Suggested Ontology for Pharmacogenomics (SO-Pharm): Modular Construction and Preliminary Testing

Adrien Coulet, Malika Smaïl-Tabbone, Amedeo Napoli, Marie-Dominique Devignes

Suggested Ontology for Pharmacogenomics (SO-Pharm): Modular Construction and Preliminary Testing

Adrien Coulet,[1,2] Malika Smaïl-Tabbone,[2] Amedeo Napoli,[2] and Marie-Dominique Devignes[2]

[1] KIKA Medical,
35 rue de Rambouillet, 75012 Paris, France
[2] LORIA (UMR 7503 CNRS-INPL-INRIA-Nancy2-UHP),
Campus scientifique, BP 239, 54506 Vandoeuvre-lès-Nancy, France

Abstract. Pharmacogenomics studies the involvement of interindividual variations of DNA sequence in different drug responses (especially adverse drug reactions). Knowledge Discovery in Databases (KDD) process is a means for discovering new pharmacogenomic knowledge in biological databases. However data complexity makes it necessary to guide the KDD process by representation of domain knowledge. Three domains at least are in concern: genotype, drug and phenotype. The approach described here aims at reusing whenever possible existing domain knowledge in order to build a modular formal representation of domain knowledge in pharmacogenomics. The resulting ontology is called SO-Pharm for Suggested Ontology for Pharmacogenomics. Various situations encountered during the construction process are analyzed and discussed. A preliminary validation is provided by representing with SO-Pharm concepts some well-known examples of pharmacogenomic knowledge.

1 Introduction

Pharmacogenomics is the study of genetic determinants of drug responses. It involves relationships between at least three actors of interindividual differences in drug responses: genotype, drug, and phenotype (Fig. 1)[1]. Relevant genotype features are mostly genomic variations and particularly Single Nucleotide Polymorphisms (SNP). The latters are one-nucleotide substitutions occurring in a studied population with a minimum frequency of 1 %. Such genomic variations modulate drug effect, and have consequences on individual phenotype from the microscopic level (gene expression, protein activity, molecule transport, etc.) to the macroscopic level (clinical outcomes, etc.).

At present, best-recognized and completely developed examples of genomic variations altering drug response in human are monogenic traits acting on drug metabolism. Nevertheless, description of complex polygenic systems has recently proven that regulatory networks and many non genetic factors (e.g., environment, life style) also influence the effect of medications. Consequently, the discovery of new pharmacogenomic knowledge is a challenging task that necessitates

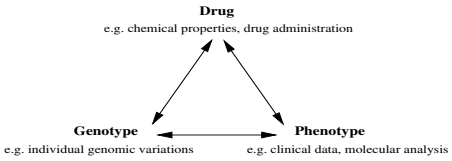

Fig. 1. Triangular schematization of the pharmacogenomic domain.

the management of complex data. For example, the design of a clinical trial in pharmacogenomics relies on the selection of genes involved in drug response, selection of associated relevant genomic variations and on knowledge about the phenotypes associated with these genomic variations [2]. An interesting research direction is the integration of biological data stored in public annotated biological data banks, and clinical data resulting from clinical trials. This integration may allow, in a second stage, the discovery of pharmacogenomic knowledge thanks to the KDD process.

The KDD process is aimed at extracting from large databases information units that can be interpreted as reusable knowledge units (such as RDF/OWL triples). This process is based on three major steps: (a) the datasets are extracted from selected data sources and prepared for data mining, (b) are mined (with symbolic or numerical methods), finally, (c) the extracted information units are interpreted by domain expert to become reusable knowledge units [3]. All along this process, domain knowledge, embedded within an ontology, can be used to guide the various steps:

a) During the preparation step it facilitates integration of heterogeneous data.
b) During the mining step, domain knowledge guides the filtering of input and output data.
c) In the interpretation step, it helps the experts for reasoning on the extracted units.

In order to achieve KDD in pharmacogenomics, we decided to develop a knowledge-based approach and therefore to explicit domain knowledge within an ontology. More and more biomedical ontologies are being developed today and often cover overlapping fields. To favor reuse of and access to ontologies, most biological ontologies are freely available. For instance, the Protégé ontologies library [4] provides various formal ontologies and the Open Biomedical Ontologies (OBO) portal [5] gathers many controlled vocabularies for the biomedical domain. Although the associated ontology is not available, the PharmGKB project has led to the construction of a valuable structured repository for pharmacogenomic data, aimed at catalyzing scientific research in this domain [6]. It provides a data model and a partial vocabulary for genotype and phenotype data of individuals involved in pharmacogenomic studies. In previous work, we developed the SNP-Ontology as a formal representation of genomic variation domain [7].

This paper describes the construction process of a "Suggested Ontology for pharmacogenomics" (SO-Pharm), that reuses existing ontologies designed for

pharmacogenomics sub-domains: genotype, drug, and phenotype. Section 2 describes the method used to build SO-Pharm and its content. Section 3 presents a preliminary testing of the ontology thanks to assertions of some established pharmacogenomic knowledge. Section 4 concludes on the work.

2 SO-Pharm Construction

2.1 Methodology Choice

Semiautomatic methods such as classification, itemset search, association rule extraction, text mining can be employed for ontology construction [8]. However, a manual construction is preferred here because of the objectives assigned to the ontology. In addition, the complexity of the field has favored a close collaboration with domain experts, nicely compatible with manual construction. Indeed, one difficulty consists in choosing and defining adequate concepts and properties for expressing pharmacogenomic knowledge. Manual construction is associated here with the use of a clearly defined methodology. Outlines of iterative processes for ontology construction have been described in [9,10,11]. We adapt here these methodologies to the case of pharmacogenomics, based on four steps:

(i) specification, embedding definition of ontology domain and scope;
(ii) conceptualization, that includes definition of list of terms and of concepts, and their articulation with existing ontologies;
(iii) formalization, i.e., the translation of the conceptualization in a knowledge representation formalism (e.g. description logics);
(iv) implementation, i.e., coding the formalized ontology in a knowledge representation language (e.g. OWL).

In the next sections we analyze and discuss the original orientations adopted during the SO-Pharm construction process.

2.2 Construction Issues

Specification. Domain and scope of SO-Pharm are primarily defined as follows. The domain considered should cover pharmacogenomic clinical trials. The ontology has to precisely represent individuals and groups of individuals involved in trials, their genotype, their treatment, their observed phenotype and the potential pharmacogenomic relations discovered between these concepts. Currently, SO-Pharm concepts do not cover epigenotype features, regulatory networks or metabolic pathways. SO-Pharm scope is to guide KDD in pharmacogenomics. According to the various steps of KDD process, SO-Pharm should reveal helpful in the following situations:

- integrating complementary data from various scopes: e.g. protein annotations and enzyme activity measurement;
- reconciling heterogeneous data: e.g. heterogeneous descriptions of genomic variations pertaining from locus specific databases and dbSNP;

– guiding data mining: for instance selection of a given class of genomic variations according to relations between these variations and the focus of the study;
– expressing data mining results as knowledge units: in order to compare with existing knowledge units and to infer new knowledge units;
– reusing of discovered pharmacogenomic knowledge: e.g. knowledge sharing between several independant projects.

During the specification step some strict nomenclature guidelines (e.g., for naming classes, associations, concepts, properties) are defined for the whole construction process. Then lists of domain terms are established. In the case of SO-Pharm ontology, the domain expert constitutes four primary term lists thanks to his own knowledge regarding respectively clinical trial, genotype, treatment, and phenotype descriptions. In parallel, data or knowledge resources in the domain are listed. These highly heterogeneous resources, including conceptual data model (in UML or UML-like), XML schemas, databases, ontologies, controlled vocabularies are displayed in Table 1 (*n.b.*: * are OBO ontologies). The study of their structure and content allows to considerably enrich the term lists.

The previous resource list is then refined for selecting relevant reusable knowledge resources according to following criteria (Table 2). First, it has been decided to take into account OBO ontologies, which are mostly used and known. Second, we have preferred the ontologies involved in the OBO-Foundry project that tries to adopt quality principles in ontology development [12]. The current resource list may be extended in the future and enriched with other interesting resources such as GO, Pathway Ontology, NCI, eVOC, Amino Acid Ontology, GandrKB.

Conceptualization. A UML class diagram is used here for representing the conceptual model of SO-Pharm. Term lists are exploited to identify ontology concepts which are assigned a name and a precise definition (free text). In SO-Pharm, a clinical item (or clinical data, or item) is defined as the measurment of a quantity for a given person, during a particular event, according to a measurment method. As well, a drug is composed of chemical compounds and may be included in a drug treatment and may have a commercial name. When concepts are identified, their hierarchical and non-hierarchical (i.e. object properties) relations are modeled by UML class diagrams. These diagrams are well adapted for conceptualization of domain knowledge because of their expressiveness and openness [13]. Fig. 2, 3 and 4 display UML class diagrams designed during SO-Pharm construction.

Articulation between the SO-Pharm concepts and external ontologies concepts is also established during this step (see Table 2 for prefix legend in UML class diagrams). The kind of relation (i.e. *embedding* or *extension*) invoked for reusing an ontology depends on its type [10]. Indeed, the majority of ontologies in biomedical domain may be organized into three categories: *meta-ontologies* providing domain-independent concepts and properties to be used as compounds for more specific ontologies (e.g. DOLCE, SUMO); *domain reference ontologies* representing a particular domain of reality and sorting entities of the domain

Table 1. List of explored resources for constructing term lists of the various domains

Resource name	*Resource type*	*Domain*	*URL*
dbSNP	XML schema, data model	genotype	http://www.ncbi.nlm.nih.gov/projects/SNP/
HapMap	XML schema	genotype	http://www.hapmap.org/
HGVBase	DTD, data model	genotype	http://hgvbase.cgb.ki.se/
OMIM	Data resource	genotype, phenotype	http://www.ncbi.nlm.nih.gov/omim/
OMG SNP	Data model	genotype	http://www.omg.org/technology/documents/formal/snp.htm
MECV	Controlled vocabulary	genotype	http://www.ebi.ac.uk/mutations/
PharmGKB	XML schema, data model	genotype, drug, phenotype	http://www.pharmgkb.org/
Pharmacogenetics Ontology	Controlled vocabulary	genotype, phenotype	http://www.pharmgkb.org/home/projects/project-po.jsp
Sequence Ontology	Controlled vocabulary*	genotype	http://song.sourceforge.net/
Gene Ontology	Controlled vocabulary*	genotype	http://www.geneontology.org/
PubChem	Data resource	drug	http://pubchem.ncbi.nlm.nih.gov/
RX-Norm	Controlled vocabulary	drug	http://www.nlm.nih.gov/research/umls/rxnorm/index.html
CDISC	XML schema	phenotype	http://www.cdisc.org/
ICD-10	Controlled vocabulary	phenotype	http://www.who.int/classifications/icd/
Disease Ontology	Controlled vocabulary*	phenotype	http://diseaseontology.sourceforge.net
Mammalian Phenotype	Controlled vocabulary*	phenotype	http://www.informatics.jax.org/searches/MP_form.shtml
PATO	Controlled vocabulary*	phenotype	http://obo.sourceforge.net/
ChEBI	Controlled vocabulary*	drug	http://www.ebi.ac.uk/chebi/
Pathway Ontology	Controlled vocabulary*	genotype, phenotype	http://rgd.mcw.edu/tools/ontology
SNOMED-Clinical	Controlled vocabulary	phenotype	http://www.snomed.org/snomedct/glossary.html

Table 2. List of selected resources for constructing SO-Pharm

Ontology name	*Description*	*Prefix*	*Namespace*
MECV	genomic variation classification	MECV	http://www.loria.fr/~coulet/ontology/mecv.owl
SNP-Ontology	genomic variations	SNPO	~/ontology/snpontology.owl
Pharmacogenetics Ontology	describes genotyping and phenotyping methods	PO	~/ontology/pharmacogeneticsontology.owl
Disease Ontology	a classification of disease	DO	~/ontology/diseaseontology.owl
Mammalian Phenotype	phenotype features	MPO	~/ontology/mammalianphenotypeontology.owl
PATO	attributes and values for phenotype description	PATO	~/ontology/pato.owl
ChEBI	molecular compounds	CHEBI	~/ontology/chebi.owl

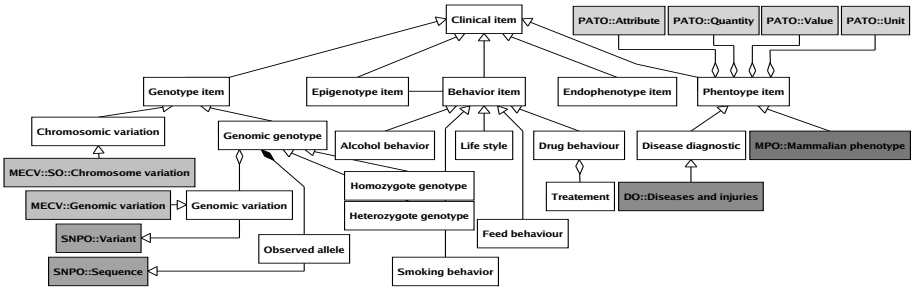

Fig. 2. UML class diagram for clinical item.

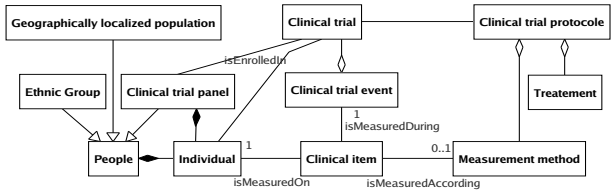

Fig. 3. UML class diagram for clinical trial.

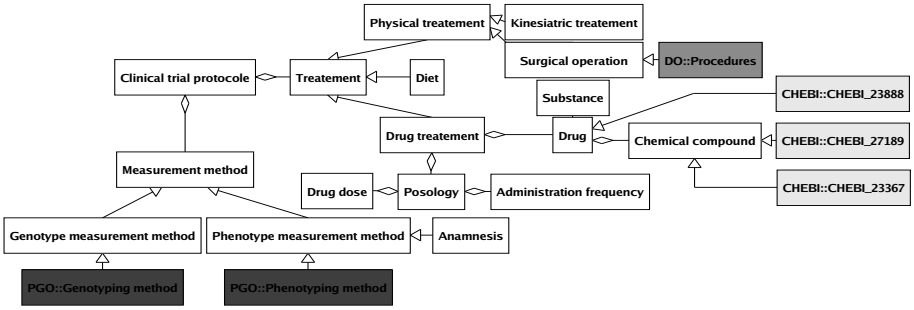

Fig. 4. UML class diagram for clinical trial protocol.

according to constraints expressed in a formal language (e.g. description logics); and *terminology-based application ontologies* which are controlled vocabularies often designed to annotate biological databases [14]. Most of OBO ontologies belong to this third family -except for the PATO ontology that can be considered as a meta-ontology. In SO-Pharm, several highly specialized vocabularies such as Disease Ontology are *embedded* meaning that these ontologies are reused in an ontology having a wider scope. On the opposite, formal ontologies, such as SNP-Ontology, are high level domain representations *extending* definitions of more specific concepts pertaining from other ontologies. For example, the *variant* concept of SNP-Ontology subsumes (i.e. extends the definition of) the *genomic variation* concept in SO-Pharm. The latter itself subsumes other more specific concepts from the OBO Sequence Ontology (e.g. *single deletion*).

In summary, SO-Pharm construction involves the design of modules favoring the reuse of concept definitions existing in other ontologies. Besides these reused concepts, additional SO-Pharm concepts and properties are defined locally.

Formalization and Implementation. SO-Pharm is implemented with the Protégé knowledge editor and coded in OWL. Formalization and implementation steps are nested. On the basis of previously designed UML class diagrams, concepts and (object and datatype) properties are formally defined in the Protégé framework. For example:

$$\begin{aligned}(1)\ clinical\_item \sqsubseteq \quad & \exists\ measuredOn.individual \\ & \sqcap \exists\ measuredDuring.clinical\_trial\_event \\ & \sqcap \exists\ measuredAccording.measurement\_method\end{aligned}$$

$$\begin{aligned}(2)\ drug \sqsubseteq \quad & \forall\ isComposedOf.chemical\_compound \\ & \sqcap \exists\ isPartOf.drug\_treatment \\ & \sqcap \exists\ isCommercialisedAs.substance\end{aligned}$$

Unfortunately, no system allows an automatic conversion of UML class diagrams into OWL statements. Simple classes and associations are easily converted, but complex ones need particular attention. For example, since the description logic formalism on which OWL is based is limited to binary relationships, the translation of UML n-ary relationships is not straightforward. The most common way to represent n-ary relationships in an ontology formalism is reification [15]. In our work, conceptualization prevents n-ary relationships by preferring addition of new classes or association classes with several binary relationships.

Apart from SNP-Ontology, ChEBI and Disease Ontology which have been directly downloaded in OWL (http://www.fruitfly.org/~cjm/obo-download/), most external ontologies are not available in OWL. They had to be translated first. Pharmacogenetics Ontology has been manually coded in OWL from text sources. Because of redundancies, Mutation Event Controlled Vocabulary and Sequence Ontology have been manually integrated and implemented in OWL. PATO and Mammalian Phenotype Ontology have been converted from OBO

format to OWL thanks to the BONG-Protégé plugin [16]. OWL-translated ontologies are then associated to namespaces and are prefixed (Table 2) for being virtually imported in SO-Pharm where they are articulated by concepts definitions:

(3) $CHEBI : molecular\_entities \sqsubseteq chemical\_compound$
(4) $MECV : genomic\_variation \sqsubseteq genomic\_variation \sqsubseteq SNPO : variant$

The consistency and the class hierarchy of SO-Pharm including reused ontologies have been validated with Racer 1.9 at each stage of the implementation thanks to standard reasoning mechanisms [17]. Manual construction and expert contribution appear as solid advantages for articulating existing ontologies in a sensible way. It allows a proper use of reasoning mechanisms despite of unclear/various purpose concepts that co-exist or overlap in ontologies.

3 Preliminary Testing of SO-Pharm semantics

As a preliminary validation of the ontology, several examples of published pharmacogenomic knowledge have been expressed with the SO-Pharm concepts. This is performed by asserting individual cases presenting genotype, treatment and phenotype features described in the litterature. The assertions of individuals and related information (clinical trial, treatment) lead us to refine SO-Pharm concepts. Genotype (encompassing several genomic variation), homozygosity/heterozygosity, poor/rich metabolizer, anamnesis, treatment effect are examples of concepts added during the first round of testing in order to be able to handle the representation of selected precise pharmacogenomic examples. Groups of individuals have been artificially constituted to gather individuals presenting common traits. Three groups of individuals are presented in expression (5), (6) and (7):

$$
\begin{aligned}
(5)\ \ & demyelinised\_patient \sqsubseteq \quad person \\
& \sqcap \forall presentsGenotype. \quad (\exists isTheGenotypeObservedFor.(\ni \{rs1142345\}) \\
& \qquad\qquad \sqcap \exists isComposedOf. \ni \{G\}) \\
& \sqcap \forall presentsPhenotype. \ (\forall measuredAccording.(\ni \{6TGN\_proto\}) \\
& \qquad\qquad \sqcap \forall PATO : hasAttribute.(\ni \{6TGN\_conc\}) \\
& \qquad\qquad \sqcap \forall PATO : hasValue.(\ni \{high\})) \\
& \sqcap \forall isEnrolledIn. \qquad (\forall isDefinedBy.(\forall isComposedOf. \\
& \qquad\qquad (\ni \{mercaptopurine\_treatment\}))
\end{aligned}
$$

The meaning of (5) is that demyelinised patients are persons who present both the allele G for the genomic variation rs1142345, a high concentration in 6-TGN

and are treated with mercaptopurine in a clinical trial.

$$
\begin{array}{lll}
(6) & over\_anti\_coagul\_patient & \sqsubseteq person \\
& \sqcap \forall presentsGenotype. & (\forall isTheGenotypeObservedFor.(\ni \{rs1057910\}) \\
& & \sqcap \forall isComposedOf.(\ni \{C\})) \\
& \sqcap \forall isComposedOf. & (\ni \{CYP2C9\_2\}) \\
& \sqcap \forall presentsPhenotype. & (\forall measuredAccording.(\ni \{bleeding\_obs\}) \\
& & (\forall PATO : hasAttribute.(\ni \{bleeding\}) \\
& & \sqcap \forall PATO : hasValue.(\ni \{high\_bleeding\})) \\
& \sqcap \forall isEnrolledIn. & (\forall isDefinedBy.(\forall isComposedOf. \\
& & (\ni \{warfarin\_treatment\}))
\end{array}
$$

Patients with an over anti-coagulation (6) are persons who present both the allele C for the genomic variation rs1057910, the CYP2C9*2 genotype, and important bleeding and are treated with warfarin in a clinical trial.

$$
\begin{array}{lll}
(7) & venous\_thrombos\_patient & \sqsubseteq person \\
& \sqcap \forall isComposedOf. & (\forall sex.(\ni \{female\})) \\
& \sqcap \forall presentsClinicalData. & (\forall measuredAccording.(\ni \{drug\_anamnesis\}) \\
& & \sqcap \forall isComposedOf.(\ni \{oral\_contraceptive\})) \\
& \sqcap \forall isComposedOf. & (\ni \{F2\_A20210\} \sqcup \ni \{F5\_A1691\})
\end{array}
$$

The patient group with venous thrombosis (7) are women who are using oral contraceptive and present the F2_A20210 or the F5_A1691 genotype. Additional assertions have led to localize and fix a few mistakes in the ontology instantiation, and to precise restrictions on object and datatype relationships. The number of required modifications decreased with each new assertion until the quasi-stability of the ontology was reached.

In view of expressing pharmacogenomic knowledge units, SO-Pharm was enriched with a simple property *mayBeRelated* that allows to link genotype item, phenotype item and chemical compound. Every required modification in the ontology is done according to a new construction iteration by updating the conceptual model, looking for reusable concepts, and finally modifying the ontology.

SO-Pharm is a crucial component for a future knowledge-based application dedicated to pharmacogenomic knowledge discovery. A complete validation has now to be conducted in the frame of the intended knowledge-based application, i.e. aimed at evaluating how SO-Pharm is able to guide the KDD process. A significant issue will be to develop appropriate wrappers to achieve heterogenous data integration as in [7].

SO-Pharm and external ontologies it includes are available (in OWL format) at http://www.loria.fr/~coulet/ontology/sopharm.owl. We plan to submit SO-Pharm to OBO portal to gain in visibility and facilitate further improvements.

4 Conclusion

Much of the quality of the SO-Pharm ontology relies on the initial extensive enumeration of term lists and use cases (specification and conceptualization steps).

Expert interviews and overview of existing ontologies are necessary for that purpose. Interestingly case studies aimed at expressing already existing knowledge extracted from the litterature lead us to enrich SO-Pharm with additional concepts in an iterative process.

Embedding and extension strategies are used to anchor existing ontologies to SO-Pharm concepts. This conceptualization task will become more and more important since more and more autonomous ontologies are produced in the biomedical domain, e.g. for representing phenotype with formal ontologies.

References

1. Evans, W., Relling, M.: Pharmacogenomics: moving toward individualized medicine. Nature **429** (2004) 464–468
2. Russ B. Altman, R.,Klein, T.: Challenges for biomedical informatics and pharmacogenomics. Annu. Rev. Pharmacol. Toxicol. **42** (2002) 113–33
3. Frawley, W., Piatetsky-Shapiro G., Matheus, C.: Knowledge Discovery in databases: An Overview, Knowledge Discovery in Databases. AAAI/MIT Press (1991) 1–30
4. Protégé ontology library [OnLine]. http://protege.cim3.net/cgi-bin/wiki.pl?ProtegeOntologiesLibrary
5. OBO web site [Online]. http://obo.sourceforge.net/
6. Oliver, D., Rubin, D., Stuart, J., et *al*: Ontology development for a pharmacogenetics knowledge base Pac. Symp. Biocomput. **7** (2002) 65–76
7. Coulet, A., Smaïl-Tabbone, M., Benlian, P. et *al*: SNP-Converter: An Ontology-Based Solution to Reconcile Heterogeneous SNP Descriptions. In proceedings of the 3rd Workshop on Data Integration in the Life Sciences (DILS'06), Hinxton, UK (2006)
8. Omelayenko, B.: Learning of Ontologies for the Web: the Analysis of Existent Approaches. In Proceedings of the Internat. Workshop on Web Dynamics, 8th Conference on Database Theory ICDT'01 (2001)
9. Noy, N., McGuinness, D.: Ontology development 101: A guide to creating your first ontology. Stanford Knowledge Systems Laboratory Technical Report KSL-01-05 (2001)
10. Gabor, N.,: WP3: Service ontologies and service description. DIP (Data, Information and process Integration with SW services), FP6-507483 (2005)
11. Uschold, M., King, M.: Towards a Methodology for Building Ontologies. In Workshop on Basic Ontological Issues in Knowledge Sharing, IJCAI-95 (1995)
12. OBO Foundy web site [OnLine]. http://obofoundry.org/
13. Kogut, P., Cranefield, S., Hart, L., Dutra, M., Kokar,M. Smith, J.: UML for Ontology Development. The Knowledge Engineering Review **17,1** (2002) 61–64
14. Rosse, C., Kumar, A., Mejino, J., et *al*: A Strategy for Improving and Integrating Biomedical Ontologies. AMIA Symposium Proceedings (2005) 639–43
15. Noy, N., Rector, A.: Defining N-ary Relations on the Semantic Web. [OnLine]. http://www.w3.org/TR/2006/NOTE-swbp-n-aryRelations-20060412/
16. Wroe, C., Stevens, R., Goble, C., Ashburner,M.: A Methodology to Migrate the Gene Ontology to a Description Logic Environment Using DAML+OIL. Pac. Symp. Biocomput. **8** (2003)624–636 Pac. Symp. Biocomput. **7** (2002) 65–76
17. Haarslev, V., Moller, R.: RACER System Description. In First Internat Joint Conference on Automated Reasoning (IJCAR'2001) 701–706 (2001)