However, selecting the sources to be explored and the information to extract is still too much grounded on the user's own skills and responsability. The current challenge is to provide a more automatized and flexible integration. A formal Web ontology language like OWL, and mediators or Peer-based distributed integration seem to be promising techniques. Main challenges are now to combine them, and to provide a language for mappings. Another bottleneck is to represent huge ontologies like GO and the UMLS in OWL and source mappings definitions for so multiple sources. Partial automatization seems the only reasonable solution.

7 References

1. Baker,P. et al. (1999) An ontology for bioinformatics applications. Bioinformatics, 15, 510–520.
2. Bechhofer S., et al. OILEd: a Reason-able Ontology Editor for the Semantic Web. Proc KI2001, 396-408. 2001.
3. Bernstein. P. Applying model management to classical meta data problems. Proc of the Conf. on Innovative Database Research (CIDR), 2003.
4. Bernstein P et al. Data management for peer-to-peer computing: A vision, Workshop WebDB 2002.
5. Cantor MN, Lussier YA. A knowledge framework for computational molecular-disease relationships in cancer. Proc AMIA Symp 2002;101-5
6. Diehn M et al. SOURCE: a unified genomic resource of functional annotations, ontologies, and gene expression data. Nucleic Acids Research, 2003, 31, 1: 219-23
7. Goasdoue, F, Lattes, V, Rousset, MC. The Use Of Carin Language and Algorithms for Information Integration: The PICSEL System Int J Coop Infor Sy stems, 9(4), 383-401, 2000.
8. Golbreich, C et al. Web ontology language requirements w.r.t expressiveness of taxononomy and axioms in medecine, ISWC 2003, Springer.
9. Golbreich, C., B., Burgun A. Challenges for Biomedical Information. Position Statement Paper. Semantic Integration Workshop, ISWC 2003
10. Guérin E et al. UML modeling of Gedaw: A gene expression data warehouse specialised in the liver. JOBIM; 2002 France, Saint-Malo. p. 319-334.
11. Hahn, U., Schulz S. Turning Lead into Gold? Feeding a Formal Knowledge Base with Informal Conceptual Knowledge. EKAW 2002: 182-196
12. Halevy A. Y. Answering queries using views. The VLDB Journal, 10(4):270-294, 2001.
13. Halevy AY, Zachary G, Ives Suciu D, Tatarinov I. Schema mediation in peer data management sy stems. ICDE, 2003.
14. Kimberly J.B. et al. MatchMiner: a tool for batch navigation among gene and gene product identifiers. Genome Biology, 2003 4(4):R27
15. Kirk, T, Levy, AY., Sagiv Y, D .Srivastava. The Information Manifold, Information Gathering from Heterogeneous, Distributed Environments, AAAI Spring Symposium Series, Stanford University, March 1995.
16. Levy A. Y, Logic-Based Techniques in Data Integration Logic Based Art Int , J Minker. Ed Kluwer,, 2000
17. Levy A. Y, Rousset MC, The Limits on Combining Recursive Horn Rules with Description Logics, AAAI/IAAI, Vol. 1 (1996)
18. Li Q, Shilane P, Noy NF, Musen MA. Ontology acquisition from on-line knowledge sources. Proc AMIA Symp. 2000;497-501.
19. Lindberg DAB, Humphreys BL, McCray AT, The Unified Medical Language System. Meth Inform Med, 1993, 32(4): 281-91
20. Lucie Xyleme: A dynamic warehouse for XML Data of the Web. IEEE Data Eng Bull 24(2): 40-47 (2001)
21. Marquet G et al. BioMeKe: an ontology-based biomedical knowledge extraction system devoted to transcriptome analysis, MIE 2003
22. Povey S et al. The HUGO Gene Nomenclature Committee (HGNC).Hum Genet. 2001;109(6):678-80
23. Rector A. et al. The GRAIL concept modelling language for medical terminology. Art Int Med,9:139-171, 1997.
24. Sarkar, I et al. Linking biomedical language information and knowledge resources in the 21st century: GO and UMLS, *Pac Symp Biocomput*, 2003 **8**, 427-450.
25. Schulz S, Hahn U. Medical knowledge reengineering--converting major portions of the UMLS into a terminological knowledge base. Int J Med Inf. 2001 Dec;64(2-3):207-21.
26. Staab S Edt Ontologies'KISSES in standardization. IEEE Intelligent Systems, 70-79
27. Stevens R et al. TAMBIS: transparent access to multiple bioinformatics information sources. Bioinformatics. 2000 Feb;16(2):184-5.
28. The Gene Ontology Consortium. Creating the gene ontology resource:design and implementation. Genome Res 2001,11(8):1425-1433
29. Wiederhold G. Mediators in the architecture of future information systems. Computer, 1992, 25(3): 38-49
30. Wroe CJ, Stevens R, Goble CA, Ashburner M. A methodology to migrate the gene ontology to a description logic environment using DAML+OIL. Pac Symp Biocomput. 2003:624-35.

Acknowledgements

The authors thank F Moussouni, E Guérin, O Loréal, F Mougin for their participation in BioMeKe.