

IMISE-REPORTS

Herausgegeben von Professor Dr. Markus Löffler

H. Herre, R. Hoehndorf, J. Kelso, S. Schulz (Eds.)

OBML 2010 Workshop Proceedings

Mannheim, September 9-10, 2010

IMISE-REPORT Nr. 2/2010

UNIVERSITÄT LEIPZIG
Medizinische Fakultät

Impressum

Editoren: Heinrich Herre, Robert Hoehndorf, Janet Kelso, Stefan Schulz

Herausgeber: Prof. Dr. Markus Löffler

Redakteur: Frank Loebe

Institut für Medizinische Informatik, Statistik und Epidemiologie (IMISE)

Härtelstraße 16-18, 04107 Leipzig

Tel.: (0341) 97-16100, Fax: (0341) 97-16109

Internet: <http://www.imise.uni-leipzig.de>

Druck des Einbandes: Buch- und Offsetdruckerei Herbert Kirsten

Redaktionsschluss: 31. August 2010

© IMISE 2010 (Report als Sammelband). Das Copyright der Einzelartikel verbleibt bei den Autoren.

Alle Rechte vorbehalten. Nachdruck nur mit ausdrücklicher Genehmigung

des Herausgebers bzw. der jeweiligen Autoren und mit Quellenangabe gestattet.

ISSN 1610-7233

Proceedings of the

2nd WORKSHOP OF THE

GI – FACHGRUPPE

“ONTOLOGIEN IN BIOMEDIZIN UND

LEBENSWISSENSCHAFTEN”

(OBML)

Mannheim, Germany
September 9-10, 2010

Group Website: <https://wiki.imise.uni-leipzig.de/Gruppen/OBML>

Organizers

Heinrich Herre	University of Leipzig
Robert Hoehndorf	European Bioinformatics Institute, Cambridge, UK
Janet Kelso	Max Planck Institute for Evolutionary Anthropology, Leipzig
Stefan Schulz	University Medical Center Freiburg

Local Organizers

Heiner Stuckenschmidt	University of Mannheim
Stefanie Keil	University of Mannheim
Johanna Völker	University of Mannheim

Keynote Speaker

Heiner Stuckenschmidt	University of Mannheim
-----------------------	------------------------

Guest Speaker

Erwin Tegtmeier	University of Mannheim
-----------------	------------------------

Program Committee

Sören Auer	University of Leipzig
Franz Baader	University of Dresden
Fred Freitas	University of Mannheim
Georgios Gkoutos	University of Cambridge, UK
Heinrich Herre	University of Leipzig
Robert Hoehndorf	European Bioinformatics Institute, Cambridge, UK
Josef Ingenerf	University of Lübeck
Janet Kelso	Max Planck Institute for Evolutionary Anthropology, Leipzig
Toralf Kirsten	University of Leipzig
Frank Loebe	University of Leipzig
Axel Ngonga-Ngomo	University of Leipzig
Dietrich Rebholz-Schuhmann	European Bioinformatics Institute, Cambridge, UK
Michael Schröder	University of Dresden
Stefan Schulz	University Medical Center Freiburg

Additional Reviewers

Dimitra Alexopoulou	University of Dresden
Philipp Frischmuth	University of Leipzig
Conrad Plake	University of Dresden
Thomas Wächter	University of Dresden
Amrapali Zaveri	University of Leipzig

Authors

Nico Adams	European Bioinformatics Institute, Cambridge, UK
Tomasz Adamusiak	European Bioinformatics Institute, Cambridge, UK
Sören Auer	University of Leipzig
Martin Boeker	University Medical Center Freiburg
Tony Burdett	European Bioinformatics Institute, Cambridge, UK
Patryk Burek	University of Leipzig
Cristian Cocos	St. Francis Xavier University, Canada
Timofey Ermilov	University of Leipzig
Ingmar Glauche	University of Dresden
Niels Grewe	University of Rostock
Silvia Groß	University of Leipzig
Michael Gruenberger	University of Cambridge, UK
Peter Haase	fluid Operations GmbH
Gesine Hansen	Hannover Medical School
John Hancock	MRC Harwell, UK
Norman Heino	University of Leipzig
Christian Hennig	Hannover Medical School
Heinrich Herre	University of Leipzig
Robert Hoehndorf	European Bioinformatics Institute, Cambridge, UK
Ludger Jansen	University of Rostock
Sergej Kireyev	University of Leipzig
Hiroaki Kitano	The Systems Biology Institute, Japan
	Sony Computer Science Laboratories, Japan
	Okinawa Institute of Science and Technology, Japan
Markus Löffler	University of Leipzig
Wendy MacCaul	St. Francis Xavier University, Canada
Ann-Marie Mallon	MRC Harwell, UK
James Malone	European Bioinformatics Institute, Cambridge, UK
Michael Martin	University of Leipzig
Tobias Mathäß	fluid Operations GmbH
Jörg Niggemann	CompuGroup Medical Software
Anika Oellrich	European Bioinformatics Institute, Cambridge, UK
Helen Parkinson	European Bioinformatics Institute, Cambridge, UK
Ricardo Pietrobon	Duke University, USA
Djamila Raufie	University Medical Center Freiburg
Dietrich Rebholz-Schuhmann	European Bioinformatics Institute, Cambridge, UK
Ingo Röder	University of Dresden
Johannes Röhl	University of Rostock
Nico Scherf	University of Leipzig
Daniel Schober	University Medical Center Freiburg
Paul Schofield	University of Cambridge, UK
Stefan Schulz	University Medical Center Freiburg
Hans Rudolf Straub	Semfinder AG, Kreuzlingen, Switzerland
Luca Toldo	Merck KGaA, Darmstadt
Ravensara S. Travillian	European Bioinformatics Institute, Cambridge, UK
Alexandr Uciteli	University of Leipzig
Amrapali Zaveri	University of Leipzig

Preliminary Program

as of August 31, 2010

THURSDAY, September 9, 2010

12:00 – 13:00 (Getting together / Registration / Coffee)
13:00 – 13:20 H. Herre
Welcome Remarks

Session 1 Medical Informatics and Semantic Web Chair: Janet Kelso

13:20 – 13:45 T. Mathäß SBML2SMW: Bridging System Biology with Semantic Web Technologies for Biomedical Knowledge Acquisition and Hypothesis Elicitation
13:45 – 14:10 C. Cocos An Ontological Implementation of a Role-Based Access Control Policy for Health Care Information
14:10 – 14:35 A. Zaveri Evaluating the Disparity between Active Areas of Biomedical Research and the Global Burden of Disease Employing Linked Data and Data-driven Discovery
14:35 – 15:00 A. Uciteli An Ontologically Founded Basic Architecture for Information Systems in Clinical and Epidemiological Research

15:00 – 15:20 COFFEE

Session 2 Formal Ontology Chair: Robert Hoehndorf

15:20 – 15:45 J. Niggemann Can a Hole Be Inflamed? On the Relation of Morphologic Abnormalities and Anatomical Cavities in SNOMED CT
15:45 – 16:10 J. Röhl Representing Dispositions
16:10 – 16:35 L. Jansen Grains, Components and Mixtures in Biomedical Ontologies

16:35 – 17:00 COFFEE

17:00 – 18:00 E. Tegtmeier *Guest Talk* : Series and Order (inquired, tentative title)

19:00 – ? DINNER

FRIDAY, September 10, 2010

Session 3 Bio-Ontologies Chair: Stefan Schulz

09:30 – 09:55 R. Travillian Anatomy Ontologies and Potential Users: Bridging the Gap
09:55 – 10:20 N. Adams The Ontology of Primary Immunodeficiency Diseases (PIDs) – Using PIDs to Rethink the Ontology of Phenotypes
10:20 – 10:45 A. Oellrich A Classification of Existing Phenotypical Representations and Methods for Improvement
10:45 – 11:10 H. Herre Towards a Cellular Genealogical Tree Ontology

11:10 – 11:30 COFFEE

Session 4 Knowledge Representation and Description Logics Chair: Heinrich Herre

11:30 – 11:55 S. Schulz Pre- and Postcoordination in Biomedical Ontologies
11:55 – 12:20 D. Schober Ontology Simplification: New Buzzword or Real Need?
12:20 – 12:45 N. Grewe A Generic Reification Strategy for n-ary Relations in DL

12:45 – 14:00 LUNCH

14:00 – 15:00 H. Stuckenschmidt *Keynote* : Ontology Matching: Past, Present, and Future
15:00 – 16:00 (Working Group) *Open Discussion*

Table of Contents

Medical Informatics and Semantic Web

	Paper ID and Pages	Page in PDF
SBML2SMW: Bridging System Biology with Semantic Web Technologies for Biomedical Knowledge Acquisition and Hypothesis Elicitation <i>Tobias Mathäß, Peter Haase, Hiroaki Kitano and Luca Toldo</i>	A 1-4	9
An Ontological Implementation of a Role-Based Access Control Policy for Health Care Information <i>Cristian Cocos and Wendy MacCaull</i>	B 1-5	13
Evaluating the Disparity between Active Areas of Biomedical Research and the Global Burden of Disease Employing Linked Data and Data-driven Discovery <i>Amrapali Zaveri, Ricardo Pietrobon, Timofey Ermilov, Michael Martin, Norman Heino and Sören Auer</i>	C 1-7	19
An Ontologically Founded Basic Architecture for Information Systems in Clinical and Epidemiological Research <i>Alexandr Uciteli, Silvia Groß, Sergej Kireyev and Heinrich Herre</i>	D 1-6	27

Formal Ontology

Can a Hole Be Inflamed? On the Relation of Morphologic Abnormalities and Anatomical Cavities in SNOMED CT <i>Jörg Niggemann, Hans Rudolf Straub and Heinrich Herre</i>	E 1-4	33
Representing Dispositions <i>Johannes Röhl and Ludger Jansen</i>	F 1-5	37
Grains, Components and Mixtures in Biomedical Ontologies <i>Ludger Jansen and Stefan Schulz</i>	G 1-4	43

Bio-Ontologies

Anatomy Ontologies and Potential Users: Bridging the Gap <i>Ravensara S. Travillian, Tomasz Adamusiak, Tony Burdett, Michael Gruenberger, John Hancock, Ann-Marie Mallon, James Malone, Paul Schofield and Helen Parkinson</i>	H 1-4	47
The Ontology of Primary Immunodeficiency Diseases (PIDs) – Using PIDs to Rethink the Ontology of Phenotypes <i>Nico Adams, Christian Hennig, Robert Hoehndorf, Anika Oellrich, Dietrich Rebholz-Schuhmann and Gesine Hansen</i>	I 1-4	51
A Classification of Existing Phenotypical Representations and Methods for Improvement <i>Anika Oellrich and Dietrich Rebholz-Schuhmann</i>	J 1-4	55
Towards a Cellular Genealogical Tree Ontology <i>Patryk Burek, Heinrich Herre, Ingo Röder, Ingmar Glauche, Nico Scherf and Markus Löffler</i>	K 1-5	59

Knowledge Representation and Description Logics

Pre- and Postcoordination in Biomedical Ontologies <i>Stefan Schulz, Daniel Schober, Djamila Raufie and Martin Boeker</i>	L 1-4	65
Ontology Simplification: New Buzzword or Real Need? <i>Daniel Schober and Martin Boeker</i>	M 1-4	69
A Generic Reification Strategy for n-ary Relations in DL <i>Niels Grewe</i>	N 1-5	73

SBML2SMW: bridging System Biology with semantic web technologies for biomedical knowledge acquisition and hypothesis elicitation

Tobias Mathäβ†, Peter Haase‡, Hiroaki Kitano*, Luca Toldo†

‡fluid Operations GmbH, Walldorf, Germany; *The Systems Biology Institute, Japan; Sony Computer Science Laboratories, Japan; Okinawa Institute of Science and Technology, Japan; †Merck KGaA, Darmstadt, Germany;

ABSTRACT

Motivation: Generation of biomedical hypotheses is a very important task in the pharmaceutical industry since it serves the whole drug development pipeline: which are the physiopathological mechanisms underlying a disease, which biomarkers shall be measured in a clinical trial in order then to be able to do proper data mining and then deliver best service to the patients, which unexpected events could occur if one inhibits a certain molecular pathway, which could be new indications for a given compound. These are only few of the many questions which require discovery of hidden links. In this work we describe our experiences and a small tool we developed and make freely available which bridges in bidirectional way semantic wiki and system biology technologies. This CellDesigner plugin therefore enables easy share and reuse of knowledge. Source code available at <http://code.google.com/p/sbml2smw/>

1 INTRODUCTION

The invention of a new drug therapy is a high-risk knowledge-intensive process which lasts often a dozen of years and involves a large number of people from different organisations. Different kinds of knowledge are required(1): **Marketing** knowledge is needed to put a business in perspective and deliver business plans, **Medical** knowledge is needed to specify physiopathological processes involved in the disease; **Genetic** knowledge is needed to identify biomarkers to stratify the patients; **Biological** knowledge is needed to identify pathways and molecular entities to specifically “target” by limiting unexpected adverse events; **Chemical** knowledge is needed to identify the appropriate scaffolds to exploit; and many more. The same variety of knowledge types is encountered when one has the task of proposing new indication for a given compound (e.g. Cladribine for Multiple Sclerosis, although originally developed for Hairy Cell Leukaemia)

Although semantic technologies have since long passed the academic stage and are well exploited in the industry (e.g. planning of elevator cablings; aviation industry; soft-

ware and IT service industry; etc) the pharmaceutical industry still is exploiting them only in “vertical” scenarios and almost only as terminology resources and not as knowledge modelling resources (e.g. Gene Ontology in target discovery; MeDDRA terminology for coding adverse events).

More recently, several vendors have approached the pharmaceutical industry exploiting very large knowledge networks and offer them either as technology (e.g. GeneStruct, BioWisdom SOFIA suite, Cellucidate Rule Studio), or as knowledge repositories (e.g. Ingenuity Pathway Assistant; GeneGO MetaCore; Biobase Knowledge Library; GVK Bioscience Biomarker knowledge base), or exploit them within consultancy services (e.g. BMSystems; Life Biosystems GmbH; BioWisdom).

The Semantic Media Wiki(2) technology allows very flexible creation of knowledge bases, and collaborative sharing for knowledge, and it comes at no costs. Several implementations of wiki exists in biology (e.g. wikigenes, wikiproteins, wikipathways) having the aim of creating a “scientific wiki” having an emphasis more on genes or on protein or on “networks”. At the same time, Payao system(3) enables more systematic community-based annotation and curation with SBML and SBGN compliance. The Project HALO is showing how low-cost highly-scalable modeling of basic scientific knowledge in biology could be appropriately handled with Semantic MediaWiki (SMW).

System biology scientists have the purpose of qualitatively or quantitative modelling the biology and eventually physiology of human beings, and for this reason acquire knowledge in form of networks, which then are studied quali-quantitatively, therefore would very much need using the SMW, through their tool of choice, CellDesigner(4). However, at the moment no “bridge” was possible to exploit these 2 technologies together.

In this work we describe SBML2SMW: a small plugin which combines freely available state-of-the-art software from System Biology and open source semantic wiki technology in order to:

- enable semantically enriched, distributed, biomedical knowledge acquisition
- share and reuse knowledge networks

†To whom correspondence should be addressed.

in the context of the pharmaceutical biomedical hypothesis services.

2 SEMANTIC WIKI: NOT ONLY FRONTEND

Wikis have been rapidly established as collaborative tool for sharing information, reducing the burdening of community creation to simply filling boxes with content. Semantic wikis are wiki “with an underlying model of the knowledge described in its pages”. Their current exploitation in biology has increased rapidly (NETTAB2010 conference) however their intrinsic formal power has been rarely reported in depth in the biomedical domain.

Semantic Media Wiki (SMW) is a free extension of MediaWiki that adds semantic annotations allowing therefore the wiki to function as a collaborative database with Semantic Web- tagged content. The use of SMW for R&D in Pharma has already been reported e.g. for the purpose of self-service portal to compile reports for drug lots(5), and in the “Pfizerpedia”: an internal tool used in Pfizer for tracking project, people and knowledge focused on patent information(6). In our work, we use SMW as core biomedical knowledge base, performing both knowledge acquisition and hypothesis generation. Therefore, we do not use it only as “front-end” for user content, but also as “back-end” for sorting and managing relations and entities that have been entered in the knowledge base by whatever mean (e.g. using the SMW import mechanisms and/or the CellDesigner “client”)

3 CELLDESIGNER: NOT ONLY CELL BIOLOGY

CellDesigner is the state-of-the-art structured diagram editor for drawing gene-regulatory and biochemical networks. Its intuitive user-interface helps drawing diagrams in rich graphical representation with personalized design. Networks are drawn based on a state transition diagram, proposed by Kitano and recent version comply with SBGN Process Description Diagram(7). Designed as a stand-alone tool, this powerful software is however network-aware, and therefore can connect to several major databases (DBGET, SGD, iHOP, Genome Network Platform, PubMed, Entrez Gene, SABIO-RK) and retrieve models from BioModels.net. The internal representation format that CellDesigner uses is the standard Systems Biology Markup Language (SBML) with CellDesigner specific annotation section to retain specific information needed for layout and other special features, and it has direct integration to the powerful Systems Biology Workbench (SBW) for performing quantitative simulations of cellular networks. Beyond system biology, CellDesigner can also be used for modelling system physiology(8) and it is projecting along those lines that we are exploiting it for improving our formal understanding on

physiopathological processes and rationalise drug discovery(1).

Although CellDesigner is publically available, it is not open source, however through extension tag specification and through the API it can be very much extended without the need of changing the main source code.

In this work, we report the extension of CellDesigner with the plugin “SBML2SMW” that enables a bidirectional exploitation of the semantic content of the SMW, and allows the use of CellDesigner as front-end for entering data in SMW.

4 SBML ONTOLOGY: ENABLING KNOWLEDGE REUSE

To bridge between the internal data representation of the CellDesigner and the Semantic MediaWiki, we have developed an ontology for the representation of the SBML knowledge models to be stored and reused.

The ontology has been modeled in OWL. It is intentionally kept small and concise, covering exactly those aspects of SBML models that are intended to be reused by others.

We identified the types of entities, all relations between these entities and all the plain information necessary to reconstruct the relevant parts of SBML models. A graphical representation of the resulting ontology can be reviewed in figure 1.

At the core of the ontology are the following four classes: Species, Reaction or Modification. We will describe these classes and the properties associated to these classes in further detail:

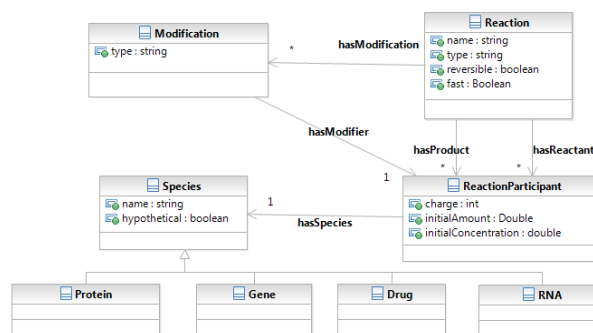


Figure 1: Overview of the SBML-ontology

- **Species:** A species is the top level type of all chemical compounds in a model. The species class has several subclasses, like Protein, Gene or Ion Channel, which allows a very detailed modeling and classification of these specieses.

- **ReactionParticipant:** Since the same Species (e.g. a certain protein) can be associated to several reactions, a further abstraction layer between the Species itself and the reaction it contributes to has to be introduced. A ReactionParticipant links a species to a certain reaction and assigns it its specific initial amount or initial concentration and other reaction-specific properties.
- **Reaction:** The reaction class links a set of species, its reactants, with another set of species, its products. For each reaction, further information like the fact if it is reversible or if any modifications to the reaction are in play, can be modeled.
- **Modification:** A reaction optionally can have one or more associated modifications, e.g. if this reaction demands the presence of a catalyst. Such a modification associates a set of modifiers to the modified reaction and contains information about the type of this modification.

5 SBML2SMW: BIDIRECTIONAL PLUGIN

SBML2SMW enables users of CellDesigner to persist arbitrary information from a graphical model in CellDesigner and to make it available for other users having access to the underlying Semantic MediaWiki. This makes it possible to reuse the facts from any model stored by any user in any other CellDesigner model. To achieve this, we use a species-centric mapping from models stored in SBML to the ontology-based representation described in section 4. Using this representation, we are not only able to load complete models saved by a user back into CellDesigner, but we also can load context-dependent information, e.g. all the reactions in the database a certain species participates on. The elements from the ontology-centric representation are then mapped to a SMW representation. Therefore, all the entities are mapped to SMW pages, the links between these entities are stored as semantic links between these pages. This way, we preserve the whole semantics of the SBML-ontology-instance.

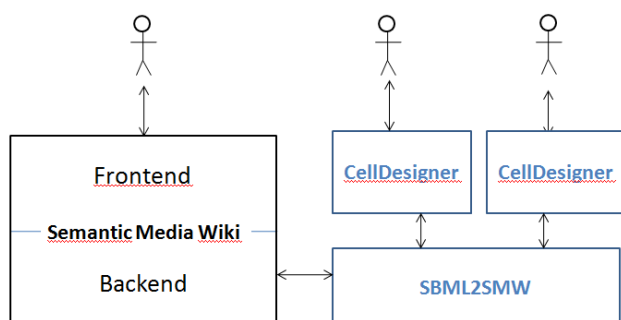


Figure 2: Conceptual Architecture

The users of this plugin should not only be able to use the store and load functionalities of the plugin completely transparently, but they also should be given the opportunity to add and correct information directly in the wiki.

Figure 2 shows the different ways of interaction our implementation supports. The SBML2SMW-plugin operates directly on the SMW backend, but data manipulation is also possible via the Webbrowser-based SMW frontend.

We now want to take a closer look on the functionality of the plugin from the user's perspective:

Store: The user first designs his model in CellDesigner by drawing all the important species, interlinking them to model the reactions taking place with their participation, setting the intended properties for all the species and adding modifications to the reactions. Next, he decides which parts of the whole model are worth storing into SMW and sharing with his colleagues. He selects these parts (i.e. the species and reactions) in CellDesigner, starts the SBML2SMW plugin and uses the store function. The selected parts of the model are extracted from the model, translated in a OWL representation and written to the SMW.

Load: If another user starts to create a model in CellDesigner, he now has the ability to add a Species, select it in CellDesigner, start the plugin and "expand" this species. The plugin accesses the SMW store, finds the selected species and retrieves all the stored information, i.e. all the reactions having this species as a reactant or a product. The plugin loads all these reactions, with all their other reactants and products and renders them into the CellDesigner window, together with all the associations between them.

Edit/Delete: Change or delete operations are not supported by using the plugin. We decided that if information was stored by any user, it has its legitimation to remain in the database, even if it has been modified in a certain model by any user. Deletion of facts has to be performed directly on the corresponding pages in the wiki.

Figure 3 shows the store and load operations in CellDesigner. Figure 3a) shows a model designed by user A, he decided to store a Protein X and a State Transition Reaction transforming it into another Protein Y.

In Figure 3b) we see how another user B added the Protein X to an empty model, and expanded this protein. The before stored State Transition Reaction is retrieved from the database and added to B's model

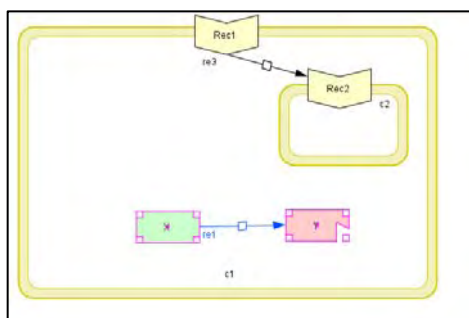


Figure 3a) CellDesigner model, X and Y selected to be saved

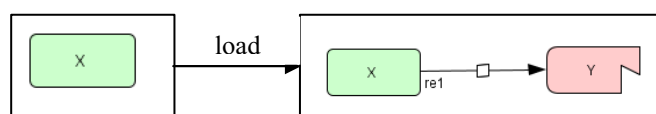


Figure 3b) before and after “load” on Protein X

6 CONCLUSIONS AND NEXT STEPS

In this work a free enabling technology is described, which extends the leading *free* system biology and system physiology CellDesigner platform by exploiting bi-directionally the leading *open source* Semantic Media Wiki technology. Knowledge can now be imported in the Semantic Media Wiki and made automatically available to the system modelers, and as-well they can share each of the relations they discover themselves (e.g. reading literature, or based on experiments or other observations), without having to interact with the SMW platform.

All the knowledge that is encoded in the CellDesigner pathways is now fully exposed to the SMW; and as well the whole knowledge contained in the SMW is now exposed to CellDesigner. This minor technological bridging thus removes any possible “knowledge gap” and maximises knowledge reuse both for an individual and for an organisation.

The technology we are making now freely available could be exploited through academic partners and have a public server sharing then properly formatted knowledge, thus seamlessly moving from a “wiki of science” to a “wiki of scientific knowledge”.

In spite of the appealing scenario that the work here reported is offering, we are aware that this report is at the moment only a development report. We are currently exploiting the technology here reported, and are confident in few months to be able to report on a detailed quantitative benefit/exploitation results.

ACKNOWLEDGEMENTS

We would like to thank Lina Yup Sonderegger (Merck Serono) for reviewing the manuscript, Abdul Mateen Rajput

for extensive use of the tool and debugging it in real-case applications.

REFERENCES

LINKS:

http://semantic-mediawiki.org/wiki/Semantic_MediaWiki
<http://sbw.kgi.edu/>
<http://www.projecthalo.com/>
<http://www.wikigenes.org/>
<http://www.wikiproteins.org/>
<http://www.nlm.nih.gov/research/umls/>
<http://www.cellucidate.com/>
<http://www.nettab.org/2010/>

ARTICLES

1. Kant CS, Ibberson MR, Scheer A. Building a disease knowledge environment to lay the foundations for in silico drug discovery and translational medicine. *Expert Opinion on Drug Discovery*. 2010;5(2):117-22.
2. ; [cited]; Available from: <http://www.mediawiki.org/wiki/MediaWiki>.
3. Matsuoka Y, Ghosh S, Kikuchi N, Kitano H. Payao: a community platform for SBML pathway model curation. *Bioinformatics*. 2010 May 15, 2010;26(10):1381-3.
4. Funahashi A, Matsuoka Y, Jouraku A, Morohashi M, Kikuchi N, Kitano H. CellDesigner 3.5: A Versatile Modeling Tool for Biochemical Networks. *Proceedings of the IEEE*. 2008;96(8):1254-65.
5. OntopriseGmbH. An SMW+-based self-service portal for R&D in Pharma. Karlsruhe; 2010 [updated 2010; cited 2010 15-06-2010]; Available from: http://smwforum.ontoprise.com/smwforum/index.php/R%26D_portal_in_pharma_industry.
6. Walsh D, Berridge A, Gardner B, editors. *Pfizer-pedia Patents Semantic MediaWiki - The How, What, When, Who and Why of patents*. The International Conference for Science & Business Information; 2009 18-21 October 2009; Spain. Infonortics.
7. Novere NL, Hucka M, Mi H, Moodie S, Schreiber F, Sorokin A, et al. The Systems Biology Graphical Notation. *Nat Biotech*. 2009;27(8):735-41.
8. Kitano H. Grand challenges in systems physiology. *Frontiers in Systems Physiology*. [Original Research Article]. 2010 2010-May-07;1.

An Ontological Implementation of a Role-Based Access Control Policy for Health Care Information

Cristian Cocos* and Wendy MacCaull ({ccocos,wmaccaul}@stfx.ca)

Centre for Logic and Information, St. Francis Xavier University, Nova Scotia, Canada

ABSTRACT

We provide a description of an access control ontology used in the context of a particular healthcare delivery framework. Its essential components are presented in detail, and its interaction with the system is highlighted.

1 INTRODUCTION

Researchers at the St. FX Centre for Logic and Information (CLI) are currently developing a workflow management system for palliative and seniors' care, which is scheduled to be deployed at several hospitals in rural Nova Scotia in the near future ([5]). This system aims at streamlining workflow by improving documentation and communication. In order to comply with the desideratum of patient-centered health care delivery, this endeavor requires a major commitment to ensuring data privacy and confidentiality. It is our belief that an ontology-structured knowledge base constitutes an ideal mechanism for the implementation/representation of a role-based access control (RBAC) policy. Here we present the design of such an RBAC ontology, which is intended to fulfill the commitment to privacy and confidentiality. While this paper may be viewed as having predominantly a practical value, several issues are of certain theoretical interest, most of which deal with our treatment of the .owl version of SNOMED-CT.

Representing the roles of an RBAC policy as classes of an access control ontology (ACO) emerges as a natural choice, especially when the ontology in question accommodates roles natively. We have, hence, decided to adopt an upper-level ontology where roles figure prominently in the asserted hierarchy; among the advantages of adopting a top-level ontology, the chief reason, in our view, is interoperability with other ontologies based on the same backbone. Basic Formal Ontology (BFO) has enjoyed a certain success in semantic web and ontology circles, and lends itself quite naturally to the type of enterprise that we have engaged in. From our perspective, four BFO classes are of immediate interest: BFO:role, BFO:object, BFO:process, and BFO:generically_dependent_continuant. For a presentation of BFO the reader is invited to consult [1] and [4]. Other terminological resources that occupy a central role in our

global venture are SNOMED-CT and the Intl. Classification for Nursing Practice (ICNP). Both standards are currently used extensively with the aim of building a palliative and seniors' care ontology (PSCO) that will guide and offer decision support for the palliative and senior's care workflow.

The workflow management system interacts with, among others, both ACO and PSCO in order to obtain access control clearance for the system user and, respectively, to guide the workflow in deciding between competing trajectories on the basis of palliative and seniors' care knowledge.

The paper is structured as follows: we first undertake a brief recount of the particular access control scenario to be implemented; this is followed by an overview of the access control ontology and its interaction with the workflow. We conclude by outlining some directions for future work.

2 ACCESS CONTROL SCENARIO

Two major types of entities ("resources") are subject to access control under the ACO scenario: informational items and actions. Database fields are paradigm examples of the former category (e.g., patient ID, patient name, primary diagnosis etc.), though reports also constitute informational content that carries access restrictions (e.g., incident reports). The latter class of entities has been assembled from actions implemented by the workflow management system, and represents actions that the system user can (or cannot) invoke, such as form/report printing, faxing, phoning, appointment scheduling, etc.

Corresponding to these two types of resources, two mechanisms have been employed, so that users' access control credentials will be checked at login time, and consequently some of these information fields and/or actions will be cleared for access. Here are the defining features of the access control policy implemented:

- (1) Roles are organized hierarchically, hence permissions granted at higher levels of the hierarchy are inherited by roles lower in the hierarchy ("hierarchical RBAC" in [3]);
- (2) Resources are organized hierarchically—e.g., allowing access to a whole form means allowing access to all of its fields;

* To whom correspondence should be addressed.

- (3) Access control constraints can be provided for each form field and action individually;
- (4) Unlike a purely linear hierarchy, we have disjoint roles, whereby not all roles intersect with respect to the range of permissions;
- (5) Database fields can be accessed as both *read only* and *write*;
- (6) System users may have multiple roles with regard to the same patient; in such circumstances, the clearance level assigned is a union of permissions for all roles involved.

3 IMPLEMENTATION

The BFO:role branch contains the main mechanism of our RBAC implementation, whereby roles relevant from the point of view of access control in a palliative/seniors care setting have been grouped on categories representing clearance and permission levels for informational items and actions respectively. As mentioned above, the “role” branch is intended to accommodate roles such as “Patient,” “Caregiver,” “Medical Doctor” etc. Most of the classes that populate this branch have been imported straight from SNOMED-CT: we thus saw fit to import SNOMED’s “Occupation (occupation)” (SCTID_14679004) branch in its entirety, as it contains a significant portion of relevant roles. Interestingly enough, SNOMED does not comprise a properly titled “role” class, nor does it contain any class to this effect, hence most of the classes that we have assembled under the “role” branch have been culled from various (and quite heterogeneous) places in SNOMED—mostly “occupation” and “person.” As such, our endeavor can also be viewed as a SNOMED-CT restructuring enterprise by bringing it in line with BFO—an undertaking that, as of this writing, is being contemplated under the auspices of the IHTSDO (the SNOMED-CT custodian). HealthCareRole (see figure 1 below) is another ACO-specific class that groups several SNOMED-CT classes such as “Caregiver (person),” “Clinical trial participant (person)” etc. The SNOMED “Relative (person)” class (SCTID_125677006) comprises the usual relative roles.

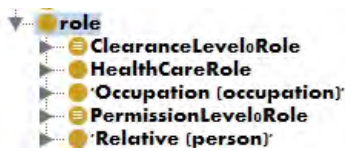


Fig. 1: Roles in ACO.

The core ACO mechanisms reside under the ClearanceLevel0Role and PermissionLevel0Role branches. These defined classes contain five, respectively two, defined (sub)classes labeled ClearanceLevelxRole, ClearanceLevelxwRole and, respectively, PermissionLevelyRole (figure 2), where x takes values from 1 to 4, and y and w , 1 and 2. (The number of clearance and permission levels is subject to

future amendments.) Due to spatial considerations we will refrain from elaborating the labeling scheme, though the labeling principles should emerge clearly from the illustration below (figure 2). We will detail the content of the clearance level hierarchy, and occasionally point out the differences between it and the permission level hierarchy.

Each of the ClearanceLevelxRole classes is defined as a union of roles that have a certain security clearance level, which is to be dictated by real-life security and privacy considerations. ClearanceLevel2Role, for example, is a union of the following SNOMED_CT classes: “Physiotherapist/occupational therapist (occupation),” “Social worker (occupation),” “Community nurse (occupation),” “Pain management specialist (occupation)” and “Pharmacist (occupation)” (see figure 3).

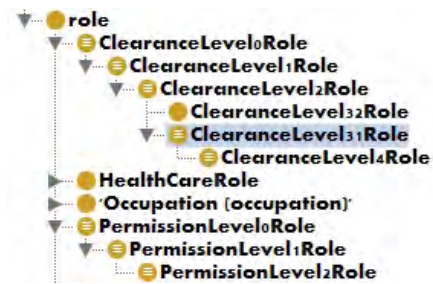


Fig. 2: ACO clearance and permission levels.



Fig. 3: ClearanceLevel2Role Definition.

A corollary of this particular chain-like structure is that all ClearanceLevelyRole roles have also clearance level x for $x \geq y$, in other words, all information that is accessible to a community nurse (say) is also accessible to a clinical oncologist (a ClearanceLevel0Role role), but not vice versa; in general, a parent inherits the clearance level of its children. Furthermore, figure 2 shows that the ClearanceLevel2Role class has two children: ClearanceLevel32Role and ClearanceLevel31Role, which have been stipulated as disjoint, such that the two actually constitute a partition of ClearanceLevel2Role. A similar story goes for permission levels (see also figure 2), and while we have not designed any disjoint permission levels, these can be captured in a manner completely analogous to clearance levels.

We further added classes that represent actual information and actions that are subject to access control, and devised a method to tie clearance and permission level roles with that information. We have thus opted to add six relations (“object properties”) together with reciprocal (“inverse”) relations:

- (1) **hasWriteAccessTo** and **hasReadAccessTo** are relations from “role” to “DatabaseField” (see below), and, respectively, to “Report (record artifact)” or

DatabaseField.” The former is a child of the latter, hence whoever has write access to some resource obviously has read access to it. Their reciprocals are **writeAccessibleTo** and **readAccessibleTo**, with the former being a child of the latter;

- (2) **invokableBy** is a relation from ACO:SystemProcedure to BFO:role, and plays the same role as either of **read/writeAccessibleTo** vis-à-vis action permissions;
- (3) **hasRole** is a relation whose range is BFO:role, and is intended to tie roles with their bearers; its reciprocal is **roleOf**;
- (4) **hasClearanceLevel** ties the SNOMED-CT class of “Homo sapiens (organism)”—a subclass of BFO:object—with a clearance level. Its reciprocal relation is **clearanceLevelOf**;
- (5) **permissionLevelOf** connects permission levels with elements of “Homo sapiens (organism).”

The classes that represent controlled *information* are children of the BFO:generically_dependent_continuant, which is the BFO entity designed to account for informational and other abstract entities. ACO imports the Informational Artifact Ontology (IAO) under the BFO:generically_dependent_continuant class, and entities of the IAO are used to define ACO’s metadata fields. The root of the IAO is “information content entity” (IAO_0000030). We have added the following chain of ACO-specific subclasses as a child of IAO_0000030: “Proposition,” “DeclarativeProposition,” and “DatabaseField” (see figure 4 below). The type of information that makes the subject of access control policies has been, at this point, limited to community palliative and seniors’ care records. Our focus has been to extract information fields from Guysborough Antigonish Strait Health Authority’s (GASHA) palliative and seniors’ care programs. The GASHA workflow collects patient information using several forms. Given the provisional status of most of the forms as of this writing, we chose to represent in the ACO taxonomy six of the more stable ones—see the children of DatabaseField in figure 4—while the rest of them will be added as they reach a reasonably stable status. The children of each form class are the fields that make up the form, such that, from an access control point of view, each form is represented as a class of fields.

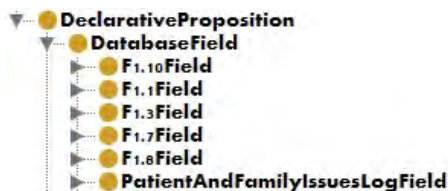


Fig. 4: Database fields.

Similarly, classes that represent controlled *actions* comprise the ACO:SystemProcedure class, which is a child of BFO:process (figure 5).

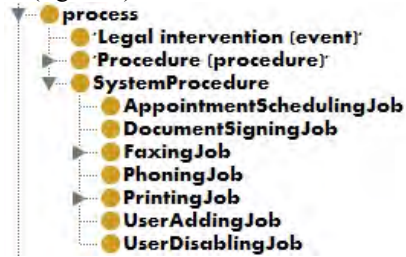


Fig. 5: System procedures

Finally, all the classes that make the subject of access control policies have been endowed with restrictions that outline their corresponding clearance/permission level. Here are a few examples:

The HospitalizationHistoryField (ACO_0000101), a child of the F1.8Field class (i.e., a field on GASHA form F1.8) has been designed so as to *not* be accessible to a clearance level 2 role—which would obviously make it inaccessible to any role in a higher clearance level category (i.e., 31, 32 and 4) as well. Figure 6 shows this:



Fig. 6: Hospitalization History field restrictions.

F1.10 form will not be accessible to a clearance level 3 role, which means that all fields on this form will bear the same restriction. This is a case where the restriction has been implemented at the parent class level (F1.10Field (ACO_0000129)), instead of adding it to all component fields individually (see figure 7).



Fig. 7: Form F1.10 restrictions.

PhoningJob will only be invokable by a PermissionLevel2Role:



Fig 8: PhoningJob restrictions.

The last of the relevant BFO classes is BFO:object, whose only direct child is the “Organism (organism)” SNOMED class (SCTID_410607006). This one, in turn, contains “Homo sapiens (organism)” as an only subclass

(SCTID_337915000), which represents the main ACO role-bearer.

ACO currently contains approximately 10,000 classes, and its level of DL expressivity is *SROLF*, which is N2ExpTime-hard ([7]).

4 INTERACTION WITH THE WORKFLOW

According to the basic interaction scenario, access control clearance is checked at login time, by querying ACO upon user login.¹ ACO provides in return a list of GASHA form fields and reports whose access is forbidden to the user, and a list of system actions permitted. The workflow system—that, by design, is in possession of the list of all actions and informational items—acts accordingly, by blocking access to the requisite actions and information entities.

An example of a DL query is the following: say Individual₁ is an instance of the “Homo sapiens (organism)” class that has been defined as a psychotherapist—i.e., is an instance of “**hasRole** some 'Psychotherapist (occupation)'.” A query such as “not (**accessibleTo** some (**roleOf** value Individual₁))” will reveal all the form fields that are *not* accessible to Individual₁—that is, it will return a list of inaccessible field *names*. Another way of defining an individual is to assign it a clearance level outright, without going through the detour of specifying its role, thus sparing the reasoner the effort of figuring out the individual’s clearance level. If Individual₂ is defined as “**hasClearanceLevel** some ClearanceLevel0Role,” the query “not (**accessibleTo** some (**roleOf** value Individual₂))” will return all the (names of the) form fields not accessible to Individual₂.

A few words about the interaction with the PCSO: this will be done in tandem with the patient database (EHR), as the ontology only contains generic information, but no information pertaining to individual patients. PCSO will provide logic-based guidance for the workflow at the so-called decision points, which are to be suitably chosen from those points in the workflow where it branches, and where palliative and seniors’ care knowledge is involved in the decision. The sequence of actions is the following: (a) the workflow reaches a decision point; (b) the ontology is queried with the patient data contained in the EHR, and furnishes information regarding the workflow branch that the process is supposed to follow; (c) the process follows the path indicated by the ontology query.

5 CONCLUSIONS AND FUTURE WORK

We have outlined the access control policies and implementation principles that lay the foundation of the RBAC ontol-

ogy built as part of a palliative and seniors’ care workflow project presently developed in Nova Scotia. Currently in pilot phase, this project makes essential use of semantic web techniques, and constitutes a living expression of the efficacy of ontological knowledge bases and their employment in concrete everyday situations, and of the usefulness of standards such as SNOMED-CT. It is our belief that semantically-structured knowledge has yet to bear its fruit as a large-scale implementational venue for health informatics systems, and we view our present endeavor as a contribution towards streamlining efforts targeting massive adoption of such techniques.

The evolution of the interaction between the workflow system and the ACO includes a customization phase, which requires implementing a workflow mechanism that queries the patient/client on specific access control preferences during several predetermined phases of the workflow, plus a mechanism that builds new patient-specific access control ontologies that will be combined with the default ACO described above in order to customize the access control policy for each patient. Further, more sophisticated, access control scenarios may require implementing task-based access control policies; this transcends the scope of role-based access control, and may require tools more expressive than OWL-DL. Among others, we envisage implementing an emergency “break the glass” mechanism, according to which certain system users can bypass security policies in case of emergency.

From a purely theoretical point of view, our approach in building the ACO can be catalogued as a special case of the “Roles as Classes” strategy detailed in [3], though our focus has been quite categorically implementational, in contrast to the latter’s heavy theoretical bent, that ultimately aims at assessing the suitability of OWL for RBAC *and beyond*.

ACKNOWLEDGEMENTS

Our industry partners, Markus Latzel and Bryan Kramer of Palomino System Innovations Inc., have provided valuable feedback and support.

REFERENCES

- [1] Bittner, T. and Smith, B. (2004) Normalizing Medical Ontologies using Basic Formal Ontology, in *Kooperative Versorgung, Vernetzte Forschung, Ubiquitäre Information* (Proceedings of GMDS Innsbruck, 26-30 September 2004), Niebüll: Videel OHG, pp. 199–201.
- [2] Bouamrane, M.-M., Rector A. and Hurrell, M. (2009) A Hybrid Architecture for a Preoperative Decision Support System Using a Rule Engine and a Reasoner on a Clinical Ontology, in Polleres, A. and Swift, T. (Eds.): RR 2009, LNCS 5837, pp. 242–253, Springer-Verlag Berlin Heidelberg 2009.
- [3] Finin, T. et al. (2008), ROWLBAC - Representing Role Based Access Control in OWL, in SACMAT’08, June 11–13, 2008, Estes Park, Colorado, USA.

¹ Note that the ontology is not meant to enforce the access of individual users to the system; it only controls the type of information accessible to various user roles—hence in order to be in position to query the ontology, one must already be logged into the system, which can be achieved only if one has declared his/her credentials at the time of login.

- [4] Grenon, P., Smith, B. and Goldberg, L. (2004) Biodynamic Ontology: Applying BFO in the Biomedical Domain. In D. M. Pisanelli (ed.), *Ontologies in Medicine*, Amsterdam: IOS Press, 2004, pp. 20–38.
- [5] Miller, K. and MacCaull, W. (2009) Toward Web-based Careflow Management Systems, *Journal of Emerging Technologies in Web Intelligence*, vol. 1, no. 2, pp. 137-145.
- [6] Tsoumas, B., Dritsas, S. and Gritzalis, D. (2005) An Ontology-Based Approach to Information Systems Security Management in V. Gorodetsky et al. (eds.): *MMM-ACNS 2005*, LNCS 3685, Springer-Verlag Berlin Heidelberg, pp. 151 – 164, 2005.
- [7] Kazakov, Y. (2008) *SRIQ* and *SR_QIQ* are Harder than *SH_QIQ*, in Baader, F. et al. (eds.), *DL 2008*. Vol. 353 of *CEUR Workshop Proceedings*.

Evaluating the disparity between active areas of biomedical research and the global burden of disease employing Linked Data and data-driven discovery

Amrapali Zaveri[†], Ricardo Pietrobon[‡], Timofey Ermilov[†], Michael Martin[†], Norman Heino[†] and Sören Auer[†]

[†]Universität Leipzig, Institut für Informatik, Johannsgasse 26,D-04103 Leipzig, Germany,
{lastname}@informatik.uni-leipzig.de
<http://aksw.org>

[‡]Duke University, Durham, NC, USA
rpietro@duke.edu

ABSTRACT

Although biomedical research has brought substantial benefit to people all over the world, by dramatically improving their life expectancy and the quality of life, the distribution of this benefit is not equitable. An important contributor to this is the current absence of accurate, interlinked data and information that enables a precise description of the degree of inequality between current efforts in biomedical research and global health care needs. In this position paper we present an approach for evaluating this disparity, which involves converting and inter-linking relevant datasets into Linked Data, and analyzing them to represent the disparity as a visual map. We identify different data sets, relevant for answering the research question. Since bio-medical statistical data is of paramount importance in this data integration project, we describe a tool and methodology of representing such bio-medical statistical data as RDF. We perform an preliminary integration of the datasets and outline how prospective queries can be formulated. We conclude by discussing the limitations of the approach taking the current bio-medical data curation landscape into account.

1 INTRODUCTION

According to the World Health Organization (WHO)¹, more than one billion people (i.e. one sixth of the world's population) suffer from one or more neglected tropical diseases [11]. This shows a significant imbalance between the research intensity invested for the investigation of certain diseases and their prevalence. Although much of this discrepancy can be attributed to the market driven aspects of biomedical research conducted by pharmaceutical companies, it is also caused by the lack of knowledge with regard to the current status and intensity of this disparity. An important contributor to this lack of information is the current absence of accurate, interlinked data and information, that enables a precise description of the degree of inequality between current efforts in biomedical research and global health care needs. In this position paper we describe our plan to evaluate the disparity between active areas of biomedical research and the global burden of disease through the use of Linked Data and data-driven discovery. After giving a brief overview over the research context in Section 2, we describe our envisioned methodology in Section 3 and describe

the process of representing bio-medical statistical data as RDF in Section 4. We describe the related work in Section 5, mention the limitations of this study in Section 6 and conclude in Section 7.

2 RESEARCH CONTEXT

In this section we describe three important aspects that are to be considered, in general, while dealing with interlinking and publishing new datasets on the Semantic Web. These are: (a) biomedical data publishing and discovery; (b) knowledge interlinking and fusion and (c) assessment of data quality.

2.1 Biomedical data publishing and discovery

There are numerous websites^{2,3,4,5,6,7} and governmental efforts [22, 21] which contain important information on health-care but this information is mainly published in textual and non or only partially machine-readable formats. Although these efforts bring together relevant resources to either retrieve citations about health-care disparities or display the statistics as charts or maps, the data is not linked to other data sources. They also do not address the research and health-care disparity issue, which would help to evaluate the intensity of the disparity for a particular condition, for example, the trial or the research going on in that area.

2.2 Knowledge Interlinking and Fusion

Interlinking occurs in the literature under a dozen of terms [1] such as Deduplication [10], Entity Identification [19], Record Linkage [15] and many more. Encountered problems are generally caused by data heterogeneity [7]. The processes of data cleaning [13] and data scrubbing [27] are common terms for resolving such identity resolution problems. As a new challenge, the Linked Data paradigm provides the means necessary to skip the data preparation step as they have already proliferated a shared

¹ <http://www.who.int>

² <http://www.ncbi.nlm.nih.gov/sites/entrez?db=pubmed>

³ <http://www.ahrq.gov/data/dataresources.htm>

⁴ http://phpartners.org/health/_stats.html

⁵ <http://minorityhealth.hhs.gov/templates/browse.aspx?lvl=2&lvlid=5>

⁶ <http://ims.missouri.edu/moims2008/step1AOI.aspx>

⁷ <http://apps.who.int/globalatlas/>

structural representation of data. DBpedia and other knowledge bases maintained by the AKSW research group⁸ are available as a crystallization point for new datasets [3]. Combined with the proposed quality assessment and interlinking approaches, this allows lowering the access barrier for new open biomedical datasets to evolve into interlinked knowledge bases and join the network ecosystem. In a recent survey on Data Fusion [4], the semantic heterogeneity is considered as the greatest challenge for data integration and fusion, data fusion being the last step of a Data Integration process (preceded by schema mapping and duplicate detection) [20]. While ontologies or thesauri already play a major role when integrating several sources to overcome the semantic heterogeneity [28], the problem of creating a complete, concise and consistent integration has not been sufficiently addressed. In the context of the emerging Web of Data, new challenges include:

- On-the-fly integration with a priori unknown data based on the discovered schema,
- Consideration of provenance and trust based on provided metadata,
- Creation of specific metrics to merge structured data based on the given vocabulary and data quality,
- Handling of inconsistencies in structured data (a key difference here is the defined semantics of ontology relations compared to e.g. relations in databases).

2.3 Assessment of data quality

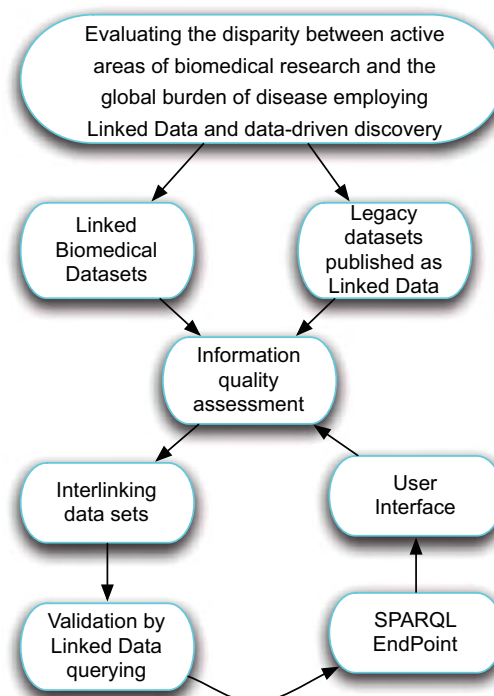
Assessing the quality of data is essential due to the multiple and autonomous data sources that can be linked, which affects data accessibility and usability [12]. The dimensions of data quality commonly found in the literature include accuracy, consistency, timeliness, completeness, relevancy, inter-operability and trustworthiness. Ensuring high quality data requires two kinds of processing: (a) data validation and (b) consistency validation on knowledge fragments. However, values may often be missing in biomedical data due to several reasons such as lost or corrupted samples, patients not showing up for scheduled appointments or failing of measuring instruments. If there are too many missing values, ignoring them may invalidate the analysis that is performed [8]. One of the remedies is to fill in the gaps with suitable replacements such as either (a) fixed values or (b) existing values at random or (c) averaging neighboring values. There are numerous efforts employed for quality checking such as (a) data mining techniques, (b) comparing data on the web versus a gold standard, (c) using provenance information about the data on the web to assess the quality and trust-worthiness [18], (d) using a metrics for quality assessment [23] and others. Thus, assessing data quality is one of the important prerequisites for publishing linked data on the web.

3 METHODOLOGY

The project described in this position paper will be conducted in two cycles. In the first cycle, we will explore the relevant datasets to be used, to identify properties that can be used to interlink them. As a next step, those datasets which are available in unstructured or semi-structured formats are converted into RDF.

⁸ <http://www.aksw.org>

Fig. 1. Overview of the methodology



After that we will assess the quality of the datasets for completeness, conciseness and consistency. The next step will be to explore different mapping approaches to interlink the datasets. We then validate the interlinking by creating suitable queries and verifying the query results. In order to make the integrated data available and easily navigable for users, we create a user interface and make the data available via a SPARQL endpoint. We will re-iterate all these steps in the second cycle to account for new or improved datasets that may be introduced after the first cycle was started. Figure 1 summarizes these steps in a nutshell. The details of the steps are described in detail in the following sections.

3.1 Datasets – Identification and Conversion

As the first step, we identified three main datasets that are relevant for our purpose, namely (1) *ClinicalTrials.gov* (2) *PubMed* and (3) WHO's *Global Health Observatory* (GHO). *LinkedCT*⁹, is the Linked Data resource of *ClinicalTrials.gov*. It contains information about 61,920 governmentally and privately funded clinical trials conducted around the world (amounting 9.8 million triples). *PubMed*¹⁰ is a service of the US National Library of Medicine that includes bibliographic information and abstracts of over 19 million publications from MEDLINE and other life science journals. It covers the fields of medicine, nursing, dentistry, veterinary medicine, the health care system and preclinical sciences. *Bio2RDF*¹¹ is a mashup of different bio-medical knowledge bases aiming to facilitate the creation of bioinformatics information

⁹ <http://linkedct.org/>

¹⁰ <http://www.ncbi.nlm.nih.gov/pubmed>

¹¹ <http://www.bio2rdf.org>

systems. It also contains data from PubMed in RDF with about 797 million triples. As a result, both LinkedCT and PubMed (via Bio2RDF) are available as SPARQL endpoints^{12,13}. The GHO dataset¹⁴, on the other hand, contains statistical information regarding the global burden of disease and is available as Excel sheets. In order to convert the GHO dataset into RDF, we used the SCOVO vocabulary [16], which is a vocabulary particularly designed to describe statistical information. We developed a plug-in in OntoWiki [2] to facilitate us in converting CSV data from GHO into RDF. OntoWiki is a tool which supports collaborative creation, maintenance and (Linked Data) publication of RDF knowledge bases. The description of the plug-in and the SCOVO vocabulary is described in Section 4.

3.2 Interlinking Datasets

Our next task was to interlink the datasets with regard to the three main concepts: (1) diseases, (2) publications and (3) countries. Figure 2 shows the links between the three datasets. The property linking the classes is The thick black lines with arrows on both sides represent the link between the different classes. Some of which are already present such as `owl:sameAs` links for publications between LinkedCT and PubMed with 42,219 links¹⁵. However, the links between GHO and PubMed and LinkedCT for diseases and countries need to be established. For this purpose, we used Silk 2.0 [26], a tool for discovering relationships between data items within different Linked Data sources. The Silk framework provides a declarative language for specifying (1) the types of RDF links that should be discovered between data sources and (2) the conditions which the data items must fulfill in order to be interlinked. The details of the interlinking results are displayed in Table 1. The string similarity was matched using the *Jaro distance* metric. The maximum and minimum threshold used for all the matching tasks was 0.95 and the filter threshold used was 0.90. Thus, the accepted links are those with the threshold value between 0.95 and 1.0 and the to-be-verified links are those with the threshold value between 0.90 and 0.95.

3.3 Knowledge-base Creation and Querying

We will create an integrated knowledge-base from the interlinked datasets as well as a SPARQL endpoint that will enable users to query the knowledge-base. We will then validate the interlinking by creating suitable queries and verifying the query results. In order to illustrate this process we created a prototypical knowledge-base by integrating some sample data. Examples of the envisioned queries are shown below:

(1) Find all diseases in India with an incidence rate higher than 70%.

```
PREFIX who: <http://who.int/>
PREFIX ct: <http://data.linkedct.org/resource/linkedct/>
PREFIX pubmed:<http://bio2rdf.org/ns/mesh#>

SELECT DISTINCT ?disease ?incidence ?country
WHERE {
  ?x      who:country      "India" .
  ?x      who:incidence    ?incidence .
  ?x      who:disease      ?disease .
```

¹² <http://data.linkedct.org/snorql/index.html>

¹³ <http://mesh.bio2rdf.org/sparql>

¹⁴ <http://www.who.int/gho/en/index.html>

¹⁵ <http://esw.w3.org/HCLSIG/LODD/Interlinking>

```
FILTER(?incidence>70)
}
```

Listing 1. SPARQL query retrieving diseases with incidences in India

This query described in Listing 1 returns a list of diseases in India along with their incidence. When executing this query with the integrated knowledge-base, we might find that Tuberculosis is the disease with the highest incidence in India. For the following queries we will use “Tuberculosis” as our disease of interest.

(2) Find all the countries, along with the number, of clinical trials conducted for Tuberculosis.

```
SELECT DISTINCT ?disease ?country ?noOfTrials
WHERE {
  ?disease  who:disease  "Tuberculosis" .
  ?y        ct:disease   ?disease .
  ?y        ct:noOfTrials ?noOfTrials .
  ?y        ct:country   ?country .
}
```

Listing 2. SPARQL query retrieving countries and number of trials trials for Tuberculosis

This query described in Listing 2 returns a list of all countries along with the number of clinical trials for Tuberculosis. When executing this query, we might find that in India there are fewer trials conducted for Tuberculosis than in the USA.

(3) Find the country with the most number of publications related to clinical trials for Tuberculosis.

```
SELECT ?country COUNT(?reference)
WHERE {
  ?disease  who:disease  "Tuberculosis" .
  ?z        ct:disease   ?disease .
  ?z        ct:country   ?country .
  ?z        pubmed:reference ?reference .
}
GROUP BY ?country
```

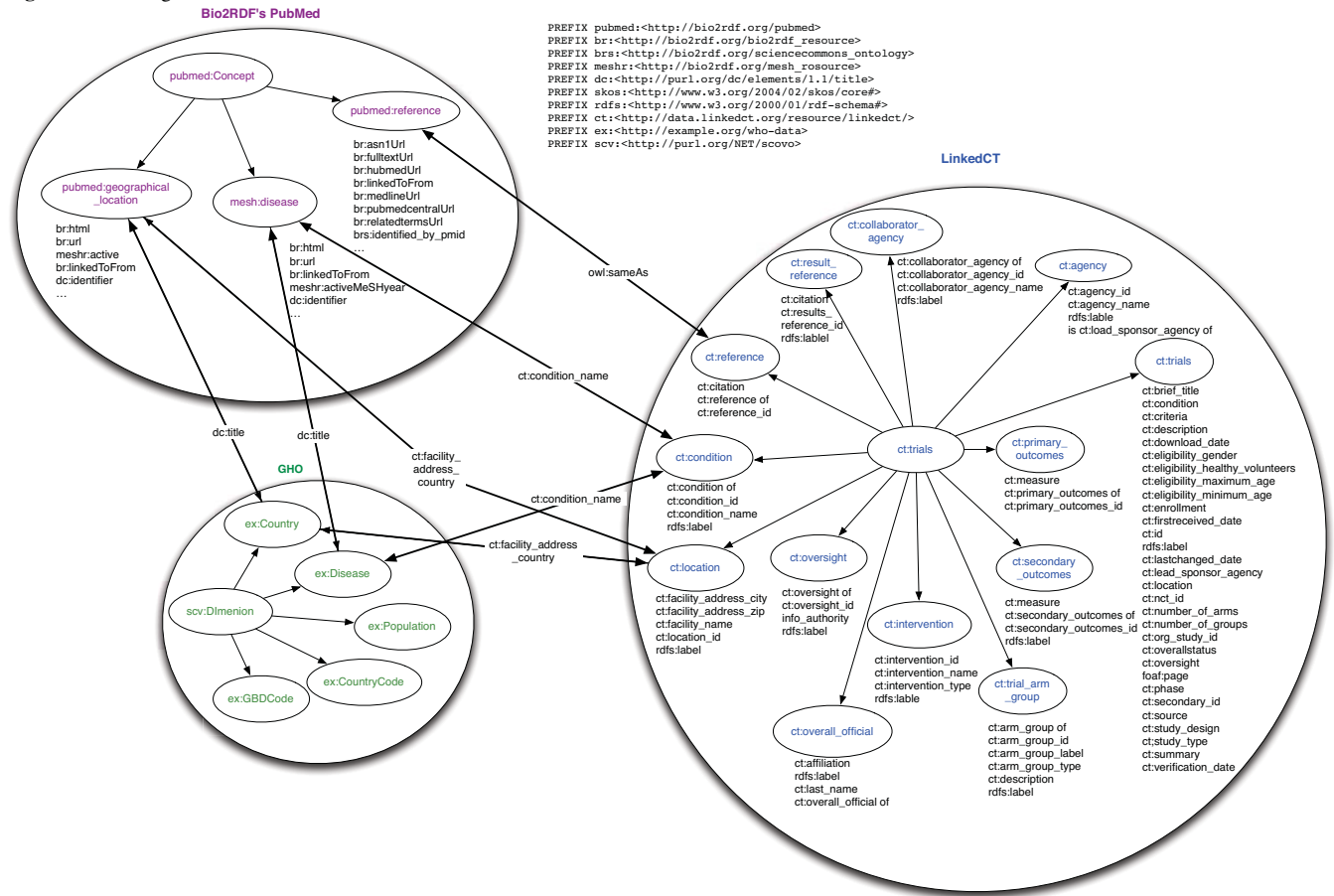
Listing 3. SPARQL query retrieving publications and countries for Tuberculosis

This query returns a list of publications along with the countries. When executing this query, we might find that the most number of publications for Tuberculosis are in Africa. Once we are able to execute these queries on the integrated knowledge-base, this might reveal that the country where the disease is most prevalent, the country where the most number of clinical trials are conducted and the country where the most research is performed are different. This accounts for the disparity. However, the evaluation of health-care disparity should take a number of other aspects into account as well (which we did not yet consider in the examples above). These aspects include, for example, mortality factor, economic costs, number of years of healthy life lost to a disease, the correlation with the amount of money spent for research (cf. [24]). Taking all these different aspects into account to truly evaluate the disparity is the major challenge of our project.

3.4 Creating User Interface

In order to make the integrated data available on the internet and easily navigable, we will create a user interface. We will employ the HTTP mechanism called content negotiation in order to cater clients such as web browsers and search engines to display the information they request in a human-readable format. The user

Fig. 2. Interlinking datasets



Datasets A ↔ B	Link ID	Instances of A	Instances of B	Accepted Links	Links to be verified
PubMed ↔ LinkedCT	Disease	23618	5000	1240	183
WHO ↔ PubMed	Disease	134	23618	201	12
WHO ↔ LinkedCT	Disease	134	5000	19	13
WHO ↔ PubMed	Country	192	23618	201	12
PubMed ↔ LinkedCT	Country	23618	5000	4999	0
WHO ↔ LinkedCT	Country	192	5000	4993	0

Table 1. Number of links between datasets

interface will provide many ways of exploring the integrated data, for example, as maps, texts, links and graphs. The data will be displayed as icons on a world map so that potential users, such as health-care researchers and health-care policy makers, are easily able to evaluate the disparity. Additional information about each resource will be displayed as text and links to relevant pages. There will also be an option to render statistical information as graphs. The proposed user interface is shown in Figure 3.

4 RDF REPRESENTATION OF BIOMEDICAL STATISTICAL DATA

Biomedical statistical data is often represented by describing a single data item in several dimensions. A simple row-based transformation of the said data into RDF is thus not applicable. SCOVO (Statistical Core Vocabulary), introduced in [16], is designed particularly to represent multidimensional statistical data using RDF. It consists of the following three central concepts:

- *Dataset*: A dataset represents the container of the data, such as a table holding data in its cells.

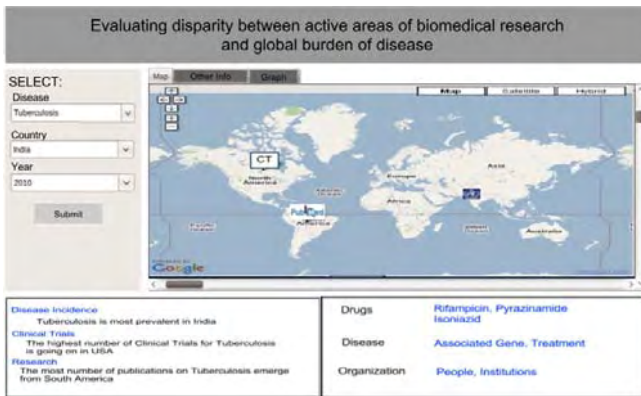


Fig. 3. Mock User Interface to display health-care disparity on a map

- *Data Item*: A data item represents a single piece of data, such as a cell in a table.
- *Dimension*: A dimension represents the unit of a single piece of data, such as a time period, location or a disease.

A statistical dataset in SCOVO is represented by the class *Dataset*. Single data items belong to a certain dataset with its dimensions—which are instances of the *Dimension* class—attached through the *dimension* property.

The data in GHO is published as Excel sheets, which can be converted and stored as Comma Separated Values (CSV). In the CSV format, it is not possible to query the data efficiently. Also, interlinking this statistical information with the other RDF datasets is difficult. Thus, in order to transform the CSV data into RDF, we used the SCOVO vocabulary. However, generating such a transformation in a fully automated way is not feasible, since publication formats often contain implicit assumptions that have to be discovered by humans. For example, dimensions are often encoded in the heading or label of a sheet or figures may be given as a fraction of 1000 so as to save space. Therefore, we developed a semi-automatic approach that facilitates the aforementioned transformation. We integrated the algorithms as a plug-in extension into OntoWiki, the semantic collaboration platform developed by our research group [2]. In the following example we will illustrate how data from GHO is converted into RDF using OntoWiki. As illustrated in Figure 4, the parsed CSV files are represented as HTML tables. This representation of the data gives the users the ability to configure (1) dimensions by manually creating them and selecting all elements belonging to a certain dimension and (2) the range of statistical items according to the dimensions. It is also possible to save and reuse these configurations for other CSV files, which are of the same structure (e.g. for data in consecutive years). On the basis of this configuration, the algorithm converts the data into RDF as displayed exemplarily for *item 127* in Listing 4.

```
prefix ex:<http://example.org/who-data>.
prefix scv:<http://purl.org/NET/scovo>.

ex:Country rdfs:subClassOf scv:Dimension;
rdf:type rdfs:Class;
dc:title "Country" .
```

```
ex:Disease rdfs:subClassOf scv:Dimension;
rdf:type rdfs:Class;
dc:title "Disease" .

ex:Afghanistan rdf:type ex:Country;
dc:title "Afghanistan" .

ex:Tuberculosis rdf:type ex:Disease;
dc:title "Tuberculosis" .

ex:c1-r6 rdf:type scv:Item;
rdf:value 127;
scv:dimension ex:Afghanistan;
scv:dimension ex:Tuberculosis .
```

Listing 4. RDF representation of data using SCOVO

The GHO table “*Estimated total female DALYs, by cause and WHO Member State in 2004*” contains 5 dimensions and 22384 statistical data items. After converting the whole data an RDF model containing 152000 triples was created. However, there may be some Excel sheets that contain taxonomies only readable by humans. It may also not be feasible to convert these automatically. In addition to ontology engineering tasks, OntoWiki provides ontology evolution functionality, which can be used to further transform the newly converted data. Furthermore, OntoWiki provides various interfaces (in particular Linked Data and SPARQL interfaces) to publish and query RDF data.

5 RELATED WORK

Calculating the health-care disparity allows research policy makers to determine whether the research strategy established by National Innovation Systems is aligned with the nation’s healthcare needs. For example, the Agency for Healthcare Research and Quality (AHRQ)¹⁶ in the US publishes two annual National Healthcare Disparities reports¹⁷. However, these reports focus on only a few demographic groups and a limited number of clinical conditions. Even though there are evidence that suggest that the investments in health-care research have been cost-effective over the past decades, the Commission on Health Research for Development highlighted the great imbalance between these investments and the global burden of disease in 1990 [14]. The US government also reported the lack of a perfect correlation between the disease burden and amount of funding allocated for diseases [25]. Although this is a concern in the US, there exists a much more severe problem in developing countries which is a great risk for poor, adolescents, and women [9]. However, due to the lack of a reliable observatory, the disparities cannot be monitored and thus still persist. One of the methods to measure this disparity is to perform cross-sectional studies comparing estimates of disease-specific funding with data on the different measures of the burden of disease [6]. Other methods are to use various statistical measures on samples of data to quantify the disparity [5]. The methods to calculate the disparity are not only cumbersome and time consuming but also are limited due to the fact that they use only a sample of the data for the analysis. In contrast, our approach aims to employ Linked Data to provide a real-time observatory for the health-care research disparity. All the data will be available for analysis along with different means and ways to evaluate the disparity, for global data. Moreover, by making the

¹⁶ <http://www.ahrq.gov>

¹⁷ available at <http://www.ahrq.gov/qual/qdr09.htm>

The screenshot shows a configuration window for importing CSV data. It features a table with columns for WHO Country codes and rows for GBD causes. Callouts highlight specific dimensions and ranges within the table.

WHO Country code	3010	4005	1010	4008	1020	2010	2020	4007	5020	4010	4012	2030	3020	3025
GBD cause (b)														
Tuberculosis	127	0	7	0	40	0	6	2	0	-	-	0	0	672
STDs excluding HIV	52	1	75	0	44	0	29	2	3	1	7	0	0	258
a. Syphilis			6	0	12	0	2	0	0	0	0	0	0	24
b. Chlamydia			38	0	16	0	17	2	3	1	5	0	0	146
c. Gonorrhoea	15	0	30	0	15	0	9	0	0	0	1	0	0	85
HIV/AIDS	0	0	5	0	163	0	16	0	0	1	0	2	0	1
Diarrhoeal diseases	1,206	3	97	0	866	0	23	3	3	1	33	0	0	1,117
Childhood-cluster diseases	170	0	19	0	108	0	1	0	0	0	0	0	0	365
a. Pertussis	110	0	16	0	77	0	1	0	0	0	0	0	0	62
b. Poliomyelitis	0	-	0	-	0	-	0	-	0	0	0	-	-	-
c. Diphtheria	2	0	0	-	2	0	0	0	-	-	-	0	0	2
d. Measles	6	0	2	0	16	0	0	0	0	0	0	0	0	236

Fig. 4. Excerpt of the configuration table used for converting “Estimated total female DALYs, by cause and WHO Member State in 2004” from GHO

data available in machine readable formats and with the possibility of further linking to other data sets, calculations will be less time consuming and more reliable.

6 LIMITATIONS

This section discusses the different limitations of our project. We have specifically mentioned these limitations to sensitize the biomedical community about the issues one might encounter while dealing with large biomedical datasets.

- **Information quality.** The information quality of ClinicalTrials.gov is relatively low [17] as critical information from trial registration, such as study contact, trial end date, and primary outcome are not consistently reported. Also, all trials are not published and indexed in PubMed or even if they are published, the respective citation is missing in ClinicalTrials.gov. On the other hand, although the data in PubMed is relatively complete, the RDF version needs improvement as there are links to non-RDF URI’s for most of the resources. In case of the GHO data, the information quality is reliable but there are issues concerning the data coverage, as described below.
- **Coverage.** When integrating data and performing analysis on the integrated information, the coverage of the base data is important and subsequent querying and analysing the integrated data has to take limited coverage into account. In our case, the GHO data is currently only available till the year 2004. Therefore, we used data only till 2004 from the other two datasets, so as to get meaningful results.

- **Interlinking quality.** The number of interlinks obtained between the datasets for diseases was still relatively low, as is illustrated in Table 1. This was because the datasets do not contain standardized identifiers for naming diseases. For example, “AIDS” in LinkedCT could not be matched with “Acquired Immunodeficiency Syndrome” in PubMed using basic string similarity.
- **Propagation of errors.** Since two of the datasets are of relatively low quality, the interlinked dataset is greatly affected. Also, since the coverage of data in WHO is specific for a certain time period, important information from the other two data sets may be left out while calculating results. This amounts to the propagation of errors in the interlinked datasets and thus affects the final results.

These different limitations of a semi-automatic information integration approach as described in this paper can not be addressed by one party individually, but should be taken into account, when publishing new versions of the datasets. In particular the use of shared identifiers (i.e. IRIs) would improve the data integration greatly.

7 CONCLUSION

As a next step, we will use different methods to improve the interlinking quality between the datasets. Also, as and when new data is published, we will update our knowledge-base. The resultant knowledge-base will contain adequate information to answer the following research questions important for health-care research:

- Which disease has the highest percentage of health-care disparity with respect to the burden of disease and the clinical trials conducted in a particular country?
- As a research policy maker, which research area would it be most beneficial to allocate funds?
- Who are the key people doing most research for a particular disease?
- What has been the trend, over time, for the health-care disparity for a particular region?

Thus, our project will help in evaluating the disparity between active areas of biomedical research and the global burden of diseases all over the world.

REFERENCES

- [1]Thor A. *Automatische Mapping-Verarbeitung auf Webdaten*. PhD thesis, Universität Leipzig, 2008.
- [2]Sören Auer, Sebastian Dietzold, and Thomas Riechert. Ontowiki - a tool for social, semantic collaboration. In 736 749, editor, *The Semantic Web - ISWC 2006, 5th International Semantic Web Conference*. Springer, 2006.
- [3]Christian Bizer, Jens Lehmann, Georgi Kobilarov, Sören Auer, Christian Becker, Richard Cyganiak, and Sebastian Hellmann. Dbpedia - a crystallization point for the web of data. *Web Semantics, Scienc, Services and Agents on the World Wide Web*, 7(3):154–165, September 2009.
- [4]Jens Bleiholder and Felix Naumann. Data fusion. *ACM Comput. Surv.*, 41(1):1–41, 2008.
- [5]Arthur J. Bonito, Celia R. Eicheldinger, and Nancy F. Lenfestey. Health disparities: Measuring health care use and access for racial/ethnic populations. Technical report, RTI International, 2005.
- [6]Gerard F. Anderson Cary P. Gross and Neil R. Powe. The relation between funding by the ntaional institute of health and the burden of diseases. *The New England Journal of Medicine*, 340(1881-1887), 1999.
- [7]Abhirup Chatterjee and Arie Segev. Data manipulation in heterogeneous databases. *SIGMOD Rec.*, 20(4):64–68, 1991.
- [8]Dianne Cook and Deborah F. Swayne. *Interactive and dynamic graphics for data analysis*. Springer, 2007.
- [9]Barbara Crossette. Disparities in health: Inequities create great risks for poor, adolescents, and women in developing countries. Technical report, Disease Control Priorities Project, August 2006.
- [10]A. Culotta A., McCallum. Joint deduplication of multiple record types in relational data. In *14th ACM international conference on Information and knowledge managemen*, 2005.
- [11]Boraschi D, Abebe Ma nd Aseffa A, Chiodi F, and Chisi J. Immunity against hiv/aids, malaria, and tuberculosis during co-infections with neglected infectious diseases: Recommendations for the european union research priorities. *PLoS Neglected Tropical Diseases*, 2(6), June 2008.
- [12]Monica Scannapieco Elisa Bertino, Andrea Maurino. *Data Quality in the Internet Era*, volume 14, chapter 4, pages 11 – 13. IEEE Computer Society, July 2010.
- [13]Ahmed K. Elmagarmid, Panagiotis G. Ipeirotis, and Vassilios S. Verykios. Duplicate record detection: A survey. *IEEE Transactions on Knowledge and Data Engineering*, 19:1–16, 2007.
- [14]John R. Evans. Essential national health research - a key to equity in development. *N Engl J Med*, 323:913 – 915, September 1990.
- [15]Ivan P. Fellegi and Alan B. Sunter. A theory for record linkage. *Journal of the American Statistical Association*, 64(328):1183–1210, December 1969.
- [16]Michael Hausenblas, Wolfgang Halb, Yves Raimond, Lee Feigenbaum, and Danny Ayers. Scovo: Using statistics on the web of data. In *ESWC*, pages 708–722, 2009.
- [17]Ross JS, Mulvey GK, Hines EM, Nissen SE, and Krumholz HM. Trial publication after registration in clinicaltrials.gov: A cross-sectional analysis. *PLoS Med*, 6(9), 2009.
- [18]Knoesis. *Using Web Data Provenance for Quality Assessment*, 2009.
- [19]Prabhakar S Lim E P, Srivastava J and Richardson J. Entity identification in database integration. In *Proceedings of the Ninth International Conference on Data Engineering*, 1993.
- [20]Bleiholder J. Naumann F., Bilke A. and Weis M. Data fusion in three steps: Resolving schema, tuple and value inconsistencies. *IEEE Data Eng. Bull.*, 29(2):21–31, 2006.
- [21]Public Health Information Development Unit. *The value of linked data for policy development, strategic planning, clinical practice and public health: An Australian perspective*. PHIDU, 2003.
- [22]Public Health Information Development Unit. *The value of linked data for policy development, strategic planning, clinical practice and public health: An international perspective*. PHIDU, 2003.
- [23]Smith L C Stvilia B, Twidale M B and Gasser L. Assessing information quality of a community-based encyclopedia. In *In Proceedings of the International Confernece on Information Quality*, pages 442 – 454, Cambridge, MA, 2005.
- [24]Harold Varmus. Evaluating the burden of disease and spending the research dollars of the national institutes of health. Editorial, June 1999.
- [25]Harold Varmus. Funding allocation for disease research. Statement, National Institutes of Health, May 1999.
- [26]Juilius Volz, Christian Bizer, Martin Gaedke, and Georgi Kobilarov. Discovering and maintaining links on the web of data. In *ISWC*, 2009.
- [27]Jennifer Wisdom. Research problems in data warehousing. In *CIKM '95: Proceedings of the fourth international conference on Information and knowledge management*, pages 25 – 30. ACM, 1995.
- [28]Patrick Ziegler and Klaus R. Dittrich. Three decades of data integration - all problems solved? In *In 18th IFIP World Computer Congress (WCC), Building the Information Society*, volume 12, pages 3 – 12, 2004.

An Ontologically Founded Basic Architecture for Information Systems in Clinical and Epidemiological Research

Alexandr Uciteli, Silvia Groß, Sergej Kireyev, Heinrich Herre*

IMISE, University of Leipzig

ABSTRACT AND MOTIVATION

This paper reports on an ontologically founded basic architecture for information systems which are intended to capture, represent, structure, apply, and maintain metadata for various domains of clinical and epidemiological research. The specification of data and their documentation and application in clinical and epidemiological study projects represents a significant expense in the project preparation and has a relevant impact on the value and quality of these studies.

An ontological foundation of an information system provides, among others, a precise semantics for the entities which are represented in this system. This semantics should be grounded, according to our approach, on a suitable top-level ontology. Such an ontological foundation leads to a deeper understanding of the entities of the domain under consideration and contributes to a rigorous formal semantics.

The intended information systems will be applied to the field of clinical and epidemiological research and will provide, depending on the application context, a variety of functionalities. Core functions include the capturing, storing, representing, and retrieval of data of several kinds (items, phenotypes, phenes, complex properties, et al.). Further envisaged functionalities are based on reasoning, as, for example, support for diagnosis finding and other medical decision support.

In the present paper we restrict to the basic architecture which might be common to all such information systems. A prototype implementation of this architecture realizes, at present, the most fundamental functionalities for such information systems, including the capturing of data, the structured representation of items, and their retrieval.

1 INTRODUCTION

The specification of data collection and documentation in clinical and epidemiological study projects, as implemented in many clinical trial centers, as well as in ZKS Leipzig ([7]) or in the LIFE project ([8]) represents a significant

expense in the project preparation. A precise definition of the documentation features has an important impact on the value and quality of these studies. The use of internationally established features by referencing existing external standards (eg CDISC [13], LOINC [11], ICD-10 [12]) has proved to be useful in diverse applications.

Clinical research will be supported thereby in the re-use of such features. For this purpose, study items must be captured and specified in semantically correct way, and computationally presented such that they can be efficiently retrieved and to be re-used. It turns out that the precision of the notion of a study item is a difficult task and that a broad, ontologically oriented, view is needed to capture, structure and present the very complex data and information in the field of medicine and biomedicine.

In the present paper we expound the basic structure of an architecture for information systems intended to capture, present, structure, and specify meta-data of the domain of clinical research. This architecture is founded on a top level ontology, since we defend the approach that a top level ontology may provide a well-founded semantics for the entities (items, data, phenotypes etc.) under consideration. An ontological semantics - based on top-level or a core ontology - can be established by principles of ontological reduction [2]. The simplest case of an ontological reduction is a mapping into a top-level or core ontology. An ontological analysis reduces and reconstructs the notions of the considered domain in the framework of the categories and relations provided by a top-level ontology.

There are several ISO-standards providing systems of categories and relations as a framework for specifying data. According to our strategy, these standards must be taken into consideration. On the other hand, these standards have, usually, an insufficient semantics. Hence, our ontological approach makes a further step to an ontologically founded semantics for such standardized systems, and thus, build a bridge between the rigor of formal ontology and the semantic fuzziness of thesaurus-style metadata representation. In the frame of our project, we decided to take the ISO/IEC 11179 standard ([6]) as our initial system because it is tailored to the description and representation of data. This standard is specified by a generic system of categories and relations which must be, usually, adapted to the particular domain under consideration. This adaption

* alexander.uciteli@imise.uni-leipzig.de

leads to the introduction of additional categories and relations.

2 BASIC PRINCIPLES OF THE ISO/IEC 11179 STANDARD

2.1 Overview

The standard ISO/IEC 11179 ([6]) - Metadata registries (MDR)- addresses the semantics of data, the representation of data, and the registration of the descriptions of that data. It is through these descriptions that an accurate understanding of the semantics and a useful depiction of the data are found. The purposes of ISO/IEC 11179 are to promote the following:

- Standard description of data
- Common understanding of data across organizational elements and between organizations
- Re-use and standardization of data over time, space, and applications
- Harmonization and standardization of data within an organization and across organizations
- Management of the components of data
- Re-use of the components of data

Understanding data is fundamental for its design, harmonization, standardization, use, re-use, and interchange. The underlying model for an MDR is designed to capture all the basic constituents of the semantics of data, independent of any application or subject matter area.

2.2 Basic Elements of the ISO/IEC 11179 standard

At the beginning, we summarize and analyze the basic elements, the basic building blocks of the ISO/IEC 11179 standard.([6]) The most basic entities are *data element*, *data element concept*, *conceptual domain* and *value domain*. We review the definitions of the main entities as expounded in the standard.

A *data element* (DE) contains two main parts: a semantic one (called *data element concept* / DEC) and a syntactic one (*value domain* / VD). A DEC may be separated in two components: the *object class*, which is a set of ideas, abstractions, or things in the real world that can be identified with explicit boundaries and meaning and whose property and behavior follow the same rules; and a *characteristic*, which can be attributed to the members of the object class. A data element concept will be associated with exact one *conceptual domain* (CD). Conceptual domains come in two (non-exclusive) sub-types. An *enumerated conceptual domain* is specified as a list of *value meanings*. Value meanings are possible values for the characteristic. A *described conceptual domain* is specified by a description. Conceptual domains will be represented by *value domains*. A value domain is a set of permissible

values, which consists of a value meaning (element of the representing CD) and this value meaning representing value.

3 ONTOLOGICAL FOUNDATION OF THE ISO/IEC 11179 STANDARD

This ISO standard claims that it provides a semantics for data, though, it turns out that the concepts introduced are lacking a real semantics. The purpose of the current section is to show - by analyzing a selection of examples - how a deeper founded and more precise semantics can be achieved by using a suitable top-level ontology.

3.1 Categories in GFO

We are using for our ontological analysis the top-level ontology *General Formal Ontology* (GFO) ([1]). In contradistinction to other top-level ontologies, for example, DOLCE ([9]) or BFO ([10]), the ontology GFO provides an ontology for categories. Distinctive features of this part of GFO are the following: GFO allows for different kinds of categories, among them, concepts, symbolic structures, and universals. Furthermore, it allows for categories of higher order. DOLCE does not include, at the present stage, a sufficiently developed ontology of concepts/categories, whereas in BFO concepts are explicitly excluded. The investigation of the structure of categories, in particular of concepts, exhibits a bridge to the field of cognitive science and cognitive psychology. The GFO ontology of concepts/categories was already exploited in several investigations ([3],[4]).

3.2 Architecture of the MDR and the Metamodel

In the present paper we propose an architecture for a metadata repository which is intended to be used for clinical and epidemiological research. The architecture of the Metadata Repository (MDR) consists of three levels with associated modules (Fig. 1). The *abstract level* includes the ISO/IEC 11179 standard, but at the same time, it extends it and presents an ontological foundation by the top-level ontology GFO. Furthermore, the notions of an item, of a property and others are analyzed and described at this level. Finally, we proved that the full ISO standard can be, essentially, embedded / mapped into GFO. For example, the category of data elements is a subcategory of the GFO meta-category *category* and the category of items is a subcategory of the category of data elements.

In other words, every concrete item (instance of the item category) will be interpreted as particular data element and at the same time as particular GFO category. This mapping is useful because it provides a precise semantics for the ISO standard which itself contains, from ontological point of view, only a small fragment of the GFO framework. The module of the abstract level presents the metamodel. The

level of elementary concepts presents the concepts associated to the items. These concepts are instances of the meta-model categories from the abstract level like Data Element Concept, Object Class, Characteristic (in GFO – property or, more general, attributives) or Item itself. The third level is presented by the level of complex concepts. The complex concepts are constructed from elementary ones by different relations.

Example (Fig 1):

abstract level: ISO:DEC is a subcategory of GFO:Category

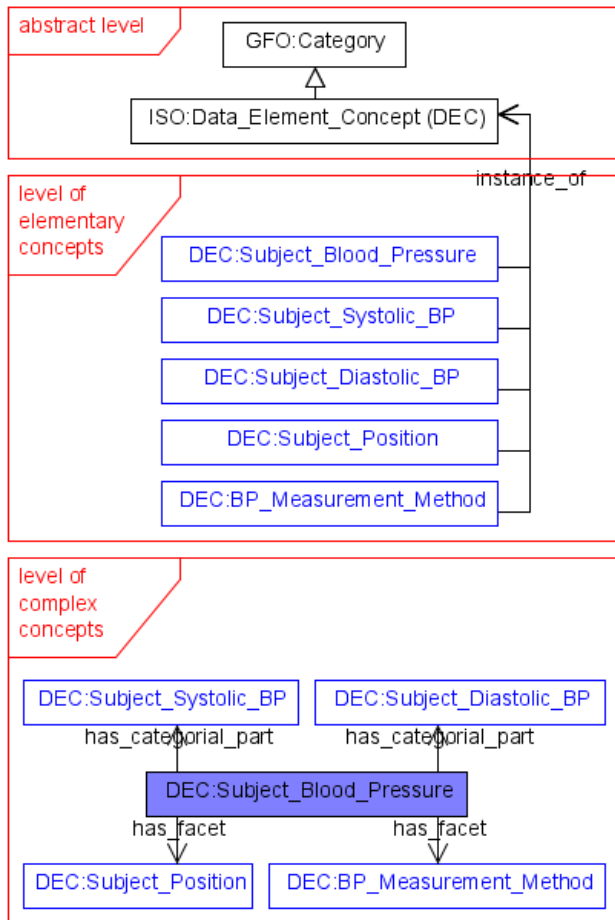


Fig. 1 Architecture (example)

level of elementary concepts: Concrete DEC's will be defined (as instances of ISO:DEC):
 DEC:Subject_Blood_Pressure, DEC:Subject_Systolic_BP,

level of complex concepts: The concepts of the second level will be combined to complex concepts by various applicable relations (The complex concepts have components like categorial parts or facets). For example, the complex concept DEC:Subject_Blood_Pressure has two

categorial parts and two facets. Subsequently, we demonstrate our method by few examples, in particular, we investigate and analyze the notion of a data element in more detail. The whole/full meta-model exhibits an analysis of many other categories of the ISO standard, and provides further categories which are not yet present in the standard. A data element has two constituents/components, a semantic one, called data element concept, and a syntactic one, called value domain (Fig 2). Since the data element itself is a category in GFO we must clarify how its components relate to the whole. For this purpose we introduce a suitable part-of relation, called constituent-part.

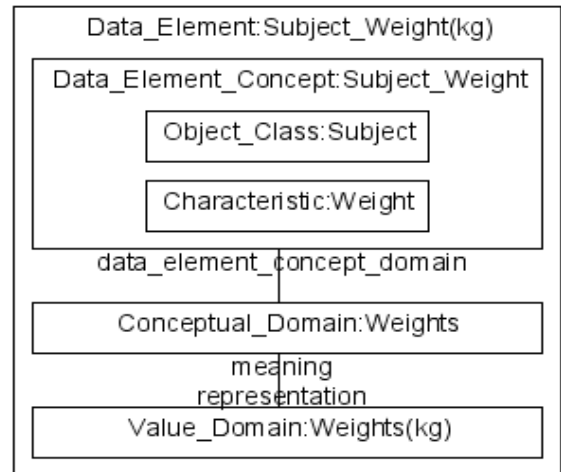


Fig. 2 Data Element

Then, data elements are GFO-categories with certain constituents (the category of data elements is a subcategory of GFO category).

A data element concept DEC includes an object class ObC(DEC) and a property P(DEC). An object class is a category whose instances are entities of the real world. The property, being a constituent of the data element concept (DEC), can be attributed to the instances of the class ObC, i.e. every instance of ObC(DEC) has the property P(DEC). To a data element concept there is associated a uniquely determined conceptual domain. This conceptual domain is a set of entities which serve as the values meanings of the property P. At this place, we must clarify what value meanings of a property are. Let us consider as an example the property weight, denoted by W. The instances of this property are qualities, being individual properties that inhere in objects which are instances of ObC. We may

is considered as an observable characteristic or trait of an organism, such as its morphology, development, biochemical or physiological properties, behavior, function, and products of behavior (such as a bird's nest). Such a single phenotype can be considered – in the framework of GFO – as a property, i.e., a category whose instances are qualities, inhering in individual spatio-temporal organism. We call such properties phenes, a term used in [20], in analogy to the notion of gene. Another term is *primitive phenotype*, which we use equivalently for the term phene. A *complex phenotype* is composed of a set of primitive phenotypes., analogously, we introduce the notion of *complex phene*. Furthermore, we generalize these notions to arbitrary spatio-temporal entities, for example, processes, diseases etc. The set of all phenes of an entity is then called the phenome or the *complete phenotype* of it.

There is yet another important notion which is related to idealizations. An anatomical atlas of a human being, for example, displays idealizations of the organs and of the whole human organism. Such an idealization is based on a selection of *essential* primitive phenotypes which are composed to an ideal entity, which is independent of space and time. We call this composition of properties a *phenomenal archetype*. It is an open question for which sets of entities a phenomenal archetype can be defined and what kind of an ontological status it has. This problem is related to an old, mainly unsolved, problem in philosophy [21].

In the framework of our basis architecture the primitive phenotypes or phenes relate to the elementary concepts of the second level, whereas the complex phenotypes or complex phenes are presented on the level of complex properties. The representation of complex phenotypes need further ontological relations which glue together the primitive phenotypes/phenes of the second level. For the representation of phenomenal archetypes we need further means to define idealized wholes. We believe that the solution of this problem must use results and theories of Gestalttheory and cognitive psychology.

5 IMPLEMENTATION

For tests of the theoretical model we implemented a software prototype whose first version allows to create and to manage all fundamental concepts of the ISO standard such as Conceptual Domain, Value Domain, Data Element Concept, Data Element and so on. Moreover, some extensions of the standard like Items, Item Groups and Item Variants were implemented. The graphical representation of the concepts is also possible. But the most important function is that the simple concepts can be connected by suitable relations to complex ones (see *level of complex concepts* in Fig. 1). These relations can come, e.g., from top-level ontologies (in our case from GFO).

The software is web-based and uses a relational database. The first version is developed in PHP with MySQL database connection. For the further developments is also a RDF Triple Store considered..

6 DISCUSSION AND CONCLUSION

We outlined basic ideas and results about an ontologically founded basic architecture for information systems to be applied in the field of clinical and epidemiological research. The information systems, based on this architecture, are usually related to particular domains, or classes of domains, and are aimed at a variety of applications. The principles, set forth in this paper, can be applied also to other domains, which are different from the field of clinical or epidemiological research. What is common to all of these information systems is the basic functionality to capture, structure, present and retrieve meta-data associated to a domain of interest. The ISO-11179-standard is a valuable initial system for such a basic architecture. Though, it turns out that the semantics of this standard is insufficiently developed.

According to our approach, we used a top-level ontology - in the present paper the ontology GFO- to elaborate an ontological foundation which establishes a more precise and fine-grained semantics for the ISO-standard. Hence, in this way we established a bridge between the standard and the rigorous methods of formal ontology. The ideas of this paper can be applied to arbitrary standards, and were, actually, already partially used in [15] for a semantic underpinning of the Unified Modelling Language (UML). Finally, we implemented a prototype for this architecture which realizes the functionality to capture and present items, and to retrieve items taken from clinical studies which were captured and stored in the system.. The development of this tool will be continued to include more functions, for example, reasoning capabilities.

A standard for a meta-model, usually, provides a system of categories and relations to be used to specify, to structure and to organize conceptualizations for a class of domains. It turns out, that standards exhibit interesting features which can be used, conversely, to support the development of ontologies of categories/concepts in the frame of a top-level ontology. Hence, the investigation of ISO standards contributes to a fruitful inter-relation between top-level ontologies and standards of meta-models. On the one hand, we achieve new insights in the structure of categories which can be included in top-level ontologies, on the other hand, a top-level ontology, based on stable logical methods, may contribute to a better founded and formal semantics for this standard. Hence, any of these fields of research may benefit from the results and methods of the other.

There is a number of promising open problems and tasks for further research. One of the most important tasks is the

ontological analysis of the notion of phenotype and related concepts, as phene, phenome, and phenomenal archetype. On the representational side, the introduction of new relations is relevant, which may be used to compose primitive phenotypes to complex ones. The envisaged information systems may present ontologies of phenotypes for various fields of clinical research.

Reasoning capabilities open a wide field of applications. Methods from artificial intelligence, deduction and non-monotonic reasoning could be used to develop decision support systems and diagnostic expert systems. It turns out, that new forms of non-monotonic reasoning are needed to achieve adequate reasoning procedures. These forms of non-monotonic reasoning are especially related to the reasoning with phenomenal archetypes.

ACKNOWLEDGEMENTS

The research of this paper is inspired by the BMBF-project “Specification and prototype implementation of a metadata repository for clinical and epidemiological research in Germany (MDR)”. We thank three referees for their valuable remarks that contribute to the quality of the paper.

REFERENCES

- [1] Herre, H. (2010) General Formal Ontology (GFO): A Foundational Ontology for Conceptual Modelling. In: Poli, R. Obrst, L. (ed.) *Theory and Applications of Ontology*. Vol. 2, Berlin: Springer, July 2010.
- [2] Herre, H., Heller, B. (2006) Semantic Foundations of Medical Information Systems Based on Top-Level Ontologies. *Journal of Knowledge-Based Systems* Vol. 19(2), pp. 107-115.
- [3] Hoehndorf, R., Loebe, F., Poli, R., Herre, H. Kelso, J. (2008). GFO-Bio: A biological core ontology. *Applied Ontology* Vol. 3(4), pp. 219-227.
- [4] Hoehndorf, R., Loebe, F., Bacher, J., Backhaus, M., Gregorio Jr., S., Pruefer, K., Visagie, J., Uciteli, A., Herre, H., Kelso, J. (2009) BOWIKI: an ontology-based wiki for annotation of data and integration of knowledge in biology. *BMC-Bioinformatics* Vol. 10 (Suppl 5):S5.
- [5] Loebe, F. (2007) Abstract vs social roles: Towards a general theoretical account or roles. *Applied Ontology*, vol. 2, n.2., p. 127158, 200.
- [6] ISO/IEC 11179 standard. (2010) Information technology — Metadata registries (MDR) — Part 3: Registry metamodel and basic attributes (http://jtc1sc32.org/doc/N1951-2000/32N1983Ta-Text-for-ballot-FCD_11179-3.pdf).
- [7] Zentrum für klinische Studien Leipzig (ZKS:Leipzig) (<http://www.zks.uni-leipzig.de/>)
- [8] Leipziger Interdisziplinärer Forschungskomplex zu molekularen Ursachen umwelt- und lebensstilassoziierter Erkrankungen (LIFE) (<http://www.uni-leipzig-life.de/>)
- [9] Descriptive Ontology for Linguistic and Cognitive Engineering (DOLCE) (<http://www.loa-cnr.it/DOLCE.html>)
- [10] Basic Formal Ontology (BFO) (<http://www.ifomis.org/bfo>)
- [11] Logical Observation Identifiers Names and Codes (LOINC) (<http://loinc.org/>)
- [12] International Classification of Diseases (ICD) (<http://www.who.int/classifications/icd/en/>)
- [13] Clinical Data Interchange Standards Consortium (CDISC) (<http://www.cdisc.org/standards>)
- [14] Herre, H., F. Loebe. (2005) A Meta-ontological Architecture for Foundational Ontologies. In: Meersman, R. Tari, Z. (ed.) *On the Move to Meaningful Internet Systems 2005: CoopIS, DOA, and ODBASE: Proceedings of the OTM Confederated International Conferences, CoopIS, DOA, and ODBASE 2005, Agia Napa, Cyprus, Oct 31 – Nov 4, 2005 (Part II)*. Lecture Notes in Computer Science Vol. 3761, pp. 1398-1415
- [15] Guizzardi, G., Herre, H. Wagner, G. 2002. Towards Ontological Foundations for UML Conceptual Models. In: Meersman, R. Tari, Z. (ed.) *1st International Conference on Ontologies, Databases and Applications of Semantics (ODBASE 2002)*, Irvine, California, USA. *Lectures Notes in Computer Science* Vol. 2519, pp. 1100-1117.
- [16] Mahner, M. Kary: what Exactly are genomes, Genotypes and Phenotypes? An what about Phenomes? *Journal of Theoretical Biology* (1997), 186, p. 15-21
- [17] Feiner, N., C. Sabatti: The Human Phenome Project. *Nature Genetics*, vol. 34, May 2003, p. 15-21
- [18] Burek, P. Hoehndorf, R. F. Loebe et al. A top-level ontology of functions and its application in the Open Biomedical Ontologies. *Bioinformatics*, 2006, 22 (19)
- [19] R. Hoehndorf et al. Applying the functional abnormality ontology pattern to anatomical functions (2010) *Journal of Biomedical Semantics*. 2010, 1:4
- [20] C. L Allan. (2008) Schizophrenia: From genes to phenes to diseases. *Current Psychiatry Reports*, 10 (4), 2008
- [21] Platon. *Parmenides*. Reclam 1987

Can a hole be inflamed? On the relation of morphologic abnormalities and anatomical cavities in SNOMED CT

J.Niggemann^a, H.R.Straub^b, H.Herre^c

^a CompuGroup Software GmbH, Koblenz, Germany,

^b Semfinder AG, Kreuzlingen, Switzerland,

^c OntoMed Research Group, University of Leipzig

ABSTRACT

The Definition of "Maxillary Sinusitis" in SNOMED CT does not mention the mucous membrane. It seems as if the cavern is said to be inflamed. We use this example for an in-depth discussion of the role of the maxillary sinus as an object or a location of the inflammation and the nature of the sinus as a complex object or the cavern itself. We suggest some clarifications in SNOMED CT which will make it much more usable for clinical decision support.

1 INTRODUCTION

In an advanced hospital information system, decision support functions are going to be embedded and linked to normal user actions. For example, some countries force by law or good practice regulations, that for every prescription of a drug there must be a justification. Instead of forcing physicians to enter a diagnosis code again and again whenever they make a drug order, an intelligent system can scan the patient's list of diagnoses and symptoms and compare this with the list of indications stored in the respective drug's entry in the drug database. To use a simple example, the drug database contains an entry "inflammation of mucous membrane" as indication for an anti-inflammatory drug such as acetylsalicylic acid (aspirin). If both are coded by SNOMED CT, together with the patient's disease such as "maxillary sinusitis", the promise of SNOMED CT is that using its multi-hierarchical structure, that match can be made. However, this is not the case, due not only to a local error in the respective SNOMED CT concept but due to the fact that a fundamental category is missing from SNOMED CT respective from its logical background: the concept of anatomical cavities.

2 BACKGROUND

2.1 SNOMED CT

SNOMED CT is a clinical terminology intended to support the encoding of medical data by means of a coded thesaurus of procedural and administrative terms[1]. It has been constructed to provide:

- semantic descriptors to annotate and encode clinical procedures, diagnoses, etc.
- standardized medical terms in different languages
- guidance for the construction of composed terminological expressions

SNOMED CT's backbone is given by a taxonomy of nodes, so-called SNOMED CT concepts. These are currently defined through the use of Description Logics (DL). The modeling problem we are discussing here however is independent from the format the system uses and can be discussed without further referral to DL.

2.2 Top Level Ontologies

These are designed to provide those abstract generalizations such as "material object" or "immaterial object" as mentioned above, together with the constraints such as "an immaterial object cannot have material objects as parts". Properly linking the ontology at hand to such an upper ontology can help to avoid unwanted situations and contradictions. But not only have the concepts to be properly linked, there also has to be some mechanism in place to enforce the constraints attached to the upper concepts.

3 PROBLEM STATEMENT

As stated in the introduction, we ask whether the use of SNOMED CT can support a simple function of a medical decision support system: Assuming that a medication system contains the information that acetylsalicylic acid is recommended for inflammation of mucous membranes, we want to see this rule triggered when the physician codes "maxillary sinusitis".

For this to work, SNOMED CT's definition of "maxillary sinusitis" must state that it is an inflammation of the mucous membrane in the sinus cavity. We should therefore expect a analogous definition of what we find with Rhinitis:

Fully Specified Name: Rhinitis(disorder)

ConceptId: 70076002

Associated morphology: Inflammation (morphologic abnormality)

Finding site: Structure of mucous membrane of nose (body structure)

Unfortunately, the definition we find for "maxillary sinusitis" does not follow the same lines:

Fully Specified Name: Maxillary sinusitis (disorder)

ConceptId: 70076002

Associated morphology: Inflammation (morphologic abnormality)

Finding site: Maxillary sinus structure (body structure)

The finding site specification in Rhinitis is one item that contains both what is affected (structure of mucous membrane) and where that is (nose). Note, however, that "of nose" implies that the piece of mucous membrane under consideration is stipulated to be a part of the nose. There is nothing said about a cavity. The matching algorithm here finds both "inflammation" and "mucous membrane" so it can match against the drug indication statement from the drug database.

In the case of maxillary sinusitis, the finding site is just "Maxillary sinus structure". There is nothing said about the mucous membrane in it, so the matching algorithm will fail and the decision support system will produce an erroneous alert to the effect that there is no indication for aspirin present in this patient. SNOMED CT does know the concept "Structure of mucous membrane of Maxillary sinus", and at first sight it seems enough to just change the definition of "maxillary sinusitis" so that it points to this instead of "Maxillary sinus structure". However, this would imply that the material structure "membrane" is part of the immaterial structure "maxillary sinus" which is not acceptable. Also, SNOMED CT claims to be founded on a rigid logical formalism that is meant to prevent or at least detect such errors automatically - which apparently has not happened in this case. Ontological basics such as the axiom that inflammation can not be situated in an immaterial object are entirely missing in SNOMED CT. So, a more fundamental solution to this problem is needed.

Coming back to the example "Maxillary sinus", related anatomical structures are:

- the skull
- the sinus proper (hole, which is hosted by the skull)
- the mucous membrane (which is contained by sinus but not part of the sinus)
- the mereological sum of skull plus mucous membrane
- and finally the hollow space delineated by the mucous membrane and not the skull

In SNOMED CT, the "Maxillary Sinus" has the following relations (SEP Triplets replaced by part_of relations, see below):

- is_a Nasal Sinus
- part_of Upper respiratory tract
- part_of Head

And there are the following structures which are part_of the "Maxillary Sinus":

- Anterior, Medial and Posterior Wall of maxillary antrum (each of which is part_of "cranial bone" and is_a "Structure of paranasal sinus wall")
- "Septum of maxillary sinus" and "Lamina propria of maxillary sinus" (which are not part_of anything else)
- "Maxillary sinus mucous gland" and "Mucous membrane of maxillary sinus"

This illustrates that, *in SNOMED CT "Maxillary Sinus" seems to be understood as the mereological sum of the sinus proper, all it contains (the glands and the mucous membrane), together with the parts of the skull that delineate it.*

On the other hand, there is the concept "Body cavity". This has subconcepts "oral cavity", "abdominal cavity" etc, but not one of the nasal sinuses. "Body cavity" itself is a primitive concept. Especially, its logical definition does not allude to the distinction of material versus immaterial objects.

4 PREREQUISITES: SOLVE THE SEP TRIPLETS AND INTRODUCE A PART-OF RELATION

SNOMED CT currently suffers from a variety of problems, to which several remedies have been proposed (one of many publications on this topic is [2]). The part-of relation is of primary interest for the topic at hand.

Currently, part-of relations are represented using the is-a subsumption and a construction called "SEP triplets" (Structure-Entire-Part, e.g. "Arm Structure", "Entire Arm" and "Arm Part"). Meanwhile the development of DL has advanced to a point where part_of can be properly handled, making the SEP construction unnecessary. We therefore postulate as a prerequisite:

The SEP triplet problem shall be solved like proposed in [3], abolishing the SEP triplets and only relying on properly defined part_of relations. (1)

In conjunction with this, we also postulate that

the relation "finding site" should be abandoned, and the relations "has_object" and "has_location" introduced instead. (2)

Suntisrivaraporn et al [3] make a similar suggestion but propose to introduce "has direct location" instead of "has_object".

In this paper, we will present examples from SNOMED CT in a way they would look like after this change, using `part_of` instead of the SEP construction. In examples from current SNOMED CT, we will continue to use "finding site", whereas in our solution proposition we will use "has_object" and "has_location" as applicable.

5 RESULTS

5.1 Foundations from an Upper Ontology

In the introduction we have mentioned that such basic concepts as "material object" and the associated constraints should be maintained in an upper ontology, and that SNOMED CT concepts should be linked back to such an ontology. We have chosen GFO [4] as the ontology of choice for these purposes. This offers precise definitions of spaces, boundaries and holes, which we will cover in the next chapter.

Boundaries and Spaces. Holes are defined by their boundaries, so we need to look into the nature of those first.

In contradistinction to an arbitrary material boundary - which can be a fiat entity - a natural boundary demarcates the material object from its environment by distinguishing properties. A tangential part (w.r.t. the natural boundary) of the object must have a property which is different from a property associated to the environment at this place [5; 6]. For the relation of objects and spaces this means:

Every material object occupies a space region and if an object Ob occupies the space region SR then the material boundaries of Ob occupy the corresponding space-boundaries of SR . (3)

A Theory of Holes. The ontology of holes - as used in this paper - is based on the notion of natural boundary as outlined above. A hole pertains to the form of the natural boundary of a material object. A hole of the material object Ob is a non-empty connected spatial span of some connected part of its natural boundary. We assume axiomatically that a hole of a material object and the object's occupied space do not overlap. Basic holes include depressions, hollows, tunnels, and cavities of a material object. From these basic assumptions the following conditions can be established:

1. Holes have a host, they are dependent entities
2. The host is not part of the hole
3. Holes can contain a material structure
4. The contained is not part of the hole, but the host of the hole can (but doesn't need to) have the contained as part. (4)

5.2 Desiderata for SNOMED CT

Based on the above foundations, we can now list the conditions that SNOMED CT must fulfill in order to correctly handle references to pathological processes which are located in anatomical cavities.

SNOMED CT should basically distinguish (5)

- *body part, which is a material object and can be a role filler for "has_object"*
- *body region, which is a space region and can be a role filler for "has_location"*
- *body cavity, which is a hole and can also be a role filler for "has_location"*

SNOMED CT's logic should govern the relations between body parts and body regions according to an ontological theory of objects, boundaries and spaces (6)

With respect to cavities,

SNOMED CT's logic should govern the relations between body parts and body cavities. (7)

The mereological sum of a hole, its walls and its contents, such as "maxillary sinus" is apparently presently understood in SNOMED CT, can only be a location, never an object. It should however be avoided altogether. (8)

The definition of maxillary sinusitis would then look like this:

Fully Specified Name: Maxillary sinusitis (disorder)

Associated morphology: Inflammatory alteration (morphologic alteration)

has_object: mucous membrane (body part)

has_location: Maxillary sinus (body cavity) (9)

6 DISCUSSION

6.1 Formal Definitions

The above listed stipulations as well as the proposed definitions of sinusitis and its related concepts can be formalized in different ways. For biomedical ontologists, some dialect of Description Logics (DL) is the informally accepted standard to do this. It is however controversial among the authors of this contribution, whether SNOMED CT should itself be regarded an ontology and whether DL definitions are really beneficial for its use.

We hold that DL definitions of the above are possible in at least two of the proposed formats, namely OWL-DL [7] and EL+ [3].

6.2 Linking to Top-Level Ontologies

We chose GFO instead of BFO [8], because it is more explicit in its handling of spaces and boundaries. BFO essentially adopts the theories expounded in [9] and [10].

The natural boundaries (of a material object) in GFO correspond to the bona-fide boundaries of BFO. In GFO natural boundaries of different material objects may touch, whereas in BFO, according to [9], this is excluded. This is a weak point in the BFO-approach, as criticized already in [11], and more detailed in [6]. In GFO holes are connected space entities (topoids) which are spanned by natural boundaries of objects. This idea is not present in BFO. GFO solves several problems, for example the problem, that a hole moves together with its host, simply, because the hole is connected to a natural boundary which moves obviously together with the object. Furthermore, material boundaries of a material object are different from the spatial boundaries of the space region occupied by the object. Material boundaries never coincide with pure space boundaries, instead, they occupy them. In BFO there is a mixing of these entities and relations whose usage may easily lead to inconsistencies. More information about this point is presented in [6].

6.3 A System Perspective

From a clinician's perspective it may be hard to understand at first sight why the mucous membrane of the stomach is part of the stomach, the mucous membrane in the maxillary sinus however is not part of the sinus. It may be easier to understand the cavity, its contents, and its wall together as a system – that works for both stomach and maxillary sinus. This would support SNOMED CT's current understanding of "maxillary sinus" as cited in chapter "Background: SNOMED CT". Linking the inflammation to this entire system however is still wrong. To further pursue this perspective, a theory of such systems would be needed. This must be a concept-oriented theory, because these systems under consideration are primarily concepts in the minds of clinicians.

7 CONCLUSION

Starting from a simple example, we have shown that quite a number of improvements must be made to SNOMED CT in order to make it usable for decision support tasks. To those requirements already known, we have added some that especially deal with the handling of anatomical cavities and pathological alterations located therein. These changes can be applied consistently if they are based on a suitable upper ontology.

ACKNOWLEDGEMENTS

The results presented in this contribution were obtained through the authors' regular work on ontologies, medical coding systems and their application. Thanks are due to the companies CompuGroup and Semfinder who sponsored the extra time needed to prepare this publication.

REFERENCES

- [1] Spackman K. SNOMED Clinical Terms Fundamentals. presentation, available from URL:http://www.ihtsdo.org/fileadmin/user_upload/Docs_01/SNOMED_Clinical_Terms_Fundamentals.pdf. 2009 Jul.
- [2] Schulz S, Suntisrivaraporn B and Baader F. SNOMED CT's problem list: ontologists' and logicians' therapy suggestions.. *Stud Health Technol Inform* 2007; 129(Pt 1): 802-806.
- [3] Suntisrivaraporn B, Baader F, Schulz S and Spackman. K. Replacing SEP-Triplets in SNOMED CT using Tractable Description Logic Operators. In: Hunter J, Bellazzi R and Abu-Hanna A (editors). *11th Conference on Artificial Intelligence in Medicine, AIME 2007, Amsterdam, The Netherlands, July 7-11, 2007. Proceedings*. Berlin: Springer; 2007. p. 287-291. (Lecture Notes in Computer Science; vol. 4594).
- [4] Herre H, Heller B, Burek P, Hoehndorf R, Loebe F and Michalek H. General Formal Ontology (GFO): A Foundational Ontology Integrating Objects and Processes. Part I: Basic Principles. *Onto-Med Report*. Research Group Ontologies in Medicine (Onto-Med), University of Leipzig, 2007.
- [5] Herre H. General Formal Ontology (GFO): A Foundational Ontology for Conceptual Modelling. In: Poli R and Obrst L (editors). *Theory and Applications of Ontology*. Berlin: Springer; 2009.
- [6] Baumann R. The Mereotopological Structure of the Brentanoraum B3 and Material Entities. Masters Thesis at Universität Leipzig, 2009.
- [7] Rector AL and Brandt S. Why do it the hard way? The case for an expressive description logic for SNOMED. *J Am Med Inform Assoc* 2008 Nov/Dec: 15(6): 744-751.
- [8] Grenon P, Smith B and Goldberg L. Biodynamic ontology: applying BFO in the biomedical domain. *Stud Health Technol Inform* 2004; 102: 20-38.
- [9] Smith B and Varzi AC. Fiat and Bona Fide Boundaries. *Philosophy and Phenomenological Research* 2000; 60(2): 401-420.
- [10] Casati R and Varzi AC. *Holes and other superficialities*. Cambridge (MS): The MIT Press; 1994.
- [11] Ridder L. *Mereologie*. Frankfurt a. Main: Vittorio Kolstermann; 2002.

Representing Dispositions

Johannes Röhl, Ludger Jansen*

Institute of Philosophy, University of Rostock, Rostock, Germany

ABSTRACT

Motivation: Dispositions and tendencies feature significantly in bio-medical knowledge. They are not only important for specific applications like an infectious disease ontology, but also in a general strategy for modeling knowledge about molecular interactions. But the task of representing dispositions in some formal ontological systems is fraught with several problems, which are partly due to the fact that Description Logics can only deal well with binary relations. The paper will discuss some of the results of the philosophical debate about dispositions, in order to see whether the formal relations needed to represent dispositions can be broken down to binary relations. Finally, we will discuss problems arising from the possibility of the absence of realizations, of multi-track or multi-trigger dispositions and offer suggestions how to deal with them.

1 INTRODUCTION

Terms for dispositions and their cognates like tendencies or propensities are important in biomedical data structures (Jansen 2007). SNOMED CT, for example, contains *Determination_of_disposition_of_blood_product (procedure)* as a procedure, *Character_trait_finding_of_sadistic_tendency (finding)* as a clinical finding, *Propensity_to_adverse_reaction_to_drug* as a disorder (January 2010 Release). Dispositions relevant for the biomedical domain comprise also the tendency of a patient for nausea, the capacity of drugs like aspirin to relieve pain, and the dispositions of molecules to undergo certain chemical reactions under certain circumstances etc. Generally, in the medical domain we are often not only interested in what actually happens, but also in what could happen, e.g. to prevent something from happening by taking precautions, and disposition ascriptions are a way to express which occurrences could take place if certain conditions are met.

2 DISPOSITIONS: WHAT AND WHY

A disposition is a property that is linked to a realization, i.e. to a behavior the individual that bears the disposition will show under certain circumstances or as response to a certain stimulus (trigger). Common examples are fragility (disposition to break when dropped) or solubility (to dissolve when put in water). Many philosophers have thought that disposi-

tion ascriptions have a dual nature (Mumford 1998, Molnar 2003, Jansen 2007b etc.). On the one hand, they *entail* counterfactual conditionals: *x has disposition D* implies: *If x was under certain Circumstances C, x would exhibit behavior R (the realization of the disposition)*. On the other hand, and primarily, they are ascriptions of such properties like fragility or solubility to their bearers – properties that are causally relevant for processes that involve the disposition’s bearer as a participant. We will briefly discuss two applications of dispositions in the ontology of biology and medicine that show why these features make dispositions crucial in many respects.

2.1 Dispositions and Interactions

A general feature of dispositions is that they can provide a connection between continuant entities and processes. To avoid having to deal with interactions as direct relations between the interaction partners of higher arity Schulz and Jansen (2009) suggest the following reification strategy: As a first step interactions are taken as occurrents, entities in their own right, to reduce the arity. The interaction process then stands in the binary relation, *has_participant*, to two or more continuants. A further suggestion is then to take interaction processes as realizations of the dispositions of the participants in interactions (in line with Ellis/Lierse 1994).

2.2 Diseases as Dispositions

The central role of dispositions is shown by the fact that recent biomedical ontologies define diseases as dispositions towards pathological processes (Scheuermann/Ceusters/Smith 2009, Goldfain/Smith/Cowell 2010, 402, following OGMS, the Ontology for General Medical Science). A disease is characterized by the three categorically distinct aspects, *disorder*, *disposition*, and *course of disease*. The corresponding *disorder* is the basis of the disposition (i.e. the structure of properties that bring with it the disposition in question) as part of an organism. The *course of disease* is the realization of the disease (as disposition), an occurrent or process. Two things should be noted: The disorder is specific for the disease. Not any presence of infectious agents in some organ means that the patient has a particular disease, but it must be the proper type of agents for the disease in question. And according to this model “disease” without further qualification is an ambiguous term: The disorder and associated dispositions may in some cases be “dormant” until it is realized in the disease course (and it may even

* To whom correspondence should be addressed.

never be realized in some patients). This distinction is important, because it helps to describe the case of patients having the disease qua disposition but no symptoms, because the realization is blocked by an additional factor (e.g., medication that counteracts the disease). A patient may carry the infectious agents and even spread them without exhibiting typical symptoms, cf. infectious mononucleosis (Pfeiffer's disease) and similar cases.

In this context it is also important that the relation between a disposition and its bearer is not confounded with the relation between a disposition and its basis. The bearer is an independent continuant (the diseased organism or part thereof), the disorder is a property (or set of properties) of the bearer. To make the dependencies clearer, the OGMS-“triangle” should be extended to a “square” including the substantive bearer, the base, the disposition and the realization. We now have to further inquire into the edges of this “square” (cf. Figure 1).

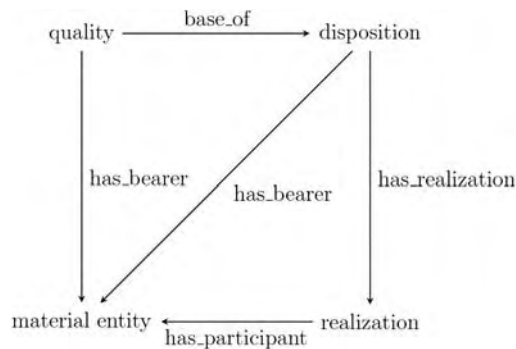


Fig. 1. Formal relations for the ontology of dispositions.

3 DISPOSITIONS AND THEIR RELATIONS

In the top-level ontology BFO (“Basic Formal Ontology”) dispositions are taken as a type of “realizable entity” (with two siblings, “function” and “role”) which in turn is a subclass of “dependent continuant” (with sibling “quality”; cf. Arp/Smith 2008). Thus there are two relations that are in any case necessary for a disposition: As a dependent continuant it needs an independent continuant as bearer. And as a realizable it needs a realization, a process entity that will be realized, given the appropriate conditions. But the most general formulation admits of more relata: A disposition ascription according to the schema “x has disposition D for realization R with trigger T under circumstances C with probability P” (Schulz/Jansen 2009) has no less than six relata: the independent continuant x that has the disposition, the disposition D, the realization R (process), the trigger T and the circumstances C, and a probability P of the realization. The Description Logics that are used by ontology languages like OWL, however, can deal well with binary rela-

tions only (cf. Baader et al. 2003: 46, 204 f.). In the following, we will try to reduce this complexity in a systematic fashion, using binary relations only. However, considerable difficulties will remain.

3.1 Starting with the simple

In order to get started, we will first consider a simplified type of disposition ascription. We will (1) concentrate on “sure fire dispositions”, i.e. such dispositions that will necessarily be realized given the respective realization conditions. Hence we will abstract from all those dispositions that even in “ideal conditions” will be realized only with a certain probability, and which are often called “propensities” (Popper 1959) or “tendencies” (Jansen 2007). This seems to be ontologically sound, as the ascription of a probability does not establish any ontological addition. For the purpose at hand, a probability is a real number from the interval $]0,1[$, and the ascription of the probability is a function of the tuple \langle disposition, realization \rangle into that interval. It seems plausible that such a function can be attached, after we have disentangled the main complex. We will also (2) ignore the distinction between “trigger” and “background conditions” and focus on triggering processes only. The ontological basis for this distinction is that triggers come along with changes (the striking of the match), while background conditions may remain constant (presence of oxygen). But to a large degree, this distinction is a pragmatic one. Usually the background conditions are taken for granted, while the trigger is thought of as the causal factor that (a) makes a difference and (b) can be or actually is influenced by human action. But both the trigger and the background conditions are necessary conditions for the realization. In this paper, we focus on the trigger, because background conditions can comprise categorially diverse kinds of entities (location, presence of oxygen, milieu of certain pH value, concentration) and it is currently not at all clear how many of these conditions are to be analyzed ontologically. It is very much desirable to integrate such conditions, once their ontological status becomes clear, because otherwise important cases are ruled out, where, e.g., the realization of a disease may be influenced by such an additional condition, either a certain condition of a patient or a drug treatment. Moreover, we will start with considering (3) cases where one type of disposition is correlated with exactly one type of realizations, i.e. we exclude dispositions that allow for several types of realizations (so-called “multi-track dispositions”). Finally, we will consider only (4) cases which involve one type of trigger only, i.e. we exclude all dispositions that can be triggered by processes of more than one type.

3.2 Binary formal relations

We now try to spell out and characterize the basic binary relations dispositions stand in. We start with the level of

particulars. We follow the OBO Relation Ontology (Smith et al. 2005) and regard as primitive the particular level relations **instance_of** (which holds between a particular and its classes, both for processes and continuants) and **has_participant** (which holds between particular processes and particular continuants). We also adopt the convention of using boldface for relations between particulars and italics for the ones between types, lower case letters for variables for instances and upper case for classes. In addition, we use the particular level relation **inheres_in**, which holds between individual instances of dependent and independent continuants (Schulz/Jansen 2009, 23). It expresses a kind of one-sided ontological dependence, i.e. if p **inheres_in** x , then it is possible that x exists without p , but not that p exists without x .

3.2.1 Has_disposition Given these primitive relations, we can define a relation **has_disposition** along the following line:

$$x \text{ has_disposition } d := \exists D d \text{ instance_of } D \wedge \\ D \text{ is_a } \textit{Disposition} \wedge d \text{ inheres_in } x$$

With the help of these particular level relations we can now start to define the wanted universal level relation *has_disposition* (which is a subrelation of *has_property* as defined in Schulz/Jansen 2009, 23). This relation will be needed to express that certain types of things have some dispositions essentially. In these cases, *all* instances will have instances of the disposition type inhering in them, e.g. all aspirin pills have the disposition to relieve pain, and all sugar has the disposition to dissolve in water:

$$A \text{ has_disposition } D := \forall x x \text{ instance_of } A \rightarrow \\ \exists y y \text{ instance_of } D \wedge y \text{ inheres_in } x$$

We have, however, to be careful when applying this relation, for in many cases only some instances of a given type will have a certain disposition. E.g., only some patients have the disposition to show an allergic reaction to penicillin, and only some mosquitoes have the disposition to transmit malaria.

3.2.2 Has_bearer The inverse relation of **has_disposition** is the relation **has_bearer**. Like all property instances, disposition instances need a bearer in order to exist, as dispositions are dependent continuants. This relation is, thus, a subrelation of the general **inheres_in** relation. But even if particular dispositions are of the same type, their bearers do not have to be of the same type. Different sorts of drugs are bearers of the disposition to relieve pain e.g. aspirin as well as paracetamol. If, however, all instances of a disposition have bearers of the same type, we can represent this by means of the *has_bearer* relation, which we can define as follows:

$$X \text{ has_bearer } Y := \forall x: x \text{ instance_of } X \rightarrow \\ \exists y y \text{ instance_of } Y \wedge y \text{ inheres_in } x$$

3.2.3 Has_realization In addition, it will be useful to have relations linking a disposition to its realization and its trigger. It is, however, not possible to define such relations according to the usual all-some semantics, because the whole point of disposition ascriptions is that not every instance of a type of disposition needs to be realized: “For all instances x of D there exists a process instance r of R that is the realization of x ” is clearly wrong! Otherwise, all dispositions would always be realized. But a particular disposition of a particular material entity may never be realized and there are even cases when *no* realization token of a disposition type is realized, because the triggering circumstances are never met. This is a feature, not a bug, as we want dispositions to be real without actual realizations. It might well be that there is a type of disposition of which no instance will ever be realized, like, e.g., the disposition of a nuclear power plant to explode, the realization of which is prevented by highly sophisticated technical machinery. (In most biomedical cases, however, usually some tokens of the disposition will be realized or will have been realized before.) We use **has_realization** as the primitive relation that connects a disposition instance with any process instance which is its realization. This is both an *ontological* and a *causal* connection, and we can roughly characterize it by saying that the realization has been brought about by the disposition. Instead of using the all-some apparatus, we introduce *has_realization* by setting up a set of value restrictions. A type of process R is the realization type of a disposition type D if and only if any instance of D is realized, then the realization is of type R :

$$D \text{ has_realization } R := \forall x (x \text{ instance_of } D \rightarrow \\ (\forall y: x \text{ has_realization } y \rightarrow y \text{ instance_of } R))$$

3.2.4 Has_trigger We will introduce two distinct relations to describe the trigger of a disposition. The trigger relation can have these two variants as we can split the fundamental nexus <disposition, trigger, realization> in two ways. When focusing on the disposition we consider the relation between the triggering process and the disposition that is realized because of the trigger, e.g. between the fragility of a glass and its dropping. On the other hand it comes natural to think of a (causal) connection holding between the triggering process and the realization of the disposition (the dropping of the glass triggers its breaking). We thus introduce a relation **has_trigger_D** that holds between the disposition and the triggering process and a relation **has_trigger_R** that holds between the realization process and the triggering process. Given our simplifying assumption that there is exactly one type of realizations and exactly one type of trigger for each disposition, we need to take only one of them as primitive;

the other one can be than expressed through a concatenation with **has_realization** as follows:

$$d \text{ has_trigger}_D t \Leftrightarrow \exists r (d \text{ has_realization}_R r \wedge r \text{ has_trigger}_R t)$$

$$r \text{ has_trigger}_R t \Leftrightarrow \exists d (d \text{ has_realization}_R r \wedge d \text{ has_trigger}_D t)$$

For the corresponding relation on the level of universals the same cautionary remarks as in the case of *has_realization* apply: Not for every particular disposition there is a particular trigger process, but only for those particular dispositions that become realized. Thus, again, we cannot use the all-some apparatus but must express our knowledge about types of trigger processes using value restrictions along the following lines:

$$D \text{ has_trigger}_D T := \forall x (x \text{ instance_of } D \rightarrow \forall y (x \text{ has_trigger}_D y \rightarrow y \text{ instance_of } T))$$

If there is only one type of possible triggers for a certain disposition, every realization must have been triggered by one of that class. For all r **instance_of** R there is some t **instance_of** T and $R \text{ has_trigger}_R T$.

3.2.5 Base_of Important discoveries about dispositions regard the question how dispositions are brought about through the interplay of micro-level structures and their properties. Such a microstructure is normally labeled as the “base” of a disposition. For example, the solubility of salt is based on the molecular structure of NaCl that allows the polarized water-molecules to break the ion bonds. While there are good reasons to assume that there are dispositions on the quantum physical level that do not have a non-dispositional “grounding”, dispositions in the biomedical domain can probably always be connected to a set of properties that account for their existence. We do not introduce a further primitive relation on the particular level, but characterize the universal level relation through the need of a disposition instance inhering in the bearer of the base quality:

$$(Q \text{ is_a } Quality \wedge D \text{ is_a } Disposition \wedge Q \text{ base_of } D) \rightarrow \forall q (q \text{ instance_of } Q \rightarrow \exists b (b \text{ bearer_of } q \wedge \exists d (d \text{ instance_of } D \wedge b \text{ bearer_of } D)))$$

While equal bases bring with them instances of the same type of dispositions, the inverse does not hold. As a rule, dispositions in the biomedical domain can be constituted by different types of base qualities. This general formulation leaves leeway for several candidates for bases. If a patient lacks a certain enzyme one could either take that absence as the basis of the disposition for the resulting pathophysiological processes. Or, to avoid “negative entities” like absences, one could say that the whole pathological (micro-)structure of an organ is the base of the respective disposition for a pathological process. Such a pathological structure can be further analysed into its constituents and their respective dispositions. Thus we get something like “a cascade” of

dispositions: The disposition to produce insufficient amounts of the enzyme leads to the bodily state characterized by lack of the enzyme and that state is in turn the base for the disposition for pathological processes resulting from that state. An advantage of disposition talk is, however, that we do not have to name or even to know the base. E.g., we often know the dispositions of a drug from clinical studies without knowing anything about the molecular mechanism that is the base of these dispositions.

3.3 Principles for Surefire Dispositions

To connect these relations among each other we now suggest a number of postulates that are intuitively plausible for surefire dispositions. The first one is the *realization principle*, i.e. the principle that if a disposition and its trigger are given, the realization will happen:

$$(\exists D D \text{ is_a } Disposition \wedge \exists T T \text{ is_a } Process \wedge \exists R R \text{ is_a } Process \wedge D \text{ has_trigger}_D T \wedge D \text{ has_realization}_R \wedge \exists d d \text{ instance_of } D \wedge \exists t t \text{ instance_of } T) \rightarrow \exists r r \text{ instance_of } R$$

The second principle is the *bearer principle* which is valid for all kinds of dispositions, i.e. the principle that the bearer of the disposition is a participant of the respective realization process. The following formulation of this principle, however, relies on the realization principle, and is thus only valid for surefire dispositions:

$$(\exists D D \text{ is_a } Disposition \wedge \exists T T \text{ is_a } Process \wedge \exists R R \text{ is_a } Process \wedge D \text{ has_trigger}_D T \wedge D \text{ has_realization}_R \wedge \exists d d \text{ instance_of } D \wedge \exists t t \text{ instance_of } T \wedge \exists b b \text{ bearer_of } d) \rightarrow \exists r (r \text{ instance_of } R \wedge r \text{ has_participant } b)$$

The converse will not hold, because a process normally is the realization of several different dispositions. A movement is the realization of both an active disposition to cause motion and a passive disposition to undergo motion. But any participant participates in the respective process in virtue of one of its passive or active dispositions.

4 OUTLOOK

When we try to extend this analysis to include more realistic cases, we face problems when it comes to multi-track and multi-trigger dispositions. A fragile thing can break into pieces, but it can also crack or splinter – and all of those would count as realizations of fragility. We can also trigger the fragility disposition in different ways, like striking, throwing onto a hard surface etc. Similar things appear in medical cases where the realization of a disease, the course of the disease and the symptoms, may cover a wide spectrum of behavior.

There are several options here. One would be to get rid of such dispositions by “fine-graining”, i.e. “splitting up” a multi-track or multi-trigger disposition into several more precisely specified ones, each with its specific triggering conditions and realization. One problem here is that this

strategy does not seem appropriate to some central applications in the biomedical domain. A disease may have different disease courses as realization, but still be the same disease, because the underlying disorder is the same. In this case, the identity of the disposition could be maintained with recourse to the underlying disorder. But such an approach cannot be extended to all dispositions, because in general one disposition can be “multiply realized”, i.e. it can be grounded in different bases: Two different pain-relieving drugs both have the disposition to lessen pain, but the chemical base for that ability may be rather different in each of these substances.

Another option would then be to give a disjunctive list of the possible realizations of multi-track dispositions. In many cases, these realizations will exclude each other, but in others they will not be disjoint: The realization of a particular disease may involve fever *or* nausea *or* both of these. For the medical domain this rather inelegant approach seems to be the most appropriate. Finally, if we take a comparably rough description we may cover several cases with one generic type of realization: Different processes are classified as disease courses of a particular disease without taking into account the differences.

ACKNOWLEDGEMENTS

Many thanks to Niels Grewe for stimulating discussions and to the referees for helpful comments. Research for this paper was funded by the German Research Foundation (DFG) within the research project “Good Ontology Design” (GoodOD).

REFERENCES

- Arp., R. and Smith, B. (2008) Function, Role, and Disposition in Basic Formal Ontology, *Proceedings of Bio-Ontologies Workshop (ISMB2008)*, 45–48.
- Bird, A. (2007) *Nature's Metaphysics: Laws and Properties*, Clarendon Press, Oxford.
- Ellis, B. and Lierse, C. (1994) Dispositional Essentialism. *Australasian Journal of Philosophy* **72**, 27-45.
- Ellis, B. (2001) *Scientific Essentialism*. Cambridge UP, Cambridge.
- Goldfain, A.; Smith, B. and Cowell, L.G. (2010) Dispositions and the infectious disease ontology, A. Galton and R. Mizoguchi (Eds.): *Formal Ontology in Information Systems*, IOS Press, 400-413.
- Jansen, L. (2007a) Tendencies and other Realizables in Medical Information Sciences, *The Monist* **90/4**, 534-555.
- Jansen, L. (2007b) On ascribing dispositions, Kistler, M. and Gnassounou, B. (Eds.): *Dispositions and Causal Powers*, Ashgate: Aldershot, 161-177.
- Mumford, S- (1998) *Dispositions*. Oxford UP, Oxford.
- Molnar, G. (2003) *Powers. A study in metaphysics*. Oxford UP, Oxford.
- Popper, K.R. (1959) The Propensity interpretation of probability, *British Journal for the Philosophy of Science* **10**, 25-42.
- Scheuermann, R, Ceusters, W. and Smith, B. (2009) Toward an Ontological Treatment of Disease and Diagnosis, *Proceedings of the 2009 AMIA Summit on Translational Bioinformatics*, 116–120.
- Schulz, S. and Jansen, L. (2009) Molecular Interactions: On the ambiguity of ordinary statements in biomedical literature, *Applied Ontology* **4**, 21-34.

Grains, Components and Mixtures in Biomedical Ontologies

Ludger Jansen^{1*} and Stefan Schulz²

¹ Institute of Philosophy, University of Rostock, Germany

² Institute of Medical Biometry and Medical Informatics, Freiburg University Hospital, Germany

ABSTRACT

In biomedical ontologies, mereological relations have always been subject to special interest due to their high relevance in the description of anatomical entities, cells, and biomolecules. This paper investigates two important subrelations of **has_proper_part**, viz. the relation **has_grain**, which relates a collective entity to its multiply occurring uniform parts (e.g. water molecules in a portion of water), and the relation **has_component**, which relates a compound to its constituents.

At first sight, collectives and compounds seem to be disjoint categories. Their disjointness, however, relies on agreement about what are uniform entities, and thus on the granularity of description. For instance, the distinction between isomeric subtypes of a molecule can be important in one use case but might be neglected in another one. We demonstrate that, as implemented in the BioTop domain upper level ontology, equivalence or subsumption between different descriptions of same or similar entities cannot be achieved. We propose a new design pattern that avoids primitive subrelations at the expense of more complex descriptions and thus supports the needed inferences.

1 INTRODUCTION

In biomedical ontologies, mereological relations between parts and wholes have always been conferred a special importance due to their relevance for describing material entities such as body parts, cells, cell components, and biomolecules [5,7,8]. Numerous subrelations of **has-part** relevant for the biomedical domain have been discussed. In BioTop, a top-domain ontology for the biomedical domain¹, the number of relations has been restricted to the minimum, mostly following the precepts of the OBO relation ontology [9]. However, the need for two distinct mereological relations, **has_grain** and **has_component**, both subrelations of **has_proper_part**, has been advocated. BioTop builds upon the formal design principles propagated by the OBO Foundry initiative² and is implemented in OWL-DL,³ the standard ontology language of the Semantic Web. We will first

present the approach of BioTop (§ 2.1), discuss it critically (§ 2.2), and then present a new suggestion (§ 3) and discuss whether it can and should be applied to mixtures, too (§ 4).

2 GRAINS AND COMPONENTS IN BIOTOP

2.1 The approach of BioTop

Schulz et al. adduce various criteria to distinguish between grains and components [8]:

- Grains are typically the constituting elements of homogeneous collections, such as the sheep in a flock or the H₂O molecules in a drop of water.
- Components are the constituting elements of a compound constituted by well identified parts, such as a bicycle being composed by frame, wheels, saddle, front set etc, or a skull composed by neatly distinct bones.

In BioTop's account of compounds and components a necessary criterion is that the compound's sortal identity depends on the exact sum of its components. In contrast, the sortal identity of a collective does not. The example given is a blood sample, from which a blood cell is removed: What is left behind, is still a blood sample, i.e. an entity of the same type. Removing a nucleotide from a gene sequence, on the other hand, brings into existence an instance of a different type. Secondly BioTop claims that grains unlike components are not spatially connected – a criterion which is not further expanded on for lack of an uncontroversial model of connection.

For their formal characterization, however, they use still other properties of the relations: In the relation **has_component**, no two components of a compound are overlapping, and the components exhaust the whole compound. In the relation **has_grain** all grains must be of the same type and any two grains are spatially disconnected, while all grains together exhaust the whole.

2.2 Problems

Schulz et al. present their formal characterization as giving both necessary and sufficient conditions for the respective relations, i.e. as definitions. However, their formal descriptions are not unproblematic. It is lost in the formalization that any complex thing can be partitioned in various ways, and any component is a component only with respect to a certain partition of the compound.

* To whom correspondence should be addressed.

¹ Cf. [2] and <http://www.imbi.uni-freiburg.de/ontology/biotop/>.

² Cf. [10] and <http://www.obofoundry.org/>.

³ Cf. [11] and <http://www.w3.org/TR/owl-guide/>.

Many of these formal shortcomings can easily be repaired. Not as easily resolved, however, is that these formal characterizations do not at all capture the criteria to draw a clear distinction between grains and compounds. Rather, as the definitions stand, all grains are also components. For having no spatial overlap is a necessary condition for being spatially disconnected, and thus everything that is spatially disconnected has no spatial overlap. As can be seen from the criteria discussed in § 2.1, this result is not at all intended.

3 GRAINS AND COMPONENTS

3.1 Defining Grains

Grains are constituent parts of pluralities: Herds are pluralities of cows, bacteria colonies are pluralities of bacteria, and water samples are pluralities of water molecules. Such pluralities of grains are often called collectives. Grains and collectives are closely related: Grains are grains of collectives and collectives are collectives of grains. To be a collective and to have grains are thus two sides of the same coin.

In order to capture this idea we use a COLL-index as an operator that takes universals and yields a universal of collectives of instances of the original universal [6]. To make things easier (and consistent with earlier work), we allow for collectives with one grain only. We can thus postulate:

$$x \text{ instance_of } X_{\text{COLL}} \Leftrightarrow \forall g (g \text{ grain_of } x \supset g \text{ instance_of } X)$$

With the help of the COLL-index and the **instance_of** relation we can now state what it is to be a grain:

$$x \text{ has_grain } y \Leftrightarrow_{\text{def}} (x \text{ has_proper_part } y) \wedge \exists X \exists Y (x \text{ instance_of } X \wedge y \text{ instance_of } Y \wedge X \text{ is_a } Y_{\text{COLL}})$$

The first conjunct guarantees that **has_grain** is a sub-relation of **has_proper_part** (and thus also a sub-relation of **has_part**). The last conjunct guarantees both that all grains in question are grains of the same type (i.e. of type Y) and that there are no grains of x that are not Y s (i.e. that the Y s exhaust the whole x).

It should be noted that we do not add any further requirement concerning the disconnectedness of grains. Hence it does not matter whether the grains of a collective happen to be connected or disconnected. A collection of water molecules may be disconnected when existing in gas state, while having connections in various degrees when in liquid or solid state. In which state whatsoever it exists, whether the grains are connected or not, it is still a collection of water molecules.

3.2 Defining Components

A component is always a component of a compound with respect to a certain partition of this compound. A protein chain, for example, can be partitioned into its constituent

amino acid monomers, but it can also be portioned into the atoms it consists of. These cases differ with respect to their level of granularity, but a partition may also arbitrarily crisscross granularity levels, e.g. by portioning half of a protein chain into monomers and the other half into atoms. Like many terms ending with “-ion”, “partition” features a product-process-ambiguity. “Partition” can denote the (mechanical or cognitive) act or process of dividing something into parts. The term can also denote the product of such a process, i.e. a collection of parts that make up the whole. These parts can be called the “segments” of a partition. The segments of a partition are pairwise disjoint and mutually exhaust the whole compound. That is:

- If p_1 and p_2 are two distinct segments of the same partition P , p_1 and p_2 have no spatial overlap.
- The mereological sum of all the segments of a partition overlaps completely the compound, and *vice versa*.

The segments of a partition form a collective entity that is an independent continuant. We can name such a partition by enumerating its segments.

Not all segments are components, because not all partitions are partitions of a compound into components. Let p be a partition consisting of the segments p_1, p_2, \dots, p_n . If all segments are instances of the same type of independent continuants, the partition can be regarded as establishing a collective consisting out of n grains. For a compound, however, it is not necessary that all components are of the same type, but their number is essential. We can thus define a compound and the relation **has_component** as follows:

- Let p be a partition consisting of the segments p_1, p_2, \dots, p_n . Then p is a compound of type X if and only if it is not possible to add further segments p_{n+1}, p_{n+2}, \dots or to subtract any of the segments in such a way that the resulting sum is still an instance of X .
- $x \text{ has_component } y \text{ with_respect_to_a_partition } p$ if and only if the partition p of x is a compound consisting of p_1, p_2, \dots, p_n and $y = p_i$ for some $1 \leq i \leq n$.
- $x \text{ has_component } y$ if and only if there is some segment p_i of some partition p of x , such that $y = p_i$.

As the relation **has_component** makes no reference to any specific partition, this relation might in many cases be too weak and thus to uninformative. If we require that the partition in question is on a certain level of granularity G (with G indicating the partition level by being a placeholder for, e.g. *Organ* or *Molecule* or *Atom*), we can restrict the granularity of the components to the required level.

Again we do not have to add any further requirement to ensure that components have no proper overlap, because this is entailed already by the logical properties of partitions. The components of a compound may or may not be connected to each other. The components of the skull are connected to each other, while the components of a chamber

music quartet are normally disconnected. Nevertheless, the chamber music quartet is a full blown compound: Under the canonical partition, it has four components, the subtraction of one of which would put an end to the quartet (for then it would be a trio). The existence of skulls or molecules, however, requires that the components are in fact connected to each other. This, however, has to be ensured with additional mereotopological means (cf. e.g. [3]).

4 COPING WITH MIXTURES

We will now apply the notions of grains and components to the area of biomedicine and biochemistry. In biological systems, we practically never encounter pure substances such as 100% alcohol or pure oxygen. The normal case, e.g. found in body substances, tissues, or in the cell protoplasm is a mixture of grains of different kinds and sizes. A complete enumeration of all those different kinds is often neither possible nor desirable. There are different ways to represent this situation using grains and components. We will describe these ways using the DL-standard, according to which formal relations are by default rendered with an all-some semantics. First we could dispose of the uniformity condition for the **has_grain** relation and allow that a collective may have grains of different sorts, e.g.:

```
BloodPlasmaSample subclassOf
  has_grain some AlbuminMolecule
BloodPlasmaSample subclassOf
  has_grain some GlobulinMolecule
```

This, however, is a severe modification of the underlying logical structure that weakens the **has_grain** relation and puts at risk the expressive power gained through the introduction of that relation in the first place. – Second, we could refrain from a sortal distinction between the grains and subsume them under their most specific common superspecies, e.g.:

```
BloodPlasmaSample subclassOf
  has_grain some PlasmaProteinMolecule
```

Third, we define mixtures as *compounds of fractions*, each fraction in turn being a collective consisting out of grains of the same sort. With option 1 we would sacrifice the ontological purity of the proposed notion of grainhood, although its simplicity would offer some advantages for ontology engineers. Option 2 would need backing by a representation of which molecules are in fact plasma protein molecules, and it would lead to extreme compromises such as to regard red blood cells and glucose molecules as entities of the same kind when describing blood. Option 3 would be the one that is most consistent with the approach proposed in this paper, as it neatly distinguishes compounds from collectives. The conceptualization of mixtures as compounds of fractions is cognitively adequate to the ways biologists and chemists think. For instance, the concept of substance concentration

hinges on a view of the proportion of different substance fractions.

Option 3, however, is prone to produce interoperability problems. In order to demonstrate this, we will discuss the use case of propanol, which is common disinfectant, a type of organic molecule which can further be refined into the isomers *I-Propanol* and *N-Propanol*.

For sake of readability we use description logics notation [1] and do not extend the expressiveness beyond the standard OWL-QL [11], in accordance with the capability of descriptions logics reasoners like Hermit, as well as the expressiveness of BioTop⁴.

One way to describe a collection of propanol, regardless which isomer it contains, is the following:

```
Propanol_Coll equivalentTo
  (has_grain some Propanol_Molecule) and
  (has_grain only Propanol_Molecule)
```

Accordingly,

```
I-Propanol_Coll equivalentTo
  (has_grain some I-Propanol_Molecule) and
  (has_grain only I-Propanol_Molecule)
```

```
N-Propanol_Coll equivalentTo
  (has_grain some N-Propanol_Molecule) and
  (has_grain only N-Propanol_Molecule)
```

A mixture of *I-Propanol* with *N-Propanol*, i.e. a compound of two fractions which are collectives is then represented as follows:

```
Propanol_Mixture equivalentTo
  (has_component some N-Propanol_Coll) and
  (has_component some I-Propanol_Coll) and
  (has_component only (Propanol_Coll or not(Molecule)))
```

The first two conjuncts assure that the mixture contains at least one molecule of n-propanol and at least one molecule of i-propanol. The third conjunct guarantees that the only molecules in the mixtures are propanol molecules.

Propanol mixtures should be classified as Propanol collections, because they contain only propanol molecules. In order to achieve interoperability between descriptions of different specificity levels, it would therefore be desirable that the subsumption of *Propanol_Mixture* by *Propanol_Coll* be computed by logical reasoning. This can, however, not directly be implemented with the available description logics, because the criteria that distinguish **has_grain** and **has_component** from its superrelation **has_proper_part** are not expressible in description logics. A practical solution which supports the desired inferences has to refrain from the use of these subrelations of **has_proper_part**. It is, however, possible to define *Propanol_Mixture* without these relations:

⁴ The ontology is available at <http://purl.org/biotop/src/propanol.owl>.

Propanol_Mixture equivalentTo
 (has_proper_part some *N-Propanol_Molecule*) and
 (has_proper_part some *I-Propanol_Molecule*) and
 (has_proper_part only (*Propanol_COLL_* or
 proper_part_of some *Propanol_COLL_*)))

Collectives can be defined analogously along the following line:

Propanol_Coll equivalentTo
 (has_proper_part some *Propanol_Molecule*) and
 (has_proper_part only
 (*Propanol_Molecule* or not (*Molecule*)))

Once collections are mixtures and collectives are defined in this way, it trivially follows that all propanol mixtures are propanol collectives, too. A description logics classifier can automatically compute a subclass relation between *Propanol_Mixture* and *Propanol_Coll*, and we thus assure cross-granular interoperability.

5 CONCLUSION

Previous work had suggested that collectives and compounds are disjoint categories. Their disjointness, however, relies on agreement about what are uniform entities, and thus on the specificity of description. While the **has_proper_part** subrelations **has_grain** and **has_component** can be used to characterize biomedical entities like mixtures, description logic reasoners fail to calculate equivalence or subsumption relations between different descriptions of same or similar entity classes. We demonstrated that using a different ontology design pattern that avoids primitive subrelations at the expense of more complex descriptions supports the needed inferences using description logics. Our example also provides evidence that compounds and collectives are no disjoint categories.

Although the distinctions between grains and components seems dispensable for the use cases for which description logics representations are adequate, relations such as **has_grain** and **has_component** are nevertheless useful for precise ontological descriptions, e.g. through providing background knowledge for the human modeler.

ACKNOWLEDGEMENTS

Research for this paper has been supported by the German Research Foundation (DFG) through the project “Good Ontology Design” (GoodOD) that is conducted cooperatively at the Universities of Rostock and Freiburg. Many thanks to the anonymous referees for valuable comments.

REFERENCES

1. Baader, F., Calvanese, D., McGuinness, D.L., Nardi, D., Patel-Schneider, P.F. (2010) *The Description Logic Handbook. Theory, Implementation, and Applications*. 2nd edition.: Cambridge University Press, Cambridge.
2. Beisswanger, E., Schulz, S., Stenzhorn, H., and Hahn, U. (2008) BioTop: An upper domain ontology for the life sciences – a description of its current structure, contents, and interfaces to OBO ontologies. *Applied Ontology*, **3**, 202–212.
3. Cohn, A.G. (1997) Qualitative Spatial Representation and Reasoning with the Region Connection Calculus. *GeoInformatica*, **1**, 275–316.
4. Cohn, A.G. (2001) Formalising bio-spatial knowledge. In: C. Welty and B. Smith (eds.), *Proceedings of the International Conference on Formal Ontology in Information Systems* (FOIS 2001), ACM Press, 198–209.
5. Schulz, S. and Hahn, U. (2005) Part-whole representation and reasoning in biomedical ontologies. *Artificial Intelligence in Medicine*, **34**, 179–200.
6. Schulz, S. and Jansen, L. (2009) Molecular Interactions. On the Ambiguity of Ordinary Statements in Biomedical Literature. *Applied Ontology* **4**, 21–34.
7. Schulz, S., Kumar, A., and Bittner, T. (2006) Biomedical ontologies: What part-of is and isn't. *Journal of Biomedical Informatics*, **39**, 350–361
8. Schulz, S., Beisswanger, E., Hahn, U., Wermter, J., Kumar, A., and Stenzhorn, H. (2006) From GENIA to BioTop. Towards a top-level Ontology for Biology. In: B. Bennett and C. Fellbaum (eds.), *Proceedings of the International Conference on Formal Ontology in Information Systems* (FOIS 2006), IOS, Amsterdam, 103–114
9. Smith, B., Ceusters, W., Klagges, B., Köhler, J., Kumar, A., Lomax, J., Mungall, C., Neuhaus, F., Rector, A.L., and Rosse, C. (2005) Relations in Biomedical Ontologies, *Genome Biology*, **6**, R46.
10. Smith, B., Ashburner, M., Rosse, C., Bard, J., Bug, W., Ceusters, W., Goldberg, L.J., Eilbeck, K., Ireland, A., Mungall, C.J., The OBI Consortium, Leontis, N., Rocca-Serra, P., Ruttenberg, A., Sansone, S.-A., Scheuermann, R.H., Shah, N., Whetzel, P.L., and Lewis, S. (2007) The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration. *Nature Biotechnology*, **25**, 1251–1255.
11. Motik, B. et al. (2009) OWL 2 Web Ontology Language Document Overview. available at: <http://www.w3.org/TR/owl2-overview>, retrieved May 2010.
12. Varzi, A.C. (2010) Mereology. In: E.N. Zalta (ed.), *Stanford Encyclopedia of Philosophy*, revised version Stanford: The Metaphysics Research Lab, <http://plato.stanford.edu/archives/spr2010/entries/mereology/>.

Anatomy Ontologies and Potential Users: Bridging the Gap

Ravensara S. Travillian¹, Tomasz Adamusiak¹, Tony Burdett¹, Michael Gruenberger², John Hancock³, Ann-Marie Mallon³, James Malone¹, Paul Schofield², and Helen Parkinson¹

¹ EMBL-EBI, Wellcome Trust Genome Campus, Hinxton, Cambridgeshire CB10 1SD, UK; ² Department of Physiology, Development, and Neuroscience, University of Cambridge, Downing Street, Cambridge CB2 3DY, UK; ³ MRC Mammalian Genetics Unit, Harwell, Oxfordshire OX11 0RD, UK

ABSTRACT

Motivation: To evaluate how well current anatomical ontologies fit the way real-world users apply anatomy terms in their data annotations. **Methods:** Annotations from 3 diverse multi-species public-domain datasets provided a set of use cases for matching anatomical terms in two major anatomical ontologies (the Foundational Model of Anatomy and Uberon), using two lexical-matching applications (Zooma and Ontology Mapper). **Results:** Approximately 1500 terms were identified; Uberon/Zooma mappings provided 286 matches, compared to the control and Ontology Mapper returned 319 matches. For the FMA, Zooma returned 312 matches, and Ontology Mapper returned 397. **Discussion:** Our results indicate that for our datasets the anatomical entities or concepts are embedded in user-generated complex terms, and while lexical mapping works, anatomy ontologies do not provide the majority of terms users supply when annotating data. Provision of searchable cross-products for compositional terms is a key requirement for using ontologies.

Supplementary information and data are available at <http://www.ebi.ac.uk/~raven/>.

1 INTRODUCTION

The need for anatomy ontologies to support researchers and clinicians in managing the explosion of experimental data is well-documented (Bard 2008). Four of the 8 Open Biological and Biomedical Ontologies (OBO) Foundry ontologies, and 39 of the site's other listed ontologies, cover aspects of the representation of anatomical knowledge. Whilst the OBO Foundry has a stated goal of “creating a suite of orthogonal interoperable reference ontologies in the biomedical domain” [<http://www.obofoundry.org>], most of these ontologies have been developed to address a species-specific need articulated by a community working with a particular model organism—for example, ZFIN by the zebrafish researcher community (Sprague 2003), FlyBase by the *Drosophila* researcher community (Drysdale 2008), and the various databases dedicated to different ways of representing the needs of mouse researchers (Smith 2007). As a result, despite the number of ontologies and the amount of knowledge represented within them, their potential for integration of data across species has not yet been realised. We undertook this study—matching terms from real-world data annotations to two major anatomical ontologies—to illuminate some of the

existing obstacles to such integration, and to propose ways of making the ontologies more usable to real-world users.

The motivation behind this study was to determine how well current ontologies fit the way users apply anatomy terms in their knowledge domains. We wished to explore whether there was a gap between available ontologies and the use cases for functional genomics data annotation. We also wished to demonstrate our process to show where any gaps and omissions lie.

This paper describes the results of a test of lexical matching anatomical terms entered by current users of an active database of experimental annotation to two major state-of-the-art anatomy ontologies. The implications of our results for the current state of usability for those ontologies in the users' experience are described. We go on to propose methods to bridge the gap between ontologies and the use cases they are intended to support.

2 METHODS

Annotations from 3 diverse public domain datasets—gene expression in the ArrayExpress Archive, Gene Expression Atlas (Parkinson 2009, Kapushesky 2010), the Europhenome mouse-specific databases (Green 2005, Mallon 2008, Morgan 2010), and the multi-species radiobiology database ERA-PRO (Gerber 2006, Tapio 2008)—were selected to provide a set of representative use cases for matching anatomical terms to terms in two major anatomical ontologies, the Foundational Model of Anatomy (FMA) and a species-agnostic ontology (Uberon) (Washington 2009).

2.1 Ontologies

The FMA is a symbolic representation of human anatomy, containing over 75,000 concepts over a range of granularity [<http://sig.biostr.washington.edu/projects/index.html#FMA>]. Uberon is a multi-species metazoan anatomy ontology with the goals of: (1) supporting translational research by allowing comparison of phenotypes across species, and (2) providing logical cross-product definitions for GO biological process terms (Haendel 2009). These ontologies were chosen for comparison on the basis of size (thus, likelihood of coverage) and importance in the anatomy ontology community.

2.2 Data Preprocessing

The ArrayExpress Gene Expression Archive [<http://www.ebi.ac.uk/gxa/>] currently contains almost 350,000 assays, covering approximately 700 different species. Of these assays, those which meet a particular standard of quality and clarity are re-annotated with ontology classes, summarized, and stored in the Gene Expression Atlas, which currently contains 39,633 assays from 22 species. Europhenome is an online resource developed to capture phenome data derived from mice using the standardised tests contained in EMPReSS (the European Mouse Phenotyping Resource of Standardised Screens) (Green 2005, Mallon 2008, Morgan 2010). The ERA-PRO database is a legacy database of the results of radiation exposure of a wide range of organisms. It contains data on more than 300,000 animals from experiments conducted in Europe, the USA and Japan from 1954-1996 (Gerber 2006, Tapio 2008).

Ontology-curated annotations from the Atlas and non-standardised ones from the Archive, along with anatomy terms from Europhenome and ERA-PRO, constituted the sample that we compared against the ontologies. There were 1537 terms in our initial sample of anatomical structures by vertebrate species.

Although these represent the set of terms as they were actually submitted by the users who annotated the use cases, we performed a small amount of manual and semi-automated preprocessing. The terms were normalised along the following lines: Species names were removed from the anatomical structures to be searched. Anatomical structure terms that had been unique only because they had different species names were resolved into a single term to remove duplicates. (This practice may have increased the risk of false positives; however, we did not formally investigate that possibility.) Inconsistent use of spaces within terms was resolved. Obvious misspellings and typographical errors, inconsistent punctuation and hyphenation were resolved. Anglicisms were converted to American spellings.

This preprocessing was carried out in order to cut down on the number of false positive matches due only to duplication, and the number of false negative match failures due to: 1) Obvious errors in the data resulting in “terms” that would not exist in any anatomy ontology, 2) duplications due to different use of spaces, sentence casing, or punctuation and hyphenation, and 3) the emphasis of both the FMA and Uberon on American spellings as the primary term.

Whilst this is not strictly verbatim annotation data, this minimal preprocessing represents simple data cleanup for which tools are readily available, so we consider it a representative model of how data would actually be submitted for matching against an ontology after automatic cleanup. After preprocessing, 1311 terms of the original 1537 remained. Inspection established that 229 terms exactly matched terms or synonyms in FMA, and 277 exactly matched terms or

synonyms in Uberon. We used these matches as test controls before comparing the dataset to the ontologies.

2.3 Comparison Tools

Zooma is an automatic ontology mapping application developed at the European Bioinformatics Institute (EBI) [<http://zooma.sourceforge.net/>]. Given a list of terms and an ontology in OWL/OBO format, or an ontology service to match against, Zooma can automatically map terms from the list to ontology terms. It can re-use existing mappings *e.g.*, from a database, discover mappings for new terms, and search many ontologies to propose mappings against terms not mapped to a reference set. It produces two time-stamped tab-delimited text files, one containing successful matches, and one containing matching failures. Zooma delegates all ontology term fetching to the OntoCAT library [<http://sourceforge.net/projects/ontocat/>]. This library provides a uniform interface to query heterogeneous ontology resources including local ontologies in OWL or OBO as well as public ontology repositories, such as NCBO BioPortal and EBI Ontology Lookup Service.

Ontology Mapper [<http://www.ebi.ac.uk/efo/tools>] is a Perl module, based on the Metaphone and Double Metaphone algorithms (Philips 2000), that normalizes the terms in the list and the ontology to a representation of their sounds, and carries out the matching of similar-sounding terms. It performs exact matches without intervention, and proposes single or multiple approximate matches to the user for curator driven matching. Our dataset was run once against both the FMA and Uberon using Zooma, and once against both ontologies using Ontology Mapper.

3 RESULTS

For Uberon, Zooma provided 286 matches, compared to the control (visual inspection) of 277 exact matches. Ontology Mapper returned 319 matches, with user curation. For the FMA, Zooma returned 312 matches, and Ontology Mapper returned 397. Precision/recall for each iteration are provided in Table 1. We examine these results in the Discussion.

Table 1. Precision and recall for matching results

	OM/FMA	OM/U	Z/FMA	Z/U
Precision	0.007	0.014	1.000	1.000
Recall	0.335	0.429	0.263	0.385

FMA=Foundational Model of Anatomy, OM=Ontology Mapper, U=Uberon, Z=Zooma

4 DISCUSSION

The first question that emerges from these results is why are there so few matches between such rich ontologies and real-life anatomical annotations from the user community? Some of the discrepancies are accounted for by the fact that the

tools were designed for strict matching to reduce the number of false positives during automated lexical processing. However, the actual mismatches between the terms as the users provide them, and as they are represented in the ontologies accounted for much of the low hit rate, and, thus, the very low precision and recall values. The strictness of the matching algorithm is a by-product of the fact that we wanted a low rate of false positives due to the known heterogeneous nature of the input. The aim is to automate these processes of mapping in future and for the use cases we have, low false positives is important. In future work we can extend the use of methods to use more fuzzy matching techniques *e.g.*, Levenshtein distance-based methods, and which will improve recall for adjectival forms.

4.1 Annotation issues

The most obvious issues with the data that would block matching to the ontologies were standardised by preprocessing before the test was run. However, even after preprocessing, there were recurring problems that interfered with successful matching.

Most of the mismatches were due to annotations containing composite terms, where one or both of the terms, taken separately, actually would have matched. For example, *Liver/Kidney* does not match any entity in either ontology, but *Liver* alone and *Kidney* alone matched in both. Sometimes only one of the terms would have matched; for example, in *Acetabulum and pelvic soft tissues*, while *pelvic soft tissues* is too vague and would have required clarification, *Acetabulum* could have been an exact match. Tools that are able to split up the user's annotations into individual anatomical terms, or even mine the user's annotations for new anatomical terms, can resolve some of these problems.

Sometimes it was unclear whether the composite term was actually referring to two different entities, or whether it was simply a redundancy, for example, *Adrenal cortex, adrenal gland*. Breaking down that composite term would have accomplished either one or two matches, depending on whether the user intended two different entities, or was using *Adrenal gland* redundantly to modify *Adrenal cortex*.

Occasionally the user would refer to a cell when it was obvious the sample referred to a tissue (example: *Adipocyte* vs. *Adipose tissue*), and whilst the cell name the user entered did not match, the intended tissue type actually would have.

Entities in anatomical ontologies are nouns for the most part, so when the user entered an adjective referring to the anatomical structure, such as *Abdominal* for *Abdomen*, or *Arterial* for *Artery*, the adjective would not match the ontology, yet was closely related to a term that was present.

Shorthand, such as *Antrum* for *Pyloric antrum*, or *Both ventricles* for *Left ventricle of heart* and *Right ventricle of heart* also kept the matching rate artificially low. Most of those examples could be expanded to full names of entities in the anatomy ontologies, but out of context, it was impossible to

determine what a few meant, such as *Ventral* or *11 different tissues*. If a human curator cannot know what the annotation means, mapping becomes an impossible task for any automatic or curated tool.

Occasionally match failures occurred from the users' utilization of named processes implying an anatomical location, such as *Colon pinch biopsy*, *BA [bronchioalveolar] lavage*, or *Bone marrow, flushed from femur*. The latter could be a composite term as well, although breaking it down into entities would require the tool to know how to deal with "flushed from", in order to find the boundary between them.

4.2 Ontology issues

Sometimes a mismatch would occur because the user made use of a synonym for a term that was actually in the ontology, but synonyms were missing. Sometimes an omission from an ontology was quite surprising—for example, the term *Anterior tibialis* was missing from Uberon and FMA. Addition of synonyms to both would improve utility.

Ontology classes tend to be expressed in the singular whilst annotations are written in singular and plural. Both tools did a better job of matching the singular terms than they did the same terms in plural; given the variation in working styles, tools that access ontologies for real-world applications will need to be better at dealing with singular-plural variation.

The FMA and Uberon are designed for different purposes, so there is no consensus between them as to exactly what entities belong in an anatomical ontology. For example, *Alveolar macrophage* is included as a discrete entity in FMA, where Uberon does not contain it, and explicitly regards it as a composite to be generated from *Alveolus* and *Macrophage*, rather than belonging in the ontology itself. These differing definitions based on design decisions, while not apparent to the user, have an impact on whether that user's terms can be expected to match terms in the ontologies being used.

Related issues of definition of scope are behind the result that Uberon handles embryological and non-human anatomical entities better than the FMA does, while the FMA is so term-rich that—curated—it provides more overall matches, a different result than for the uncurated terms, where Uberon provided more matches. For this reason, Uberon performed relatively better in matching the heterogeneous data from this community.

4.3 Mutual mismatch issues

The implicit assumption designed into the tools is that the mapping between list terms and ontology terms should be 1:1. This meant that there were sometimes approximate matches on a closely-related term, even in the absence of an exact match. For example, sometimes there was a superset or subset relationship approximate match, such as *Abdominal fat* and *Abdominal fat pad*, or *Fascia* and *Connective tissue*. Other proposed matches crossed levels of abstraction, for example, *Right lung* as opposed to *Lung*.

Additionally, a few cases of quantitative annotations, such as 75% kidney, 25% liver were present in the annotations; these were ruled out of scope, as very few present ontologies can handle quantitative data. However, it does indicate a currently-unmet user need in data annotation.

We have established that, although we were able to map almost half of the terms from the use cases to the ontologies, the process required a great deal of time, effort, and manual curation. There remains a vast gap between the way users use anatomical terms in free text annotation and the way they are represented in two of the richest anatomical ontologies. This exercise provided preliminary insight into the following issues:

- Which terms are available in which source(s);
- Which areas require concentration in ontology development in order to obtain as much coverage as other areas have;
- What maps, what does not map, and why;
- What duplications and errors our tools are able to determine in the ontologies used in the comparison;
- What suggestions we would make for additions and modifications to the source ontologies;
- What suggestions we want to make to the tool developers for functionality that would make it easier for users to obtain better matches.
- The need for error detection in source ontologies;
- The requirements for cross-products in many annotation use cases.

This insight will inform our future efforts in developing and refining ontology-matching tools.

In order for ontologies to realise their potential, they need to be used. The user must perceive the benefit from their use, whether that benefit takes the form of ease of data entry, time saved, replacement of manual inspection with automation, and so forth. The current state of anatomical ontologies leaves a gap between the needs of the user and what the ontologies are available. There is a real and growing need for tools such as Zooma, Ontology Mapper, and others that can complement the functions of ontologies in bridging that gap, removing barriers between the ontology and the community of users it is intended to serve.

5 ACKNOWLEDGEMENTS

We thank Onard Mejino, Todd Detwiler, and Melissa Haendel for providing support for UBERON and FMA. *Funding:* Gen2Phen (contract #200754), European Molecular Biology Laboratory, Biotechnology and Biological Sciences Research Council (grant #BB/G022755/1), Medical Research Council, European Commission's FP6 Programme (contract #LSHG-CT-2006-037188) for EuroPhenome as part of the EUMODIC project.

Conflict of Interest: none declared.

REFERENCES

- Bard J. (2008) Anatomical ontologies for model organisms. In: *Anatomy Ontologies for Bioinformatics*. Springer, London.
- Drysdale R. FlyBase Consortium. (2008) FlyBase: a database for the Drosophila research community. *Methods Mol Biol*, **420**, 45-59.
- Gerber GB, Wick RR, Kellerer AM, Hopewell JW, Di Majo V, Dudoignon N, Gössner W, Stather J. (2006) The European Radiobiology Archives (ERA)--content, structure and use illustrated by an example. *Radiat Prot Dosimetry*, **118(1)**, 70-7.
- Green EC, Gkoutos GV, Lad HV, Blake A, Weekes J, Hancock JM. (2005) EMPReSS: European mouse phenotyping resource for standardized screens. *Bioinformatics*, **21(12)**, 2930-1.
- Haendel M, Gkoutos G, Lewis S, and Mungall C. (2009) Uberon: towards a comprehensive multi-species anatomy ontology. Available from Nature Precedings <<http://dx.doi.org/10.1038/npre.2009.3592.1>>
- Kapushesky M, Emam I, Holloway E, Kurnosov P, Zorin A, Malone J, Rustici G, Williams E, Parkinson H, Brazma A. (2010) Gene expression atlas at the European bioinformatics institute. *Nucleic Acids Res.*, **38**, D690-8.
- Mallon AM, Blake A, Hancock JM. (2008) EuroPhenome and EMPReSS: online mouse phenotyping resource. *Nucleic Acids Res.*, **36**, D715-8.
- Morgan H, Beck T, Blake A, Gates H, Adams N, Debouzy G, Leblanc S, Lengger C, Maier H, Melvin D, Mezziane H, Richardson D, Wells S, White J, Wood J; EUMODIC Consortium, de Angelis MH, Brown SD, Hancock JM, Mallon AM. (2010) EuroPhenome: a repository for high-throughput mouse phenotyping data. *Nucleic Acids Res.*, **38**, D577-85.
- Parkinson H, Kapushesky M, Kolesnikov N, Rustici G, Shojatalab M, Abeygunawardena N, Berube H, Dylag M, Emam I, Farne A, Holloway E, Lukk M, Malone J, Mani R, Pilicheva E, Rayner TF, Rezwan F, Sharma A, Williams E, Bradley XZ, Adamusiak T, Brandizi M, Burdett T, Coulson R, Krestyaninova M, Kurnosov P, Maguire E, Neogi SG, Rocca-Serra P, Sansone SA, Sklyar N, Zhao M, Sarkans U, Brazma A. (2009) ArrayExpress update--from an archive of functional genomics experiments to the atlas of gene expression. *Nucleic Acids Res.*, **37**, D868-72.
- Philips L. (2000) The Double Metaphone Search Algorithm, *C/C++ Users Journal*, **18(6)**.
- Smith CM, Finger JH, Hayamizu TF, McCright IJ, Eppig JT, Kadin JA, Richardson JE, Ringwald M. (2007) The mouse Gene Expression Database (GXD): 2007 update. *Nucleic Acids Res.*, **35**, D618-23.
- Sprague J, Clements D, Conlin T, Edwards P, Frazer K, Schaper K, Segerdell E, Song P, Sprunger B, Westerfield M. (2003) The Zebrafish Information Network (ZFIN): the zebrafish model organism database. *Nucleic Acids Res.*, **31(1)**, 241-3.
- Tapio S, Schofield PN, Adelman C, Atkinson MJ, Bard JL, Bijwaard H, Birschwilks M, Dubus P, Fiette L, Gerber G, Gruenberger M, Quintanilla-Martinez L, Rozell B, Saigusa S, Warren M, Watson CR, Grosche B. (2008) Progress in updating the European Radiobiology Archives. *Int J Radiat Biol.*, **84(11)**, 930-6.
- Washington NL, Haendel MA, Mungall CJ, Ashburner M, Westerfield M, Lewis SE. (2009) Linking human diseases to animal models using ontology-based phenotype annotation. *PLoS Biol.*, **7(11)**, e1000247.

The Ontology of Primary Immunodeficiency Diseases (PIDs) – Using PIDs to Rethink the Ontology of Phenotypes

Nico Adams^{1,3}, Christian Hennig², Robert Hoehndorf^{1,3}, Anika Oellrich¹, Dietrich Rebholz-Schuhmann¹, Gesine Hansen²

¹European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SD, United Kingdom,

²Department of Paediatric Pneumology, Allergology, and Neonatology, Hannover Medical School, Carl-Neuberg-Strasse 1, D-30625 Hannover, Germany,

³Department of Genetics, University of Cambridge, Downing Street, Cambridge CB2 3EH, United Kingdom

ABSTRACT

Primary immunodeficiency diseases (PIDs) are the consequence of genetic disorders and usually manifest themselves in very young patients. Because of their rarity, they are notoriously difficult to diagnose both for general practitioners and clinicians. In this paper, we present the foundations of an ontology of PIDs, which will be at the heart of an expert system designed to assist the clinician in the diagnosis of these diseases. To achieve this, the PIDOntology characterises Primary Immunodeficiencies in terms of Phenotypes. While there are a number of different ontologies already available that allow the description of phenotypes and phenotypic qualities, these have a number of associated ontological problems, which we will also address as part of this paper. We use the subtype of Hyper-IgE Syndrome caused by a STAT3 defects as an example of a primary immunodeficiency and show how the clinical phenotype of the disease can be modeled in terms of other phenotypes by introducing the notion of the “phene”. Furthermore, we develop patterns for different types of phenes and show, that these patterns can be mapped onto more traditional entity-quality statements, which are the current state of the art in phenotypic modeling.

1 INTRODUCTION

Primary Immunodeficiency Diseases (PIDs) are a group of disorders caused by the absence of or by defects in genes involved in regulating and coordinating the body’s immune system. Their incidence varies from relatively common (1:1,200) to extremely rare (1:2,000,000) [Lim and SJ, 2004]. PIDs can manifest themselves in newborns and toddlers, but also much later in life, which makes their identification and diagnosis difficult for both the general practitioner and the experienced clinician alike. The fact that primary immunodeficiencies represent a large and diverse group of disorders complicates the picture even further: currently, the International Union of Immunological Societies’ classification recognises 140 different PIDs [Geha et al., 2007], although more than 200 primary immunodeficiencies have now been identified.

Informatics resources that contain condensed and structured information, which can be used for the diagnosis of PIDs remain relatively scarce, and include the ImmunoDeficiency Resource [Vliaho et al., 2002] and INFO4PI [Foundation, 2010]. A recent review provides a comprehensive overview over the existing bioinformatics services relating to PIDs [Samargithea and Vihinen, 2009]. Furthermore, few of the available services for PIDs leverage the power of semantic web technologies, and allow the semantic codification of knowledge.

1.1 Why an ontology of Primary Immunodeficiency Diseases?

One important component of the semantic web are ontologies, which are formal and – importantly – computable specifications of a shared conceptualisation. The overall aim of our research is the development of an ontology-driven expert system for clinicians, which can assist in the diagnosis of primary immunodeficiency diseases.

Apart from diagnostic tools, we also expect that the ontology will be useful in a number of other areas in the future: it is reasonable to assume, that it will form one of the components for semantically-rich publishing of academic work relating to research in primary immunodeficiency disorders and that it will also be useful for the integration with knowledge contained in other ontologies as well as the integration of existing and new data resulting from the development of relevant new research and laboratory techniques such as iterative chip-based cytometry [Hennig et al., 2009]. In this contribution, we report on the development of the underlying ontological principles of the PID Ontology. In the future, we will address the delivery of an ontology-driven and web-delivered expert system as well as formal methods for the comparison of observed and formally codified phenotypes for primary immunodeficiency diseases.

2 METHODS

2.1 Constructing the PID Ontology

The scope of the PID ontology is to serve as both a reference and an application ontology, which defines a phenotype of a primary immunodeficiency disorder in terms of a set of biomarkers. There is little restriction on the type of biomarker the ontology references. Typical biomarkers are, for example, gene defects, disorders, clinical or laboratory findings and genetic inheritance patterns. Collectively, these biomarkers form the “canonical” phenotype of a primary immunodeficiency disorder.

The PID ontology has been formalized in the OWL 2 ontology language, which has a mechanism for annotating assertions in the ontology. This is of critical importance for the development of the ontology: every assertion relating a primary immunodeficiency to a symptom is annotated with the literature source, from which the assertion was derived by a human curator. An example implementation of the ontology can be downloaded from <http://bitbucket.org/na303/pidontologyexample/>.

In the development of the ontology we aim to be term-orthogonal with respect to other established ontologies in the biomedical domain – in particular those developed under the umbrella of the OBO Foundry. Furthermore, we have decided to use the General Formal Ontology (GFO) [Herre et al., 2006] as an upper ontology due to its integration of objects and processes, its rich classification of roles [Loebe, 2005] and the expressive axioms in its OWL version.

2.2 Phenotypic characters and phenes

The dominant view in the biomedical ontology community is that phenotypic characters are qualities of entities. This is exemplified in the Entity-Quality (EQ) formalism [Mungall et al., 2010] and the databases that use EQ for their annotations, as well as in the PATO ontology [Gkoutos et al., 2004] which provides the qualities that **inhere in** an entity.

We note, however, that there are at least two fundamentally different types of qualities in the PATO: qualities of objects that can be *discovered* by an observer, and qualities of objects that are *established* by an observer. Qualities of the first kind are qualities like *red*. Qualities that are *established* by an observer are all those qualities that are defined with respect to an explicit or implicit *norm*: *increased size* or *lacking parts*.

In particular, there is a kind of quality in the PATO that should *not* be considered to be an ontological quality. PATO contains the category *count* with sub-categories *present* and *absent*. Yet *absent* cannot be a quality of some entity in the same sense as *weight* or *shape* can be qualities of an entity: qualities are existentially dependent on their

bearer, they cannot exist without the entity of which they are a quality. But in the case of *absent*, such a bearer does not exist [Hoehndorf et al., 2010]. *Absent* was introduced in PATO in order to facilitate the curators' need to annotate absent parts. We provide a formal treatment of absence using a relation to allow the inference of *has-part* and *lacks-part* relations from these qualities. This addresses the difficulty of using EQ formalisms to establish an information flow between an assertion about an entity's qualities to statements involving propositions about the parts, qualities or dispositions the entity has.

Additionally, having or lacking parts is not *primarily* a quality: first, there is the absence of parts, *not having an X as part*, a structural feature expressed using negation and the **part-of** relation, a feature of an organism that is distinct from an ontological quality.

Therefore, although qualities in the sense of PATO certainly seem to play an important role in phenotypes, and reference to qualities is often necessary to describe phenotypes, phenotypes may not always be qualities, at least not in the sense of PATO, as existentially dependent entities inhering in a bearer. To formalize PIDs, we use the EQ formalism to provide compatibility with the rich resources already available in this formalism, and also use a new method for the semantic representation of phenotypes that interoperates with OWL and Semantic Web technology and permits making the semantics of phenotypic descriptions explicit in these formalisms. To set these semantic phenotype descriptions apart from the EQ phenotypes, we call them *phenes* [Allan, 2008] in the context of this paper.

A *phene* is a basic observable characteristic possessed by an organism. Phenens are attributive individuals in the sense that they are existentially dependent on a bearer, and they are related to their

bearer by the **pheneOf** relation. The **pheneOf** relation has an inverse which we call **hasPhene**.

We use a general definition pattern for phenens. This pattern is based on the observation that having a phene means exhibiting certain features and properties. We define a category of phenens *X* as the category of phenens “of entities with the property *Y*”. The property *Y* is expressed as class-membership in description logic or unary predicates in first-order logics. The analysis of the ontological status *Y* is outside the scope of this manuscript.

```
X EquivalentClass Phene and pheneOf some Y
```

2.3 Biomarkers and Biomarker Roles

The notion of an entity acting as a biomarker is central to the PID ontology. Biomarkers are phenens that characterise variation in cellular or biological components, pathways, etc., and where the variation is objectively measured and observed. Therefore, biomarkers are phenens participating in clinical diagnosis processes in the biomarker role. The form of participation is dependent on the observer, and the PID ontology contains a classification of biomarkers based on the kind of role they play in the observation process. Generally, we may define a “biomarker” as

```
Phene and (plays-role some Biomarker_Role).
```

By specifying the type of observation process, we may further define the type of biomarker. An “imaging biomarker”, for example, is a biomarker, which is observed in a radiological observation process (e.g. projection radiography or Computed Tomography Scanning). By analogy, a cellular biomarker is a biomarker observed during a cytometric experiment. The PID ontology will provide an extensive hierarchy of biomarkers, which is useful for the further classification of phenens and any one phene will be able to assume multiple biomarker roles. In the first instance, the classification of phenens in terms of biomarkers will mirror the way in which most clinicians classify phenens. However, due to the definition of biomarkers via the role mechanism, the axiomatic construction of inferred polytaxonomies is eminently feasible. The formal integration of PID ontology with diagnostic processes is the subject of future work.

2.4 Diseases and syndromes

Although the PID ontology uses terms referring to syndromes and diseases, it does not actually classify diseases or syndromes themselves. Instead, the PID ontology is an ontology of the *disease phenotype* of primary immune deficiency diseases.

In other ontologies, diseases are sometimes classified as processes, dispositions or qualities [Scheuermann et al., 2009]. We hold that a phenotypic description of diseases is more general than each approach. Using phenens permits the representation of dispositions, qualities, processes and other attributes of organisms in a single coherent framework.

2.5 Relation to other ontologies

Whenever possible, we reuse terms from ontologies in the OBO and OBO Foundry. In particular, we use the Foundational Model of Anatomy [Rosse and Mejino, 2003], PATO [Mungall et al., 2010], the Human Phenotype Ontology [Robinson et al., 2008], Mouse Pathology Ontology and OBI [Courtot et al., 2008].

In addition to defining PIDs in terms of phenens using OWL, we provide EQ definitions of the biomarkers in the PID ontology. These serve to facilitate compatibility with those resources that have

been annotated using the EQ formalism. Additionally, we document transitional patterns for formally defining the EQ statements using the method introduced here.

3 DISCUSSION

3.1 Use-case: Hyper-IgE syndrome caused by STAT3 Defects

The Hyper-IgE syndrome (HIES), sometimes also known as “Job’s syndrome” is a collective term for a set of complex immunodeficiencies. First reported by Davis et al. [1966], it is characterised by increased serum IgE levels, chronic dermatitis and serious recurrent infections such as pneumonia and recurrent staphylococcal skin abscesses. Staphylococcal infections can also affect lungs, joints and other sites. Other signs and symptoms which have been observed are atypical eczema, pneumatoceles, and osteopenia. Furthermore, patients often have fair skin and red hair as well as “lion-like” facial features, caused by a high palate. The inheritance pattern is autosomal dominant or recessive and patients affected by the autosomal dominant form often fail to shed their primary teeth. Additionally, some patients also suffer from scoliosis [Grimbacher et al., 1999]. The presentation of Hyper-IgE syndrome varies from patient to patient and is also age-dependent. However, no patient will present with the whole gamut of clinical indicators for the disease, that have been observed over the totality of all patients.

We use the Hyper-IgE syndrome caused by STAT3 defects as an exemplar to show how this complex phenotype can be modeled in terms of phenes and how phenes themselves can be considered to be biomarkers in the context of a diagnosis process. We focus on the representation of the *canonical* phenotype in the ontology. Patients, however, can be non-canonical with respect to the canonical phenotype of the Hyper-IgE syndrome.

One phenotypic trait characterizing the syndrome is the absence of Th17 cells. Using the EQ method, this is formalized as

```
[Term]
id: absence_of_Th17_cell
intersection_of: PATO:0001557 ! lacking physical part
intersection_of: towards CL:0000899 ! Th17 cell
```

Using our method, this can be extended and formalized using the phene *Absence of Th17 cells* as

```
Phene and pheneOf some (not (has-part some CL:0000899))
```

which enables the inference that entities with this phene have no Th17 cells as part.

Another phene of the Hyper-IgE syndrome is pneumonia. In the HPO, a *pneumonia* is an inflammation of the lung:

```
[Term]
id: HP:0002090 ! pneumonia
intersection_of: PATO:0001561
! having extra processual parts
intersection_of: inheres_in FMA:7195 ! lung
intersection_of: towards MPATH:212 ! inflammation
```

Inflammation is a process and there are different forms of participation in processes. In the General Formal Ontology, these are modeled using processual roles. In an inflammation process, we

may distinguish between an inflammatory agent and an inflamed structure. In pneumonia, the lung plays the role of the inflamed structure and a definition of the inflammatory agent can be used to further differentiate the type of pneumonia (e.g. staphylococcal pneumonia, where a strain of *Staphylococcus* plays the role of the inflammatory agent). Using our notion of phenes, we may therefore rewrite the EQ definition in the following form:

```
Phene and pheneOf some (has-proper-part
some (FMA:7195 and plays-role
some Inflamed_Entity))
and (plays-role some BiomarkerRole)
```

Inflamed Entity, in turn, can be defined as

```
ProcessualRole and role-of some MPATH:212
```

A processual role is a **role-of** a process, and playing the role implies participation in this process. Furthermore, the use of “has-proper-part” as opposed to “has-part” allows the distinction between a local inflammation (pneumonia is localised in the lung) and global inflammation (e.g. hemophagocytotic syndrome) and, more generally local and global parts.

We can use phenes to model additional parts as well. A related PID, the Wiscott-Aldrich-Syndrome, is characterized by B-cell lymphocytic neoplasms, which would be formalized using EQ and PATO and MPATH as:

```
[Term]
id: B_Cell_Lymphocytic_Neoplasm
intersection_of: PATO:0002002
! having extra physical parts
intersection_of: inheres_in FMA:FMA:20394
! human body
intersection_of: towards MPATH:516 ! B-cell neoplasms
```

The use of the PATO quality *having extra physical parts* hides the semantics of the phene, i.e., that the phene’s bearer has a B-cell neoplasm as **part**. Therefore, we define a phene *Having B-Cell lymphocytic neoplasm* as

```
Phene and pheneOf some
(Human and has-part some MPATH:516)
```

This differs from the EQ statement in that it explicitly establishes a relation between having the neoplasm and the parts of the organism.

3.2 Comparison

Beyond the advantages of phenes and the problems associated with the EQ formalism discussed above, our use of the phene formalism makes much of the implicit semantics contained in entity-quality statements explicit: taking the B-cell lymphocytic neoplasm as an example, the use of the PATO quality “having extra physical parts” hides the semantics of the particular phenotype, namely that the phenotype’s bearer has a B-cell lymphocytic neoplasm as **part**. The use of phenes as shown above explicitly establishes the relationship between having the phene and the parts of the organism that bear it. Furthermore, the phene formalism allows the correct use of qualities: absence (as in absence of Th17 cells) is not a quality as currently modeled by PATO and phenes allow us to model absence as not having a *part*. In spite of the differences in approach, the phene-formalism is “backwards compatible” with the

more traditional EQ approach and should allow the mapping and interconversion of ontologies using either of these two frameworks.

3.3 Future research

Future research will pursue several different strands. Firstly, the continued enrichment of the ontology with content will be the highest priority. As discussed above, the use of relations such as **part-of** in the definition of phenes can lead to inconsistencies when combining these with canonical ontologies. In future research we will address this, by investigating how canonical ontologies formalized in OWL can be restructured using *Canonical* and *Non-canonical* classes.

The appropriate definition of phenes such as “small platelets size” (a phene for Wiskott-Aldrich Syndrome) or “thrombocytopenia” remains an unsolved issue. The former refers to the fact that the average of the platelet size distribution in a “non-canonical patient” is shifted to lower values with respect to that in the “canonical patient” and thrombocytopenia denotes the situation in which the number of platelets in the blood of a “non-canonical” patient is reduced with respect to the number found in a “canonical patient”. The current state of the art is to use the EQ framework, for which, however, no explicit semantics is currently available to formalise concepts such as “small platelet size”. Furthermore, some symptoms of a disease only manifest themselves in a fraction of all patients that share a common diagnosis.

Finally, we will address the development of formal methods for the comparison of canonical phenotypes (i.e. phenotypes encoded in an ontology) and observed phenotypes (the phenotype presented by a patient) and how these comparisons can be used to assist the clinician in the diagnosis of primary immunodeficiency diseases.

4 CONCLUSION

We have developed the ontological basis for the description of primary immunodeficiency syndromes by defining a semantically rich representation of basic observable characteristics in organisms. Building on this, we have developed a view of a primary immunodeficiency disease as a set of complex phenes, described by other simpler phenes. We show that this new formalism and the modeling of phenotypic characters using entity-quality statements are compatible.

The method we use to characterize the canonical disease phenotypes of primary immunodeficiency syndromes can be applied to all forms of disease. Thereby, the PID ontology serves as an example for the development of disease and phenotype ontologies in general. Application of the method used in constructing PID leads to an integration with canonical ontologies, allows for an explicit representation of abnormality and facilitates knowledge-based inferences over observed phenomena.

REFERENCES

Charlotte L. Allan. Schizophrenia: From genes to phenes to disease. *Current Psychiatry Reports*, 10(4), 2008.

Mélanie Courtot et al. The owl of biomedical investigations. In Catherine Dolbear, Alan Ruttenberg, and Ulrike Sattler, editors, *OWLED*, volume 432 of *CEUR Workshop Proceedings*. CEUR-WS.org, 2008.

SD Davis, J Schaller, and Wedgwood RJ. Job’s syndrome : Recurrent, “cold”, staphylococcal abscesses. *Lancet*, 287(7445),

1966.

Jeffrey Modell Foundation. Info4pi, 2010. URL <http://www.info4pi.org>.

R Geha et al. Primary immunodeficiency diseases: An update from the International Union of Immunological Societies Primary Immunodeficiency Diseases Classification Committee. *Journal of Allergy and Clinical Immunology*, 120(4):776–794, 2007.

G. V. Gkoutos, E.C.J. Green, A-M Mallon, J.M. Hancock, and D. Davidson. Using ontologies to describe mouse phenotypes. *Genome Biology*, 6(1):R8, 2004.

B Grimbacher, SM Holland, JL Gallin, F Greenberg, SC Hill, HL Malech, JA Miller, AC O’Connell, and Puck JM. Hyper-IgE syndrome with recurrent infections—an autosomal dominant multisystem disorder. *New England Journal of Medicine*, 340(9): 692–702, 1999.

Christian Hennig, Nico Adams, and Gesine Hansen. A versatile platform for comprehensive chip-based explorative cytometry. *Cytometry, Part A*, 75A(4):362–370, 2009.

Heinrich Herre, Barbara Heller, Patryk Burek, Robert Hoehndorf, Frank Loebe, and Hannes Michalek. General Formal Ontology (GFO) – A foundational ontology integrating objects and processes [Version 1.0]. *Onto-Med Report 8*, Research Group Ontologies in Medicine, Institute of Medical Informatics, Statistics and Epidemiology, University of Leipzig, Leipzig, 2006.

Robert Hoehndorf, Axel-Cyrille Ngonga Ngomo, and Janet Kelso. Applying the functional abnormality ontology pattern to anatomical functions. *Journal for Biomedical Semantics*, 2010. in press.

SM Lim and Elenitoba-Johnson SJ. The molecular pathology of primary immunodeficiencies. *Journal of Molecular Diagnostics*, 6(2):59–83, 2004.

Frank Loebe. Abstract vs. social roles: A refined top-level ontological analysis. In G. Boella, J. Odell, L. van der Torre, and H. Verhagen, editors, *Proceedings of the 2005 AAAI Fall Symposium ‘Roles, an Interdisciplinary Perspective: Ontologies, Languages, and Multiagent Systems’*. AAAI Press, 2005.

Christopher Mungall, Georgios Gkoutos, Cynthia Smith, Melissa Haendel, Suzanna Lewis, and Michael Ashburner. Integrating phenotype ontologies across multiple species. *Genome Biology*, 11(1):R2+, 2010.

Peter N. Robinson et al. The human phenotype ontology: a tool for annotating and analyzing human hereditary disease. *American Journal of Human Genetics*, 83(5):610–615, 2008.

Cornelius Rosse and Jose L. V. Mejino. A reference ontology for bioinformatics: The foundational model of anatomy. *Journal of Biomedical Informatics*, (36):478–500, 2003.

C. Samargitheat and M. Vihinen. Bioinformatics services related to diagnosis of primary immunodeficiencies. *Current Opinion in Allergy and Clinical Immunology*, 9(6):531–536, 2009.

Richard H. Scheuermann, Werner Ceusters, and Barry Smith. Toward an ontological treatment of disease and diagnosis. In *Proceedings of the 2009 AMIA Summit on Translational Bioinformatics*, pages 116–120, 2009.

J Vliaho, M Pusa, T Ylinen, and M Vihinen. IDR: the immunodeficiency resource. *Nucleic Acids Research*, 30(1): 232–234, 2002.

A classification of existing phenotypical representations and methods for improvement

Anika Oellrich¹, Dietrich Rebholz-Schuhmann¹

¹European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, CB10 1SD, UK

ABSTRACT

Phenotypic data resources have to scale with the growth of genomic data to better understand the association between the genotype and the phenotype of a given organism. Large-scale analysis concerning the connection of phenotype and genotype improve our understanding of diseases, their origins and the underlying mechanisms.

Over recent years, the number and scale of phenotypic resources has grown significantly, but no predominant solution for the annotation and analysis of phenotypes has been found, enabling species-specific and cross-species comparisons of phenotypes, the latter being necessary for evolutionary studies and for cross-species analyses of human diseases.

We suggest a classification of the existing phenotype representations to assess their advantages and disadvantages for genotype-phenotype analysis and for cross-species comparisons. Furthermore, we propose how improvements to the existing phenotype resources would support better and more accurate annotations of genomic data and would lead to a solid foundation for the analysis of phenotypes and the use of disease models from model organisms for human diseases.

1 INTRODUCTION

A phenotype is the composition of all observable characteristics of an organism, whereby observable does not necessarily mean measurable. For example, eye colour is one particular phenotypic characteristic which is not measured but instead observed. A phenotype covers molecular, cellular, metabolic, physiological, anatomical, organismal and environmental attributes. The phenome denotes the full range of all phenotypes for a single species not only one individual.

To untangle diseases mechanisms, we need to understand the connection between a particular genotype and a phenotype and the relation between a phenotype and a disease. The ever growing amount of genomic data offers the opportunity for an exhaustive analysis of both, the phenomic and genomic data in conjunction, but requires to deliver phenotypic description at the same scale and quality as the genomic data (Freimer and Sabatti 2003).

In the past decade, a number of projects have targeted to describe the complete phenome of selected model organisms, e.g. the EUCOMM project in mouse (Friedel et al. 2007). It is not only important to produce descriptions that are consistent across a selected organism, furthermore it is also required that the representations are consistent across species. This consistent representation would facilitate cross-species phenotype comparisons for efficient exploitation of the existing data resources. It would also allow for inferences of new knowledge for a given disease from a well-described model organism to humans and human diseases, e.g. mouse models for Rheumatoid Arthritis (Washington et al. 2009).

For the analysis and comparison of phenotypic data, we need comparable and interoperable ontological representations of the phenotypic data. If descriptions of phenotypes are

based on standardised ontological resources with semantic descriptions, then we can compare those phenotypic descriptions automatically, explore similar phenotypes in different organisms and can effectively derive genotype-phenotype relationships from experimental results. The most important and difficult step is the development of a consistent formal representation of phenotypes which is capable of capturing all facets of it within and across species. To date, we are still lacking comprehensive and expressive phenotypic descriptions and, even more, the existing phenotype resources use different standards for their representation with no explicit semantics.

In principle, the phenotype can be described using natural language ('narrative style') as it is common in the scientific literature, but we could also choose a formal and computer-readable representation. allows automatic consistency checking and would even facilitate computer-based reasoning for the prediction of diseases on experimental data and the inference of disease-related knowledge to a given phenotype. In the past, phenotypic description were purely narrative (see scientific literature and OMIM). In recent years, computer-readable solutions have been produced, e.g. the Gene Ontology (GO) or the Mammalian Phenotype Ontology (MP) (Ashburner et al. 2000, Smith et al. 2005), which have been improved for the reuse of existing ontological resources. One proposed solution is the use of Entity-Quality statements (EQ), i.e. phenotypic expressions that make use of existing ontological resources such as the Foundational Model of Anatomy (FMA) and the Phenotype Quality and Trait Ontology (PATO) for efficient descriptions of the phenotypes (Mungall et al. 2010, Gkoutos et al. 2004, Rosse and Mejino 2003).

In this manuscript, we categorize existing resources for phenotype descriptions based on the way how the phenotypic expressions are generated. We focussed on databases containing phenotypical content for several species and how the phenotype is constituted within those databases. In addition to that, we make suggestions for the improvement of the existing phenotype repositories to improve their interoperability and their reuse for comprehensive phenotypic analyses.

2 FROM NARRATIVE TO FORMAL DESCRIPTIONS OF PHENOTYPES

In principle, we find phenotype descriptions in the scientific literature, in databases for model organisms and annotated molecular data, such as sequences and expression profiles. Until now, no standard format or standard representation has been found suitable to be used universally across the different data resources for the representation of phenotypes. From the different projects concerned with the storage and analysis of phenotypes for each of the model organisms, different types of representations have evolved that are used for the phenotype descriptions. In principle, the descriptions

of the phenotypic resources can be categorized according to the following classification of the phenotype representations:

1. Narrative descriptions of phenotypes

In this category, the author of the phenotypic description does not have to follow any requirements for the formal representation of the phenotype. The description uses natural language and the author only abides to the syntax of natural language. The consistency of the representation is achieved through peer review from a domain expert.

2. Phenotypic representations using a taxonomic structure (ontology)

In this representation, the phenotype consists of a list of phenotypic characteristics usually stored as annotations for an individual, or a group of individuals sharing phenotypic features, of a certain species.

a. The annotations are represented with pre-composed phenotypic attributes

The annotation representing one phenotypic characteristic is based on a species-specific ontology, whereby the annotation corresponds to one concept of this ontology.

b. The annotations are represented with post-composed phenotypic attributes

The annotation, again corresponding to one phenotypical attribute, is a combination of existing ontological resources, e.g. FMA and PATO for human data. The difference to pre-composed is that the annotation does not correspond any more to a single concept of an ontology, instead it is a combination of different concepts from several ontologies. To date, the commonly used approach to define post-composed phenotypic expressions is the entity quality (EQ) approach, whereby an entity, such as an anatomical part or a process, is further specified with a quality.

The pre- and post-composed phenotypic attributes are comparable to pre- and post-coordination, which are terms coined and used in the medical domain (Elkin et al. 1998).

In principle, we would expect that formal representations of phenotypes are best suited to support automatic exploitation of the ontological resource and the annotated phenotypes, but the development of formal representations requires better support to the curators through special computer based support tools to check the consistency of the annotations during their creation and would also enable reasoning based on the phenotypic descriptions. Currently, neither the formalism nor the required tools are available to facilitate this.

The different kinds of phenotypic descriptions have advantages and disadvantages that are summarized in table 1.

2.1 Narrative phenotype description

Narrative descriptions of phenotypes, such as those in the Online Mendelian Inheritance in Man (OMIM) database (McKusick-Nathans Institute of Genetic Medicine and NCBI 2010) or FlyBase (Flybase 1999), deliver a fine-grained description of the organism's phenotype. They provide usually a rich set of details that have to be identified with automatic text processing means if it is integrated into an information technology infrastructure. The creator of the narrative phenotypic description is not bound to formal requirements in the description and therefore can choose his preferred wording

leading to a description that represents the domain knowledge of the author. The format of the phenotypic description varies across the whole database content and hence leads to difficulties to do automatic comparisons of the reported phenotypes, not only within but also across other data resources for both species-specific and cross-species comparisons. Furthermore, searches within those phenotypic content and the integration of the content with other data resources is difficult to achieve.

2.2 Pre- and post-composed ontological resources for phenotype descriptions

Pre- and post-composed phenotypic descriptions can be distinguished based on the availability of the phenotypic attribute at the time the curator uses the attribute for the annotation step. In the case of a pre-composed label, a concept, capturing the phenotypic attribute in its name, has to exist in an ontological resource for the curation step. For example, the curator can use concepts from the Human Phenotype Ontology (HPO) (Robinson et al. 2008) or Mammalian Phenotype Ontology (MP) during the curation work. In the case of post-composed descriptors, the annotator has to generate the phenotypic attribute at the time of the annotation step, i.e. the curator has to compose the phenotypic attribute from existing resources as part of his work. One possibility to create post-composed phenotypic characteristics, and to date the commonly used procedure, are EQ statements.

2.2.1 Pre-composed phenotypic attributes In a pre-composed phenotypic ontology, each concept represents itself a phenotypic attribute. The concepts are ready to use for annotation purposes and do not have to be produced at annotation time. Given that the annotations is limited to the selection of concepts in the pre-existing ontology, we conclude that the expressiveness of the annotations is thus depending on the degree of detail in the pre-existing ontology.

The main advantage in the use of a pre-composed ontology lies in the fact that phenotypes that have been annotated with the same ontology can be compared without difficulties to phenotypes reported in other data resources that have also used the same ontology. The data entries from the two data resources can be aligned through their phenotypic annotations and therefore support for example cross-species comparison, if they deal with two different model organisms.

Furthermore, the pre-composed labels and the taxonomic structure of the ontology can be exploited for computer-based solutions, for example the automatic annotation of existing data. In this case, the computer-based solution has to pre-analyse the labels in the ontology and it has to apply similarity measures to match the labels to the data (Köhler et al. 2009). These solutions enable indexing or categorisation of the phenotypic data resources, the automatic validation of annotations or the prediction of phenotypic traits for heterogeneous data resources, e.g. prediction of phenotypic traits from the literature in the context of a mentioned disease.

The comparison against the ontological resource for the phenotype can in addition exploit the structure of the ontology to improve the overall performance, for example the is-a relations in the taxonomic structure. If two different data resources use different pre-composed phenotypic ontologies then the cross-database comparison requires a mapping between the two ontological resources. In principle, such a mapping could be produced by automatic means or through manual efforts.

2.2.2 Post-composed phenotype characteristics For the post-composed annotations, we can in principle combine all existing ontological resources, but meaningful post-composed phenotypic representations can only be expected from a selection of ontological resources, e.g. the combination of concepts from FMA and PATO enables to denote anatomical modifications based on FMA.

Post-composed phenotypes have the advantage that they are flexible and can be adapted according to the needs of the curator. Hence, the curator generates a phenotypic description that is founded in the definitions in the pre-existing ontological resource and still can choose from a large selection of ontological resources to correctly choose a suitable phenotypic description. The flexibility inherited in the post-compositional approach makes it superior to the alternative of a pre-composed resource delivering 'ready-made' phenotypic descriptors. However, post-composed descriptions are often more difficult to use and employ, and therefore not as accepted as pre-composed descriptions.

3 REQUIREMENTS FOR THE IMPROVEMENT OF PHENOTYPIC REPRESENTATIONS

None of the existing data resources for the annotation of phenotypic information provides a complete and consistent definition of what a phenotype is. Such a definition, would be a prerequisite for a shared understanding about what constitutes a phenotype and which information is required for the description of a phenotype. An ideal solution would even enable us to bridge biological and clinical data and would serve as a shared base for ontological annotations in both domains, i.e. the phenotypic descriptions of patients could be linked to phenotypic dysfunctions in model organisms and to molecular defects in all model organisms.

Furthermore, a formal representation of phenotypes is required to achieve a coherent and comprehensive description across different species-specific phenotype ontologies. The formal representation would not only allow for the interoperability of the existing ontological resources but also improve the quality of annotations and facilitate exhaustive analyses of the existing phenotypic resources by means of consistency checking and knowledge inference (Hoehndorf et al. 2007).

To improve the useability of the free-text descriptions and to provide fast and automated access to it, annotations containing pre- or post-composed annotations are required which conform to the content of the text. The annotations would add all ontological benefits to the existing free-text resources without having to rebuild a new resource. One effort in this direction is the annotation of OMIM with HPO terms (Oti et al. 2009) and other resources would have to follow this example.

In addition to the annotations of free-text phenotypes, the phenotype ontologies need to be adapted to the formal representation once defined and aligned to each other to facilitate the comparison of phenotypes annotated with different ontological resources. The ontological mapping would not only facilitate the species-specific but also the analysis of data across species.

4 CONCLUSION

Various phenotypic resources exist and build a tremendous repository for further analyses. In order to facilitate the exhaustive

analysis of phenotypes, a consistent formal framework has to be developed that defines all aspects of a phenotype and can be applied to the existing resources and their annotations and allows for fine-grained, detail-rich, easy-to-use annotation but also supports the usage of reasoning techniques for consistency checks and knowledge inference.

5 ACKNOWLEDGEMENTS

The authors thank Robert Hoehndorf for valuable discussion about phenotypes, their formal representation and his contributions to this manuscript.

REFERENCES

- Michael Ashburner et al. Gene ontology: tool for the unification of biology. the gene ontology consortium. *Nat Genet*, 25(1):25–9, May 2000.
- Peter L Elkin et al. The role of compositionality in standardized problem list generation. *Medinfo Proceedings*, Jan 1998.
- Consortium Flybase. The flybase database of the drosophila genome projects and community literature. the flybase consortium. *Nucleic Acids Res*, 27(1):85–88, January 1999.
- Nelson Freimer and Chiara Sabatti. The human phenome project. *Nature Genetics*, 34 (1):15–21, 2003.
- Roland H Friedel et al. Eucomm the european conditional mouse mutagenesis program. *Briefings in Functional Genomics*, Jan 2007.
- George Gkoutos et al. Building mouse phenotype ontologies. *Pac Symp Biocomput*, pages 178–189, 2004.
- Robert Hoehndorf et al. Representing default knowledge in biomedical ontologies: application to the integration of anatomy and phenotype ontologies. *BMC Bioinformatics*, 2007.
- Sebastian Köhler et al. Clinical diagnostics in human genetics with semantic similarity searches in ontologies. *Am J Hum Genet*, 85 (4):457–64, Sep 2009.
- MD) McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins University (Baltimore and MD) NCBI, NLM (Bethesda. Online Mendelian Inheritance in Man, OMIM (TM). <http://www.ncbi.nlm.nih.gov/omim/>, 2010.
- Christopher Mungall et al. Integrating phenotype ontologies across multiple species. *Genome Biology*, 11(1):R2+, 2010.
- Martin Oti et al. The biological coherence of human phenome databases. *Am J Hum Genet*, 85(6):801–8, Nov 2009.
- Peter Robinson et al. The human phenotype ontology: a tool for annotating and analyzing human hereditary disease. *American Journal of Human Genetics*, 83(5):610–615, November 2008.
- Cornelius Rosse and José L. V. Mejino. A reference ontology for biomedical informatics: the Foundational Model of Anatomy. *Journal of Biomedical Informatics*, 36(6), 2003.
- Cynthia L Smith et al. The mammalian phenotype ontology as a tool for annotating, analyzing and comparing phenotypic information. *Genome Biol*, 6(1):R7, Dec 2005.
- Nicole L Washington et al. Linking human diseases to animal models using ontology-based phenotype annotation. *PLoS Biol*, 7 (11):e1000247, Nov 2009.

Resources	Advantages	Disadvantages	Example	Terminology
free text	<ul style="list-style-type: none"> - detailed descriptions of phenotypes - increased human readability 	<ul style="list-style-type: none"> - no standardization of the free text descriptions - contained information difficult to access in an automated fashion - searches on phenotypic content are time consuming - no formal representation of phenotypes - neither species-specific nor cross-species comparison possible 	OMIM entry 214200	OMIM, FlyBase
pre-composed	<ul style="list-style-type: none"> - developed according to the domain vocabulary - easy to use for manual annotation - enables searches for specific phenotypic labels - application of tools specific for ontologies allows analysis of data - specific pattern for the definition of phenotypes 	<ul style="list-style-type: none"> - limitation to terms defined within ontology which may be not precise enough - no cross-species comparison possible given that the ontologies are specific to certain species 	HP:0001410 'decreased liver function'	MP, HPO, LDDDB
post-composed	<ul style="list-style-type: none"> - allows for species-specific and cross-species comparisons of phenotypes - enables search functionality on phenotypic labels - variability in patterns for phenotypes as post-composed 	<ul style="list-style-type: none"> - not as easy to use for annotation purposes as no fixed patterns are defined for the creation of a phenotypic characteristic - compatibility of resources required to build phenotypes - limited formal representation through crossproducts and patterns 	entity = FMA:7197 ! liver, quality = PATO:0001624 ! decreased functionality	EQ statements

Table 1. Classification of the existing phenotypic resources into the following three groups: (1) free-text phenotypic descriptions, (2) pre- and (3) post-composed phenotypic characteristics.

Change	Reasons	Resources
Common terminology for phenotypes	<ul style="list-style-type: none"> - reduce discrepancies in understanding - facilitate communication across domains 	needs to be created
Formal representation of phenotypes	<ul style="list-style-type: none"> - coherent representation of phenotypes to facilitate species-specific and cross-species comparisons - opportunity to use reasoning for analysis of phenotypic content 	needs to be created
Adapt ontologies to formal representation	<ul style="list-style-type: none"> - allow species-specific and cross-species comparisons - reuse of existing annotations for inference mechanisms 	all existing phenotype ontologies, such as HPO and MP
Mapping between ontologies	<ul style="list-style-type: none"> - facilitate comparison and interoperability across species - ensure orthogonality of existing resources 	all existing phenotype ontologies, such as HPO and MP
Annotations to free text	<ul style="list-style-type: none"> - faster access phenotypic data - perform efficient searches - achieve formal representation with the annotations - use benefits which result from formal representation 	OMIM, FlyBase

Table 2. Listing of the required changes in the existing phenotype resources.

Towards a Cellular Genealogy Ontology

Patryk Burek^{‡*} & Heinrich Herre^{‡*}, Ingo Roeder[†], Ingmar Glauche[†], Nico Scherf[†],
Markus Löffler[‡]

[‡] Institute for Medical Informatics, Statistics and Epidemiology (IMISE), University of Leipzig, Germany

[†] Institute for Medical Informatics and Biometry, Dresden University of Technology, Germany

ABSTRACT

Motivation: In the present paper we outline basic ideas and results about a *Cellular Genealogy Ontology* (CGO). This work is aimed at providing a framework for analyzing, specifying and annotating results of experiments and of simulations in the field of stem cell research. The real world objects of these investigations are processual cellular genealogies which are studied by using, among other methods, time lapse experiments. This framework pursues three goals: Firstly, it provides a domain independent core ontology, called Simple Process Object Ontology (SPOO). SPOO is the basis for a coherent and integrative handling of objects, processes and characteristics which are the main building blocks for any domain ontology. Secondly, this core ontology is utilized for the development of a domain ontology for cellular genealogies. Thirdly, this domain ontology CGO is intended to support and enrich tracking algorithms by providing annotations of photos, storage and structuring of analysis results, and, thus, enables the discovery of correlations between qualities, visualizations of annotations. Furthermore, CGO is intended to support semantically correct data exchange (import, export).

1 INTRODUCTION

The application of time lapse video microscopy for the analysis of cell cultures facilitates the tracing of single cells, comprising all the progeny over extended time periods up to several days. This includes the temporal analysis of cell specific parameters like morphology, expression of marker genes, cell cycle time, motility or the occurrence of cell death within the population context. All these different information can be comprised into a pedigree-like structure, referred to as *cellular genealogy* (Figure 1), in which the founder cell represents the root and the progeny is arranged in the branches. In such a framework a *cell* is perceived as a *spatially and temporally extended object*. The existence of such a cell is temporally restricted by the generating division of the paternal cell and by the terminating division that generates the descending daughters. Alternatively, a cell might undergo cell death which also precludes its existence.

Automated analysis of time lapse videos from cell cultures allows the simultaneous tracking of a multitude of root cells. Automatic cell tracking procedures are based on the analysis of each individual picture taken during the time lapse experiment. Under a set of rules (characteristics e.g. size, shape, color) certain objects are identified as cell objects in every single picture taken at a particular time point t . This process is termed image segmentation.

2 PROCESSUAL CELLULAR GENEALOGIES

An accurate description and definition of the notion of a cell genealogy, as mentioned in section 1, leads to a number of ontological problems, which are subsequently discussed. It turns out that a cell admits different views which are described by the following observations.

(1) a cell, considered at a time-point t , is completely present at this time-point and has no temporal parts,

(2) a cell participates in a process which exhibits the change of this cell at different time-points.

(3) a cell persists through time, that means that this cell is the same at different time-points.

These conditions are, obviously, incompatible. This situation is, we believe, caused by the fact that the term “cell” denotes three different pairwise disjoint concepts which are closely related. A cell, satisfying the condition (1), is called a *presentic cell*, the corresponding predicate is denoted by $\text{PresCell}(x)$; behind the condition (2) there is a *processual cell*, whose concept is denoted by $\text{ProcCell}(x)$, and, finally, the condition (3) captures a type of a cell which we call *continual cell*, expressed by the predicate $\text{ContCell}(x)$.

The relation between these three concepts is specified by several axioms, introduced in GFO (Herre 2010), which use two ternary relations. If p is a process then $\text{timerest}(p,t,q)$ has the following meaning: the restriction of the process p to the time-point t yields the presentic entity d . Furthermore, the relation $\text{exhibit}(c,t,d)$ expresses the condition: the continual entity c exhibits at time-point t the presentic entity d .

We assume the following integration axioms, formulated in GFO (Herre 2010), which express fundamental interrelations between the categories $\text{Proc}(x)$, $\text{Cont}(x)$, and $\text{Pres}(x)$:

* To whom correspondence should be addressed.

$$\begin{aligned} &\forall x (\text{Cont}(x) \rightarrow \\ &\exists y (\text{Proc}(y) \wedge \forall z t (\text{exhibit}(x,t,z) \rightarrow \text{timerestr}(y,t,z)). \\ &\forall x (\text{Pres}(x) \rightarrow \exists yt (\text{Proc}(y) \wedge \text{timrestr}(y,t,x)) \end{aligned}$$

Note, that these axioms allow to interpret a continual cell as a process. The most basic entity is the processual genealogy of the cell, denoted by ProcGen(c); this is a process which happens and unfolds into branches.. The introduced entities are related to ProcGen(c) and are displayed in Figure 1.

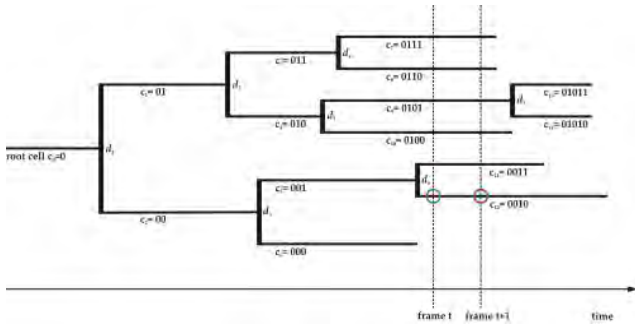


Fig. 1

Within the given five generation genealogy the thin horizontal lines represent the cells c_i whereas the divisions d_i are marked by the thick vertical bars. The horizontal dimension is time with the founding root cell c_0 indicated on the left side. Thus, the length of the horizontal lines represents the duration of the continual cell's existence and is a measure of the cell cycle time. According to the axiom IntAx the horizontal lines present a continual cell which is related to the underlying process, i.e. the processual cell. Final cells on the right side are called leaf cells. The vertical dashed lines indicate the two snapshots of ProcGen(c_0) for the frames t and $t+1$ as indicated in Figure 1. The genealogy illustrates that cell object a in frame t (blue circle) and b in frame $t+1$ (red circle) are presentic cells which participate in the same processual cell c_{12} with the binary code 0010. The representation, as displayed in Fig.1, is called *basic genealogy of the cell* c . These basic genealogies can be annotated by additional properties.

3 SIMPLE PROCESS OBJECT ONTOLOGY (SPOO)

The SPOO exhibits a modification of a part of GFO, in particular, several new concepts are introduced, and some simplifications of GFO are proposed.

3.1 Main Ontological Choices of SPOO

In this section we outline the Simple Process Object Ontology (SPOO), which exhibits a representational level for categories and provides the most general vocabulary and

structuring principles for the Cellular Genealogy Ontology (CGO). Top level of SPOO provides the most general concepts which can be used for domain modeling and, additionally, it permits to localize the concept of a process in the hierarchy of concepts.

The root category of SPOO is *Entity*, and every category of SPOO subsumes under the category *Entity* whose instances include all individuals.. The main distinction between the categories of SPOO is based on their relation to time. The ontology of time is taken in SPOO as primitive and in fact SPOO can be integrated with an arbitrary time ontology as long as it provides the notions of *time point* and *time period* or equivalent. This is a very minimal requirement on time ontology, namely that time can be organized into points and intervals.

We distinguish three categories of entities with respect to their relation to time, namely: (1) *Presentials* - those entities which are located at a time point, (2) *Temporally Extended Entities* (TEE) - entities extended through time, i.e. having some lifetime or happen in time, (3) *Abstracts* - entities, which do not have a relation to time, e.g. mathematical entities. In contradistinction to GFO, SPOO does not make a clear distinction between continuants (having a life-time), and processes (having a temporal extension)

Among abstract entities two are of particular importance, namely, *Property* and *Value*. Properties are abstract entities which we measure, observe or calculate, such as weight, color, speed or temperature. Values are, for example, volumes used in measurement, observation or calculation e.g. 10kg, green, 40\$. Often they are scalars or vectors. Both properties and values are considered as abstracts having no relation to time and independent from the entities which are characterized by them.

An individual assignment of a property and a value to an entity is called *Quality* and since this assignment characterizes an entity it is called its *Characteristic*. A host of a quality can be of an arbitrary ontological kind including a quality itself.

The second class of characteristic is *Role*, which is a notion related to the category of *Relation*, namely a role is an aspect of an entity in context of a relation. For example a cup has in context of the stands_on relation a role of a standing_object, whereas desk has the role of supporting_object.

Relations are defined as entities which glue or connect other entities via roles played by those entities in relations, e.g. stands_on relation glues a cup with a desk. Note, that in SPOO - in accordance with the GFO framework - we do not consider relations as mere sets of n-tuples of the corresponding arguments, but as concrete entities that glue their arguments together.

Relations can be contrasted to *Objects*, which are entities typically perceived and distinguished from other entities as existing to some extent independently of their surroundings. The independence of objects from their surroundings means that they can be conceived and modeled without references

to the surrounding entities. Objects typically have names without references to other entities e.g. apple, house and have characteristics by which they are conceived. This contrast objects to relations and characteristics which in turn often are named by references to their host e.g. process of goods transportation, color of apple.

In SPOO entities which are composed of at least one relation together with its players are called *Situations*. Thus e.g. a cup standing on a table is considered as a situation in which are involved two objects, namely a cup and a table, one relation: *stands_on* and two roles: *standing_object* and *supporting_object*. This situation should not be confused with *stands_on* relation which is dependent on its players but does not have them as its parts.

3.2 Matrix of SPOO entities

The matrix of SPOO entities is constructed by combining the discussed above ontological choices. In each group of entities, i.e. presentials, temporally extended entities and abstracts can be identified objects, qualities, roles, relations and situations. The summary of the SPOO entities is presented in table 1.

Table 1. SPOO Categories

Presential Entities	Time Extended Entities	Abstract Entities
Presentic Object	Object/Continuant	Abstract Object
Presentic Relation	Process	Abstract Relation
Presentic Situation	Situation	Abstract Situation
Presentic Quality	Quality	Abstract Quality
Presentic Role	Role	Abstract Role
		Property
		Value

The SPOO entity which deserves a particular attention is *Process*, which is considered as a time extended relation. Thus, in other words any binding of entities which exist through time is considered in SPOO as a process, e.g. the relation of *stands_on* if considered as existing through time would be a process which can have its own dynamics. This, very general treatment of processes significantly differs SPOO from other top level ontologies e.g. (DOLCE (Masolo, C. et al. 2003)) and enables representing a broad spectrum of processes covering both dynamic process such as movement of a body as well as static processes such as keeping goods in warehouse.

A number of roles are involved in processes, called *Process Participants*, are introduced i.e. *Process Executor*, *Process Resource*, *Process Operand*. A participant of a particular type is a process executor – this is a role of an entity, which is responsible for a process, hosts and realizes a process. It should be mentioned that SPOO is permits not only object but also other entities, including processes themselves, to play a role of executors. This is of particular sig-

nificance in dynamic modeling in engineering and in natural science when processes often hosts and execute other processes.

An executor alone seldom realizes a process, but most likely there are also other entities contributing to the process realization. Such entities are called *Process Contributors*.

Not all participants of a process are responsible for its realization or for an active contribution to it. In SPOO such participants are called *Process Resources*. Process resources are those entities involved in a process, which are not realizing a process or contributing to its realization. A resource is involved in a process intentionally and thus it is involved in the goal of a process as well. Two types of resources can be identified: *Process Products* – resources produced by a process, and *Process Operands* – resources changed or consumed by a process.

4 REQUIREMENTS

In a biological context the cellular genealogies represent unique examples of the developmental sequence originating from the root cell as it occurs under particular assay conditions. The number of research groups doing time lapse experiments and analysis of genealogies increases (Schröder 2008, Eilken 2009, Roeder I 2006, Scherf N 2008, Glauche I 2007), however, the data exchange between them is limited, due to, among other reasons, the lack of a common data format. For the data storage, exchange as well as for the statistical analysis of cellular genealogies a precise characterization of the particular data types is required. The Ontology of Cellular Genealogies (CGO) is developed for that purpose. However, in our opinion the development of the domain ontology alone is not sufficient due to following reasons.

Firstly, the potential users of a genealogy ontology have their own interests and perspectives taken on the domain, hence, there is a high risk of refactoring and restructuring of the ontology during later stages of development or during application. However, refactoring of deployed artifacts is a difficult and expensive enterprise. This fact, well known in software engineering, is also true in the field of ontology development. The recent initiatives of refactoring the biological ontologies, e.g. (Diehl A, et al. 2009) demonstrate that biological ontologies, in order to fulfill their goals, should be well-structured and founded on a solid ground.

Secondly, the development of biological ontologies is a complicated enterprise due to many factors, including the high complexity of the domain and the dynamics of knowledge increase and evolution. This is true also for time lapse experiments. It can be expected that already at an early stage of the ontology's development users will annotate the observed genealogies with more and more information, as well as with new analysis results. Thus, a core requirement for the CGO ontology is to build a model which is easily extensible. A strategy might be the take the processual genealogy of a cell as starting point, and then use als a initial

genealogy the basic genealogy, as introduced in section 2. Any properties, introduced for the the entity ProcGen(c) must be compatible with the basic genealogy BasGen(c).

To illustrate the need of solid conceptual foundations, as well as extensibility, let us consider the most straightforward conceptual model of a cellular genealogy, namely an ontology with one concept Cell and one hasParent relation attached to Cell.

This model is sufficient to represent the cellular genealogy illustrated on figure 1. Moreover, it permits for some extensions e.g. assignment to the cell concept of properties such as e.g. life time or type. However, implementing such a model, although sufficient at the early development stage, would shortly bring us to trouble when it turns out that some of the properties change during the cell life time and, more importantly, some properties do not concern cells, but processes in which they are involved, e.g. cell division or cell death.

5 CELLULAR GENEALOGY ONTOLOGY (CGO)

The Cellular Genealogy Ontology (CGO) is a part of a conceptual framework underlying the annotation schema developed for the purpose of structuring and annotating experiment and simulation results, obtained in frames of the research on the cellular genealogies which is based on the SPOO ontology.

The development of CGO is work in progress, though, in its current state it permits already the description of cellular genealogies and their components in context of research activities which produce them. The main notions of CGO are briefly discussed in the current section.

A *Cellular Genealogy* is considered as a product (called *Cell Tracking Product*) of a *Cell Tracking Process*. Two types of Cell Tracking Processes are considered, namely *Cell Tracking Experiment* and *Cell Tracking Simulation*. The former corresponds to time lapse experiments, the later to the simulations of in silico cell cultures. This distinction permits to handle with the same representation schema both the genealogies simulated and observed and yet distinguish them. Currently, a cell tracking process and cell tracking product are merely placeholders which at the later phases of the ontology development will be extended for proper handling of data describing the social context of research process such as e.g. scientist and lab data. For this purpose integration of CGO with existing schemas such as e.g. Dublin Core is planned (Dublin Core Metadata Initiative 2010).

A Cell Tracking Process results in a sequence of *Frames*. Each Frame is a presential situation describing a cell colony at a particular moment of time. In case of Cell Tracking Experiment a Frame depicts a photo of a colony taken. Each Frame consist of zero to many *Presential Cells*. A presential cell is a presential object being a part of presential situation.

Presential cells can be depicted by qualities such as e.g. *Position* or *Shape*. The SPOO mechanism of characteristics permits a user to define additional qualities of presential cells. A presential cell is characterized not only by presential qualities but also by roles it plays participating in presential relations such as e.g. *Cell-Cell Contact*.

Presential Cells belonging to different Presential Frames can be related by abstract n-ary relation of *Abstract Link* which comes in three types: *Succession*, *Division* and *Fusion*.

Succession is an abstract link between exactly two presential cells located at different Frames indicating that both cells are considered to represent the same time extended cell. In context of succession the older presential cell is called *Predecessor* and the later - *Successor*. Division is a relation which links a single presential cell located at the former frame and called a *Parent Presential Cell* with two cells in the later Frame called *Daughter Presential Cells*. Finally, fusion is a relation gluing two presential cells at a former frame with one at a later frame. All of the above relations come with their own characteristics such as e.g. *probability*, *confidence of human expert*.

Out of presentic cells and their abstract interrelations can be constructed cellular genealogies. A Cellular Genealogy is considered as a SPOO situation, that is a complex time extended entity in which other entities participate. Two types of entities participate in Genealogical trees, namely, *Cells* and *Cellular Processes*.

Cells are represented in CGO as objects, i.e. entities existing in time and to some extent independent of their background and of the processes in which they participate. Cells are constructed out of a chain of presential cells linked with succession relation. Each cell can be characterized with a number of characteristics such as qualities (e.g. *morphology*, *shape*, *lineage assignment*) and roles played by cells in relations. Qualities of cells can be either calculated out of qualities of corresponding presential cells e.g. *velocity* or can be genuine time extended qualities. Cell qualities can be time-parameterized by the temporal location which permits to document a quality value changes overtime.

Among the qualities of cells is *Cell Generation* which organize them in a genealogy. Cell Generation is itself characterized by such qualities as *Division Probabilities* (*Asymmetric*, *Symmetric*, *Undifferentiated Symmetric*) and *Cell Death Index*.

Cells participate in two types of processes: *Cell Division Processes* and *Cell Death Processes*. Cell division process is a process operating on/consuming one cell called *Parent Cell* and producing two *Daughter Cells*. Both the parent cell and the daughter cell are roles of a cell in context of a cell division process. The process of cell division is constructed out of abstract relation of division. It is worth mentioning that in CGO the abstract division relation between presential cells is distinguished from the process of division. The first is mere representation of the fact that instead of one cell at a frame two cells were observed, whereas the second represent a biological process of cell division which can be

further characterized. For example, a division process can be characterized by at most two different division classes with respect to the chosen *view on cell fate identification*. In CGO are introduced two views - *Prospective* and *Retrospective View*. The chosen view has also an influence on the cell lineage assignment, discussed above.

The notion of *Cell Death Process* indicates a process of cell death which operates on one cell.

Root Cell of a cellular genealogy is the cell from which the observation of a cellular development starts and which is a trunk of a genealogy. Technically, the root cell is a role of a cell in context of a genealogy. Out of a root cell new cells are developed by means of the *Cell Division Processes*.

In addition to the discussed above structure of a genealogies and their participants, a number of qualities depict a genealogy, namely *total number of leaf cells*, *total number of divisions*, *range of branch lengths*, *symmetry indices*, *generalized cell death index*.

6 CONCLUSION

In the current paper is discussed a work in progress on a framework for modeling and representing data on cellular genealogies. The framework consist of two ontologies – a top level ontology of SPOO which is a part of GFO tailored for conceptual modeling and the domain ontology of CGO. The former provides a general structuring principles and handling of cross-domain general notions such as object, process and characteristic. Its primary goal is to provide a design patterns for constructing well-structured and easily extensible domain ontologies. The later is a domain ontology providing the vocabulary for describing results of time lapse experiments and simulations.

Currently, the framework is utilized for first applications such as:

- An object-oriented domain model and database schema for cell tracking software developed within the frames of EuroSyStem Project (EuroSyStem Project 2010).
- An export/import format for software tools developed for analysis of time lapse experiments.

The ontology currently is tailored mainly to satisfy the requirements for the purposes of the DynaMo Research Group (DynMo). However, in the second step additional standardization effort and cooperative work with other groups working on cellular genealogies is necessary.

In addition, number of issues concerning the structure of ontologies require further work. In particular these are patterns of abstract links between presentials cells and integration with current biological ontologies e.g. Cell Type Ontology (Bard J, Rhee SY, Ashburner M 2005).

ACKNOWLEDGEMENTS

Research for this paper has been realized within EuroSyStem Project (EuroSyStem Project 2010).

REFERENCES

- Bard J, Rhee SY, Ashburner M (2005). An ontology for cell types. *Genome Biol.* 2005;6:R21.
- Diehl, Alexander, Augustine, Alison, Blake, Judith, Cowell, Lindsay, Gold, Elizabeth, Gondré-Lewis, Timothy, Masci, Anna Maria, Meehan, Terrence, Morel, Penelope, Nijnik, Anastasia, Peters, Bjoern, Pulendran, Bali, Scheuermann, Richard, Yao, Q. Alison, Zand, Martin, and Mungall, Christopher. (2009) Hematopoietic Cell Types: Prototype for a Revised Cell Ontology. Available from Nature Precedings <<http://hdl.handle.net/10101/npre.2009.3635.1>>
- Dublin Core Metadata Initiative (2010). Available <<http://dublincore.org/>>, cited 2010.
- DynaMo
http://tu-dresden.de/die_tu_dresden/fakultaeten/medizinische_fakultaet/inst/imb/forschung/MM_MS
- Eilken, H. M., S.I. Nishikawa, and T. Schroeder. Continuous single-cell imaging of blood generation from haemogenic endothelium. *Nature* 457, 896 (2009)
- EuroSyStem Project (2010). Available <<http://www.eurosystemproject.eu/>>, cited 2010.
- Glauche, I., M. Cross, M. Loeffler, I. Roeder. (2007) Lineage specification of hematopoietic stem cells: Mathematical modeling and biological implications. *Stem Cells* 25, 1791 (2007)
- Glauche, I: *Dissertation* Glauche, I. Theoretical studies on the lineage specification of hematopoietic stem cells, Dissertation, University of Leipzig, 2010
- Herre, H. General Formal Ontology- A Foundational Ontology for Conceptual Modeling. In: *Theory and Application of Ontology*, Ed R. Poli et al. Springer (2010)
- Masolo, C., Borgo, S., Gangemi, A., Guarino, N., Oltramari, A. (2003) WonderWeb Deliverable D18. Ontology Library (final). Version 1.0. Laboratory For Applied Ontology, ISTC-CNR, Trento, Italy.
- Roeder, I. m R. Lorenz. Asymmetry of stem cell fate and the potential impact of the niche observations, simulations, and interpretations. *Stem Cell Rev* 2, 171 (2006)
- Scherf, N., I. Roeder, I. Glauche. in: Proceedings of the Fifth International Workshop on Computational Systems Biology. *WCSB 2008* pp. 161-164 (2008)
- Schroeder, T. Imaging stem-cell-driven regeneration in mammals. *Nature* 453, 345 (2008)

Pre- and Postcoordination in Biomedical Ontologies

Stefan Schulz*, Daniel Schober, Djamila Raufie, Martin Boeker

Institute of Medical Biometry and Medical Informatics, Freiburg University Hospital, Germany

ABSTRACT

To what extent should an ontology provide readily useable pre-coordinated expressions? Pre-coordination is an issue in any major ontology engineering project. Reducing the incidence of pre-coordinated expressions shifts the focus to post-coordination, i.e. the restriction of the ontology to primarily provide primitive representational units, which then leaves the burden of assembling compositional expressions to the user. In this paper, different engineering requirements for ontology pre- and post-coordination are discussed in the context of description logics. On the one hand, pre-coordination can already be supported by equivalence statements drawing on Aristotelian definitions where a relatively low level of shared domain knowledge is required. On the other hand, to be user-friendly, ontologies for post-coordination require a comprehensive axiomatization of the domain using value restrictions and negations. These can be exploited to guide the user to adhere to pre-formulated design patterns and to avoid nonsensical, ambiguous, and idiosyncratic coordinations. The ontology examples are available at: <http://purl.org/biotop/src/pneumonia.zip>

1 INTRODUCTION

Pre-coordination is a general principle governing the design of terminological and ontological resources. According to Frege's Principle [1], the meaning of a complex expression is fully determined by the meanings of its constituents, if appropriate combination rules are taken into account. Hence, for instance, the meaning of the expression "bacterial pneumonia" can be clearly derived from the meaning of its constitutional parts "pneumonia" and "bacteria". Or the meaning of "pancreatitis" is fully derived from the meaning of "pancreas" (and its modified word stem "*pancreat-*") and the meaning of the suffix "*-itis*". However, not all composed linguistic expressions obey Frege's principle: The meaning of "borderline disorder", for instance, cannot be sufficiently derived from the meanings of "borderline" and "disorder". Lexicon, terminology, and ontology curators are continuously confronted with decisions about whether to include a complex term or a fully defined class (pre-coordination), or whether to leave the burden of constructing complex ex-

pressions to the user (post-coordination) [2]. A major reason for providing pre-coordinated constructs is an expected high frequency of use. Hence "femur fracture" is more likely to be found in a medical dictionary or ontology rather than, e.g. "fracture of the middle phalanx of the right index finger". It is a truism that pre-coordinating everything is doomed to failure, as a consequence of combinatorial explosion. Even admitting a certain degree of pre-coordination may multiply the editorial burden, lead to performance problems at reasoning time and, additionally, makes content retrieval difficult. A good example for this is SNOMED CT, a huge clinical terminology with more than 300,000 classes [3], with a high but unevenly distributed incidence of pre-coordination. As an example, SNOMED CT contains over 800 pre-coordinated codes for burns, including, e.g. "Third-degree burn of wrist", "Superficial burn of great toe", "Phosphorus burn of skin", which is however, still far from covering all possible combinations of burns degrees, skin regions, causes and complications. Here, providing just a few atomic building blocks, together with post-coordination guidance would render the artifact smaller and efficient.

Most biomedical terminologies in practical use are still restricted to the provision of pre-coordinated terms, which simplifies their usage but drastically reduces the level of detail available to the users. Exceptions to this rule are found with nomenclatures [4], which are combinatorial vocabularies organized along disjoint axes. Early SNOMED versions [5] and several nursing terminologies follow this principle. A formal foundation of this combinatorial architecture was first put into practice by the GALEN approach [6] and later taken up by SNOMED RT and CT [5]. As soon as an ontology supports both pre-coordination and post-coordination, intelligent systems (supporting fact retrieval, decision support etc.) need to be able to check for equivalence between existing pre-coordinated and user-generated post-coordinated expressions. Only a formal-logic grounding, supported by inference engines such as description logics classifiers can assure the detection of equivalences between pre-coordinated and post-coordinated expressions, which is a crucial task, e.g. for the re-use of coded clinical data in computerized medical records. But still widely used terminologies such as LOINC [7] allows to encode the same content in different ways without providing a formal framework for the detection of equivalences. It permits, e.g., to express "Weight at birth" by a single preexisting code

* To whom correspondence should be addressed: IMBI, Stefan-Meier-Str. 26, D-79104 Freiburg, stschulz@uni-freiburg.de

(LOINC 8339-4), and in parallel, the postcoordination of new expression like:

```
(Weight | LOINC 3141-9, Weight circumstance |
LOINC 8337-8, Birth | SNOMED F-88005).
```

SNOMED CT's compositional grammar [8] supports the post-coordination of complex expressions, e.g. to encode an insertion of a left hip prosthesis:

```
363704007 procedure site |= (24136001 | hip
joint structure |:272741003 | laterality |
=7771000 | left |) {363699004 | direct device |
=304120007 | total hip replacement prosthesis
|, 260686004 | method |=257867005 | insertion-
action |}
```

Translated into description logics and submitted to a classifier, equivalence or subsumption relations between such expressions and existing SNOMED CT classes can then be computed.

Motivated by an ongoing ontology engineering project in the field of infectious diseases and antibiotic therapy [9] we will analyze ontology design requirements regarding their use in a pre-coordinative vs. a post-coordinative context. We acknowledge that modern, ontology-based biomedical terminology systems are neither fully post-coordinative nor fully pre-coordinative, due to very pragmatic reasons.

Our hypothesis is that these two aspects – pre-coordination vs. post-coordination – pose distinct design principles and demands on shaping biomedical ontologies.

2 METHODS

In the following we will present a prototypical sketch of an ontology which provides representational units for the assembly of post-coordinated expressions on the one hand, but which already contains pre-coordinated units that cover the most common cases. We use description logics (DL) [10] as being the standard representational formalism in most biomedical ontology projects like SNOMED CT [3], NCIT [11], and OBO Foundry [12]. Out of the different description logics dialects described in the W3C OWL2 specification we will base our deliberations on the OWL2 RL profile [13], which is supported by current classifiers. In particular, the constructors used are class inclusion (\sqsubseteq), class equivalence (\equiv), class disjointness using negation (\neg), existential quantification (\exists), value restriction (\forall), conjunction (\sqcap), and disjunction (\sqcup). Our running example will be pneumonia, a disease characterized by the inflammation of lung tissue. A broad range of different types of pneumonia can be distinguished according to the following criteria:

1. Pathological/anatomical: the localization of the disease in the lung and its extension to certain tissues;
2. Course: acute or chronic;
3. Etiological: causes of the disease (infections, physical, chemical...);
4. Pre-existing conditions, which are complicated by

pneumonia;

5. The environment in which an infections pneumonia was acquired (community or hospital);
6. Signs and symptoms.

2.1 Pre-coordination requirements

The focus of pre-coordination is the provision of axioms that state equivalences between atomic classes and compositional expressions. The meaning of a class is therefore fully defined in terms of a complex expression out of other classes. In the following examples, examples, largely matching the above pneumonia classificatory criteria several full definitions are given:

Pneumonia \equiv

Inflammation $\sqcap \exists$ **has-participant**.*LungTissue*

AcutePneumonia \equiv

Pneumonia $\sqcap \exists$ **bearer-of**.*AcutenessQuality*

BacterialPneumonia \equiv

Pneumonia $\sqcap \exists$ **has-agent**.*BacteriaPopulation*

ViralPneumonia \equiv

Pneumonia $\sqcap \exists$ **has-agent**.*VirusPopulation*

As the examples demonstrate, pre-coordinated expressions require, besides the equivalence operator (\equiv), existential restriction (\exists) and conjunction (\sqcap), all of them belonging to the less expressive but computationally more efficient OWL EL specification. These constructors are already sufficient to compute the equivalence of post-coordinated classes such as between *ViralPneumonia* with *AcutenessQuality* and *AcutePneumonia* caused by *BacteriaPopulation*. However, if exclusively relying on axioms of this kind important ontological aspects are missing like:

- delineation: Our example would perfectly fit in a world in which the classes *LungTissue* and *KidneyTissue* overlap and therefore a pneumonia located in a kidney constitutes a valid model.
- Top level categories, e.g. whether *Pneumonia* is a process, a pathological structure, or both.
- Allowed values, e.g. whether the role (object property) **has-agent** in *BacterialPneumonia* can, additionally, be filled by instances of, e.g. *VirusPopulation* or *ArthropodPopulation*.
- Cardinalities of role fillers (object properties), specifying, e.g., whether an instance of *AcutePneumonia* can have, besides acute, other qualities, e.g. chronic.

The use of an ontology limited to pre-coordination axioms would therefore be restricted to use cases like retrieving information encoded by post-coordinated expressions by pre-coordinated queries (or vice-versa). For any further-reaching knowledge services, pre-coordination axioms must be embedded into an ontology that additionally supports:

- Taxonomic hierarchies structured by subclass relations, in order to allow inferences such as that every *BacterialPneumonia* is a *BacterialInflammation*;

- Mereotopological axioms, in order to allow inferences such as that every *Pneumonia* is located in some *Lung*, because every instance of *LungTissue* is part of some *Lung*;
- Disjointness statements, e.g. that processes cannot be structures, and therefore a pneumonia process is not the same as the underlying pathological structure, or that a population of viruses cannot be a bacteria population
- A strict categorization by an upper ontology.

2.2 Post-coordination requirements

The above picture changes if an ontology is primarily devised to support postcoordination to be performed by users. Post-coordinated expressions are expected to be

- valid, i.e. they constitute meaningful compositions, with nonsensical coordinations, e.g. “left pancreatitis”, or “viral pneumonia caused by bacteria” to be precluded;
- expressive, i.e. unambiguous;
- reliable, i.e. consistent between different modelers.

This poses new demands for OWL ontologies, which must be addressed by elaborated design patterns. In the following we will use the pneumonia example to illustrate such patterns. The pneumonia ontology we are developing is rooted in the BioTop upper ontology [14] and can be downloaded from <http://purl.org/biotop/src/pneumonia.zip>.

In BioTop, generalized localization is expressed by the relation **locus-of** (inverse **has-locus**), which relates a place with an entity which occurs in it, inheres in it, or is part of it. We can therefore restrict the localization of pneumonias by the following axiom:

$$Pneumonia \sqsubseteq \forall \text{has-locus}.(\exists \text{locus-of}.LungTissue)$$

Given an underlying anatomy model which, e.g. includes the axioms that only lungs have lung tissue, lungs are located in a thorax, and what is located in the lung can not be located in the abdomen or the extremities,

$$\begin{aligned} LungTissue &\sqsubseteq \exists \text{has-locus}.Lung \\ Lung &\sqsubseteq \exists \text{has-locus}.Thorax \\ \exists \text{has-locus}.Thorax &\sqsubseteq \neg \exists \text{has-locus}. \\ &\quad (Abdomen \sqcup Extremity) \end{aligned}$$

an application (e.g. which controls a dropdown menu in an ontology driven, constraint based GUI) can display exactly those terms which indicate sensible locations for pneumonia, whereas it hides all other anatomy terms. In a similar vein, the cause of bacterial pneumonia can be restricted by

$$BacterialPneumonia \sqsubseteq \forall \text{has-agent}.BacteriaPopulation$$

assuming that the class *BacteriaPopulation* is disjoint from, e.g. *VirusPopulation*, and others at the same level. In an ontology-enabled interface, the user would then only be offered subclasses of *BacteriaPopulation*, such as *MRSA-Population* or *PneumococcusPopulation*.

There are more complex patterns such as guiding the representation of secondary pneumonias, i.e. pneumonias, which exist as complications of previous conditions. This requires a more sophisticated disease model, according to the one proposed in [15] based on [16]. Using this model we define a secondary disease as a pathological process that realizes a pre-existing disposition. This disposition inheres in a pathological structure, which may exist as a congenital disorder or as some outcome of a former pathological process. One example is a lung infarction as a cause of pneumonia, others are lung edema and bronchial carcinoma.

$$Pneumonia \sqsubseteq$$

$$\forall \text{realization-of}.(PathologicalDisposition \sqcap$$

$$\forall \text{inheres-in}.$$

$$(LungInfarction \sqcup LungEdema \sqcup \\ BronchialCarcinoma \sqcup \dots))$$

A variation of this pattern is the one which refines a pneumonia process in terms of accompanying signs and symptoms, such as fever, chills, or cough. Again, the user would welcome a concise list of those signs and symptoms that are commonly attributed to the underlying disease, whereas numerous other conditions, which can, of course co-occur with a pneumonia (e.g. toothache, hernia,...) should be hidden. In these cases we consider the pneumonia a primary process, which produces a pathological structure which then bears the disposition for the signs and symptoms (the secondary processes), e.g.

$$PneumoniaProcess \sqsubseteq \exists \text{has-output}.$$

$$(PathologicalStructure \sqcap$$

$$\exists \text{bearer-of}.(PathologicalDisposition \sqcap$$

$$\forall \text{has-realization}.$$

$$(Cough \sqcup Chills \sqcup Fever \sqcup \dots))$$

A more intricate pattern is necessary when addressing the environment in which the pathogens have been acquired. Clinically, this distinction is essential, because the risk of being infected by multiple resistant bacteria is much higher inside health care institutions. Here we need to introduce a second process, *LungTissueExposureToPathogen*, the onset of which is prior to the pneumonia itself (incubation period). For the definition of hospital-acquired pneumonia, the criteria is that the organism in which this process takes place is located within a health care institution when this process initiates.

3 DISCUSSION AND CONCLUSION

Comparing the requirements for pre-coordination centered ontologies (i) vs. post-coordination centered ontologies (ii) we observe that post-coordination centered ontologies require more sophisticated design patterns in the form of axi-

omatic templates if the ontology user is to be guided to construct meaningful expressions. A further analysis draws attention to the fact that the patterns needed for ontology guided post-coordination use different constructors as required for the main reasoning cases in pre-coordinated ontologies. Whereas pre-coordination centered ontologies can perfectly be supported by inexpressive description logics, such as OWL-2 EL (which is the case, e.g., with SNOMED CT) which is sufficient to express Aristotelian definitions, postcoordination requires careful usage of a broader range of constructors, *viz.* negation, disjunction and value restriction. The more guidance the ontology offers to the modelers in order to prevent them to deviate from the path of the pre-formulated design patterns, the more accurately these constraints need to be crafted. Attention must be paid to the fact that description logics based ontologies adhere to the open world assumption. In contrast to frame-based systems, e.g. CLIPS [17] all constraints on values need to be supported by a carefully devised architecture of disjoint categories: if there is nowhere stated that a bacteria cannot be a virus, a value constraint for bacterial pneumonia is of no use. This poses new challenges for intelligent user interfaces supported by the enforcement of constraints provided by the ontology. While procedures which leverage logically defined classes, their OWL representation and the identification of patterns in candidate terms for ontology inclusion are currently being investigated (e.g. [http://obi-ontology.org/page/Quick Term Templates](http://obi-ontology.org/page/Quick_Term_Templates)), principled ontological upper level constraints as found in upper and top level ontologies already provide valuable patterns. In all these areas better support by ontology editing tools would be welcome, i.e. in order to increase inter-annotator reliability, support for compositions that are consistent between different modelers should be provided.

Acknowledgements

This work was supported by the DFG, grant agreement JA 1904/2-1, SCHU 2515/1-1, and by the EU 7th FP project DebugIT, grant agreement ICT-2007.5.2-217139.

4 REFERENCES

1. Zoltan Gendler Szabo. Compositionality. In Edward N. Zalta, editor, *Stanford Encyclopedia of Philosophy*. Stanford: The Metaphysics Research Lab, plato.stanford.edu, 2003.
2. P.L. Elkin, M. Tuttle, K. Keck, K. Campbell, G. Atkin, and C.G. Chute. The role of compositionality in standardized problem list generation. In *Stud. in Health Technology and Informatics*, volume 52, 660–644, 1998.
3. SNOMED Clinical Terms. Copenhagen, Denmark: International Health Terminology Standards Development Organisation (IHTSDO), 2010.
4. Josef Ingenerf and Wolfgang Giere. Concept-oriented standardization and statistics-oriented classification: Continuing the classification versus nomenclature controversy. *Methods of Information in Medicine*, 37(4/5):527–539, 1998.
5. Ronald Cornet and Nicolette de Keizer. Forty years of snomed: a literature review. *BMC Medical Informatics and Medical Decision Making*, 8:Suppl 1:S2, 2008.
6. Alan L. Rector, A. J. Glowinski, W. Anthony Nowlan, and Angelo Rossi Mori. Medical-concept models and medical records: An approach based on Galen and Pen & Pad. *Journal of the American Medical Informatics Association*, 2(1):19–35, 1995.
7. S.M. Huff, R.A. Rocha, C.J. McDonald, G.J. DeMoor, J.E. DeMoor, and T. Fiers. Development of the logical observation identifier names and codes (LOINC) vocabulary. *Journal of the American Medical Informatics Association*, 5:276–292, 1998.
8. Kent Spackman and John Gutai. Compositional grammar for SNOMED CT expressions in HL7 Version 3. http://www.ihtsdo.org/fileadmin/user_upload/Docs_01/About_IHTSDO/Publications/CompositionalGrammar20081223.pdf, 2008.
9. Christian Lovis, Douglas Teodoro, Emilie Pasche, Patrick Ruch, Dirk Colaert, and Karl Stroetmann. DebugIT: building a european distributed clinical data mining network to foster the fight against microbial diseases. Number 148 in *Studies in Health Technology and Informatics*, 50–59, 2009.
10. Franz Baader, Diego Calvanese, Deborah L. McGuinness, Daniele Nardi, and Peter F. Patel-Schneider, editors. *The Description Logic Handbook. Theory, Implementation, and Applications*. 2nd edition. Cambridge, U.K.: Cambridge University Press, 2007.
11. F.W. Hartel, S. De Coronado, R. Dionne, G. Frago, and Golbeck J. Modeling a description logic vocabulary for cancer research. *Journal of Biomedical Informatics*, 38(2):114–129, 2005.
12. B. Smith, M. Ashburner, C. Rosse, J. Bard, W. Bug, W. Ceusters, L.J. Goldberg, K. Eilbeck, A. Ireland, C.J. Mungall, OBI Consortium, N. Leontis, P. Rocca-Serra, A. Ruttenberg, S.A. Sansone, R.H. Scheuermann, N. Shah, P.L. Whetzel, and S. Lewis. The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration. *Nature Biotechnology*, 25(11):1251–1255, 2007.
13. Boris Motik, Bernardo Cuenca Grau, Ian Horrocks, Zhe Wu, Achille Fokoue, and Carsten Lutz. Working Group OWL 2 Web Ontology Language document overview. available at: <http://www.w3.org/STR/owl2-overview>, May 2010.
14. Elena Beisswanger, Stefan Schulz, Holger Stenzhorn, and Udo Hahn. BioTop: An upper domain ontology for the life sciences – a description of its current structure, contents, and interfaces to OBO ontologies. *Applied Ontology*, 3(4):202–212, 2008.
15. Stefan Schulz, Djamilia Raufie, and Martin Boeker. Scalable representations of diseases in biomedical ontologies. In Nigam Shah and Susanna Sansone, editors, *Proc. of the 2010 Bio-Ontologies SIG. Semantic Applications in Life Sciences*. Boston, MA, USA, July 9–10, 2010.
16. Richard H. Scheuermann, Werner Ceusters, and Barry Smith. Towards an ontological treatment of disease and diagnosis. In *Proc. of the 2009 AMIA Summit on Translational Bioinformatics*, 116–120. San Francisco, California, March 15–17, 2009.
17. Daniel Schober, Ulf Leser, Martin Zenke, and Jens Reich. GandrKB – ontological microarray annotation and visualization. *Bioinformatics*, 21(11):2785–2786, 2005.

Ontology Simplification: new buzzword or real need?

Daniel Schober^{1*}, Martin Boeker¹

¹Institute of Medical Biometry and Medical Informatics (IMBI), University Medical Center, 79104 Freiburg, Germany

ABSTRACT

Motivation: Available formal ontologies are not as abundantly used as they could be. As the main reason for this is the inherent complexity of description logics (DL) expressions in owl, we argue for the introduction, collection and propagation of simplification methods that render owl-DL ontologies more human understandable without sacrificing DL reasoning capabilities.

We review existing approaches to simplify re-occurring patterns from the kick-off to the deployment phase and investigate where the application of simplification methods can ease aspects in ontology engineering, re-use and deployment by end-users that are not DL experts.

We present characteristics of classes that render them more understandable and introduce an initial typology of simplification methods. Some new simplification approaches are illustrated in more detail, investigating requirements aligned with different entity types, user roles, tools and deployment settings.

1 INTRODUCTION

On one hand, a wide variety of ontologies relevant to the biomedical domain are available through open access portals such as the NCBO BioPortal [1], and their number is growing rapidly. On the other hand, there is a relatively low number of applications that consume and use these ontologies. The situation has not changed much ever since [2] first mentioned this deficit. This also holds true for artifacts that adhere to emerging ontology engineering good practices, e.g. as established by the OBO Foundry initiative [3] or that implement good practice ontology design patterns (ODPs, [4, 5]).

In general, development progress and usability decreases as the formality and expressive power increases. Especially for OWL-DL ontologies, the learning curve is high on the engineering, as well as on the usage side. As a consequence experienced developers and consumers for formal ontologies are difficult to find, increasing development time and decreasing the actual usage of ontologies in real life deployment settings such as data annotation, integration and querying.

We investigate if simplification ODPs can help to simplify the user-to-ontology interface and owl-DL expressions in

order to increase understandability, user compliance and hence can ultimately lead to a wider acceptance and usage of formal ontologies.

To this extend, we review existing but dispersed, implicit, sparse and difficult to find simplification strategies and add up to them. Some ideas on how to raise general awareness and how to make simplification methods widely accessible are discussed.

2 MATERIALS AND METHODS

We restrict our analysis to the OWL DL flavor¹, as its semantics is formal and rich enough to provide the benefit of logical reasoning. Its simplified **OWL-lite** flavor lacks in sufficient expressivity in this context (no disjoints).

To support our claim for selected simplifications, we will leverage on our involvements in the **DebugIT** project, a large EU project using clinical ontologies to foster international hospital data integration and comparison as needed for clinical guideline development and decision support in the field of antibiotics resistance prevention [6].

We refer to the meta-level entities of which a representational artifact (RA) consists (e.g. classes and object properties) as its representational units (RU, [7]).

An RA is understandable for humans if its RUs are understandable. Ontology **RUs are ‘understandable’** if, they are traceable and readily understandable by the user, i.e. its meaning can be grasped in a short time. This is usually the case when they:

- are instantiated in reality often, i.e. have a high everyday-usage frequency in the domain and can be easily mapped to everyday language constructs, concepts or equivalents in abundantly used terminologies
- reside in the middle level, i.e. are neither too abstract nor too special; residing in the ‘mesolevel’ that can readily and directly be perceived by human senses)
- belong to traceable and intelligible top level categories, i.e. MaterialEntity vs. DependentContinuant
- have relations to other simple classes with logical definitions that are short and built from simple RUs themselves
- are displayed via a short preferred label that is aligned with user expectations

* To whom correspondence should be addressed.

¹ <http://www.w3.org/TR/2004/REC-owl-features-20040210/#s1.3>

Making an RA more understandable is possible via simplifications of two types²

1. Simplifications that remove complexity, and which will prevent full computational exploitation of the formal semantics compared with the un-simplified model.
2. Simplifications that do not remove, but merely hide complexity at certain stages and for certain user roles. These allow full computational exploitation of the formal semantics of the original model

Type 1 simplifications e.g. mere format transformation into OWL-lite, SKOS or RDF triples will only briefly tackled in this paper. Simplifications such as transitioning from

HeparinBiosynthesis \sqsubseteq

(HeparinMetabolism \sqcap (Biosynthesis \sqcap

\exists *acts_on*. Heparin))

to a simple Heparin biosynthesis \sqsubseteq Heparin metabolism will decrease inference capabilities and query granularity, whereas in DL the reverse transitions are encouraged (untangling) to ensure compositionality, parsimony and scalability to enable reasoning. However complexity can be hidden from users at certain times during the ontology life cycle (Type 2 simplification) and that is our focus.

3 RESULTS

Here we present an initial typology for simplification methods and list some new simplification methods as examples. The full set of reviewed existing methods is available online at

<http://www.imbi.uni-freiburg.de/~schober/Simplifications/>

Towards a typology of simplification methods:

1. Syntax simplifications and normalizations
 - a. Exchange complex syntactic expressions with equivalent simpler one.
 - b. Increase readability by removing redundancy (e.g. N3 vs. OWL-RDF)
 - c. Increase readability by aligning to constraint natural language (CNL) grammar
 - d. RU Label normalisations via naming conventions
2. Structural simplification patterns
 - a. Normalization via content ODPs, e.g. for handling roles, functions
 - b. Reification of complex classes
 - i. Granular partition and normalization
 - c. Reification of object properties
 - d. Removing over-specifications
 - e. Remove hidden or implicit content redundancy
3. Shortcuts and local approximative models to ‘fold away’ complexity
 - a. Conflate property chains

- b. Simplified umbrella classes allowing for graceful evolution, i.e. which can later be untangled seamlessly
4. Views showing user and expectations-aligned subsets of RUs
 - a. Graph based visualizations
 - b. Context-based views generated via rules
 - i. Hiding abstract top level nodes
 - ii. Hiding overly specific leaf nodes
 - iii. Hiding unneeded constructs
 1. Prune the relations needed
5. Modularizations and Partitions
 - a. Namespace alignments
 - b. Slim versions, e.g. GO-slim
6. Tools simplification (through adjustable GUI setup)
 - a. Help and guide through data capture
7. Code or GUI enrichment with helper text
 - a. Tool-tips
 - b. Entailment lists to explain reasoning outcome

Existing papers that describe re-useable simplification methods are cited in the supplementary material within the appropriate Simplification category section. Additional new simplification methods are described in more detail below.

Code simplifications through syntax normalization (1.a)

The RDF source-code can be simplified by normalizing semantically equivalent constructs into coherent but simpler forms, e.g. at parsing level individual long descriptions like

```
<rdf:Description rdf:ID=" Beta-3 adrenergic receptor 949352"> <rdf:type
rdf:resource="#AdrenalineReceptor"/> </rdf:Description>
```

can be substituted with a short, but equivalent

```
<AdrenalineReceptor rdf:ID=" Beta-3 adrenergic receptor 949352"/>
```

For Class definitions, we can state that ‘AdrenalineReceptor’ and ‘VoltageGatedReceptor’ are disjoint as follows:

```
<owl:Class rdf:about="#AdrenalineReceptor">
<rdfs:subClassOf><owl:Restriction><owl:complementOf
rdf:resource="#VoltageGatedReceptor"/></owl:Restriction>
</rdfs:subClassOf></owl:Class>
```

This says that every AdrenalineReceptor is an instance of the complement of VoltageGatedReceptor, that is, no AdrenalineReceptor is a VoltageGatedReceptor. This code should be normalized into its short and hence more human readable form:

```
<owl:Class rdf:about="AdrenalineReceptor">
<owl:disjointWith rdf:resource="#VoltageGatedReceptor"/>
</owl:Class>
```

Along these lines DL assertions for equivalence like ‘A SubClass B AND B SubClass A’ should be substituted by the much shorter and traceable form ‘A EquivalentClass B’.

² In analogy to the problem of ‘lossless vs. lossy data compression’, http://en.wikipedia.org/wiki/Data_compression

In all the above examples the expressions are semantically full equivalent, and the complexity reduction is only syntactically achieved by using specialized own idioms and language constructs.

Increasing readability of class expressions via CNL (1.c)

The HeparinBiosynthesis definition from the previous section can be expressed in a syntax that omits logics specific symbolisms and hence becomes easier to read, e.g. using the Manchester OWL Syntax:

```
HeparinBiosynthesis
SubClassOf HeparinMetabolism
SubClassOf Biosynthesis AND acts_on SOME Heparin
```

The expression can be rendered completely understandable by a biologist when it is autoconverted into a constraint natural language e.g. Attempto Controlled English³:

“Every HeparinBiosynthesis is a HeparinMetabolism. Every HeparinBiosynthesis is a Biosynthesis that acts_on a Heparin.”

Simplifying relations (2.c)

In NCIT [8] long relation names can be found, e.g. “Gene_Product_Plays_Role_in_Biological_Process” which themselves constitute nearly full ontological triples, e.g. in Ovary $\sqsubseteq \exists$ *Anatomic_Structure_Is_Physical_Part_Of*. Reproductive_System

The relation name should be kept short, as its length decreases readability and the information is redundant because it already is specified via the domain definition for the relation. Also that an Ovary is an Anatomical_Structure is implicitly stated already. So here a simple ‘Ovary $\sqsubseteq \exists$ *Is_Physical_Part_Of*.Reproductive_System’ would increase readability.

Conflating distributed redundancy in restrictions (1.b)

Redundancy in restriction patterns should be avoided and is also a frequent source of errors for inadequate modeling:

```
Calcium-Activated_Chloride_Channel-2  $\sqsubseteq$ 
 $\exists$  Gene_Product_Expressed_In_Tissue.Lung  $\sqcap$ 
 $\exists$  Gene_Product_Expressed_In_Tissue.Mammary_Gland  $\sqcap$ 
 $\exists$  Gene_Product_Expressed_In_Tissue.Trachea
```

This long axiom contains a large redundant restriction part, which means literally that each individual *Calcium-Activated_Chloride_Channel-2* is simultaneously expressed in three different body parts, which is simply impossible. A simplified, shorter and also logically correct model would be the following

```
Calcium-Activated_Chloride_Channel-2  $\sqsubseteq$ 
 $\forall$  Expressed_In. (Lung  $\sqcup$  Mammary_Gland  $\sqcup$  Trachea)
```

Conflate property chains via Shortcuts (3.a)

Property chains (a new OWL2 feature) allow shortening and simplifying OWL representations by folding and compressing expressions over two or more properties. E.g if A is_son_of B and B is_brother_of C, then these two properties can be chained by a new property: A has_uncle C. In a human view existing shortcuts should be presented rather than the full property chain.

Simplified umbrella classes which can later be untangled seamlessly (3.b)

This method allows for graceful evolution of an ontology through the introduction of local proximity models. Assume one wants to create and maintain a complex domain model for diseases, such as:

```
PathologicalDisposition  $\sqsubseteq$ 
 $\exists$  inheresIn PathologicalStructure
PathologicalProcess  $\sqsubseteq$ 
 $\exists$  hasParticipant PathologicalStructure
PathologicalProcess  $\sqsubseteq$ 
 $\exists$  realizationOf PathologicalDisposition
PathologicalDisposition  $\sqsubseteq$ 
 $\forall$  hasRealization PathologicalProcess
```

Pre-coordinating such statements is labor-intensive because of the sheer amount of all pathological entities to be represented, and hence might not be realistic to implement at an early development stage. For this stage we present a pragmatic proximity model, which at this stage ignores the classic structure / disposition / process distinction, but can later evolve gracefully towards a more complex model. We introduce an intermediate umbrella-class together with all high-level relations to capture the disease / disorder hierarchy, regardless whether a distinction is made between processes, structures, dispositions:

```
PathologicalEntity  $\equiv$  PathologicalStructure  $\sqcup$ 
PathologicalDisposition  $\sqcup$  PathologicalProcess
```

All the needed relations (Pathological Structures: *part-of / located-in*, Pathological Dispositions: *inheres-in*, Pathological Processes: *has-participant / located-in*) can be captured via one super-relation (*has-locus*). This redesign of the relation hierarchy allows the connection to organism parts or locations, without commitment to structure, disposition, or process granularity. This intermediate simple model already supports important inferences, and can later be expanded without rendering the simplification false.

³ <http://attempto.ifi.uzh.ch/aceview/>

4 DISCUSSION

Compared to man-made objects, the biobdomain is more complex, as we are dealing with non-linear behaviors, feedback loops and non-classical physics laws which are all hard to grasp for the human mind [9]. So, simplification should not be realized by removing complexity because statements would become verbose. Type 2 simplifications should be seen as a way to produce more easily understandable views and excerpts of ontologies. As views alone comprise so many techniques that there could be an own paper, the importance of collecting and presenting such methods to the end-user in a coherent way seems justified.

As general non-awareness hints for a need to make simplification methods explicitly available to the wider community, the question arises how and where such methods should best be made publicly accessible. There are a few places where simplification strategies could be published:

1. the OBO Foundry initiative
2. the Ontology Engineering and Patterns Task Force of the Semantic Web Best Practices and Deployment working group [10]
3. ontology design pattern portals such as the NEON Ontology Design Pattern databases [4] or the Manchester pattern portal [5]

Investigating these resources reveals that none currently addresses ‘simplifications’ explicitly. Simplifications are rather seen as properties or qualities of general design patterns, but nevertheless are not explicitly accessible. Here, at least a way to sort patterns according to their simplification abilities seems desirable. For the time being, we will present an initial set of simplification methods on our website, but will also engage in the pattern portal community to get feedback on whether the introduction of an own simplification pattern type is feasible or whether there should merely be an additional descriptor for existing pattern types.

Although ontology understanding fundamentally depends on human cognitive abilities, it can be facilitated via tools implementing simplification strategies.

Issues that need further investigation are

- What complexities can be automatically detected and be removed or shielded from the user by parsers and tools? E.g. parsers can unify and simplify OWL code.
- Would a tool like a guided simplification finder be feasible that chooses appropriate simplifications based on user requirements?

5 CONCLUSION

The high complexity of formal ontologies restricts their own acceptance, which decreases re-use, applicability and overall ‘market penetration’ [11]. Surveying the literature has yielded an initial set of simplification methods which supports the need for a common repository to make the com-

munity aware of simplification patterns and render them widely accessible.

The set of simplification strategies presented here should be viewed as a primer, to be expanded and refined on the basis of further input from the pattern community, allowing continuous refinement, expansion and finalisation of this proposal. Subsets of the surveyed strategies are currently under evaluation in the DebugIT project.

ACKNOWLEDGEMENTS

We kindly acknowledge the members of the ODP portal and of the DebugIT project for their valuable contributions.

REFERENCES

- [1] DL Rubin, SE Lewis, CJ Mungall, S Misra, M Westfield, M Ashburner, I Sim, CG Chute, H Solbrig, MA Storey, et al: **National Center for Biomedical Ontology: advancing biomedicine through structured organization of scientific knowledge.** *Omics* 2006, 10:185-98.
- [2] Rector A. **Medical Informatics. Chapter 13, Description Logic Handbook**, edited by F. Baader, D. Calvanese, D.L. McGuinness, D. Nardi, P.F. Patel-Schneider, Cambridge University Press, 2003, pages 406-426.
- [3] Smith B, Ashburner M, Rosse C, Bard J, Bug W, Ceusters W, Goldberg LJ, Eilbeck K, Ireland A, Mungall CJ, et al: **The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration.** *Nat Biotechnol* 2007, 25:1251-1255.
- [4] V. Presutti and A. Gangemi. **Content ontology design patterns as practical building blocks for web ontologies.** In *Proceedings of ER2008*. Barcelona, Spain, 2008.
- [5] Mikel Egaña, Alan L. Rector, Robert Stevens, Erick Antezana: **Applying Ontology Design Patterns in Bio-ontologies.** EKAW 2008: 7-16
- [6] D Schober, M Boeker, J Bullenkamp, S Schulz **'The DebugIT Core Ontology: semantic integration of antibiotics resistance patterns**, accepted paper at 13th World Congress on Medical and Health Informatics Medinfo 2010, to be held from 12-15th September 2010 in Cape Town, South Africa.
- [7] Smith B, Kusnierczyk W, Schober D, Ceusters W: **Towards a Reference Terminology for Ontology Research and Development in the Biomedical Domain.** *KR-MED* 2006
- [8] de Coronado S, Wright LW, Fragoso G, Haber MW, Hahn-Dantona EA, Hartel FW, Quan SL, Safran T, Thomas N, Whiteman L. **The NCI Thesaurus Quality Assurance Life Cycle.** *Journal of Biomedical Informatics* 2009 Jan 22
- [9] Dean M. Jones, Pepijn R.S. Visser and Ray C. Paton, **Addressing Biological Complexity to Enable Knowledge Sharing**, *AAAI Technical Report WS-98-04*. AAAI (www.aaai.org), AAAI Workshop on Knowledge Sharing Across Biological and Medical Knowledge Based Systems
- [10] **Semantic web best practices and deployment group, Ontology Engineering and Patterns Task Force** [<http://www.w3.org/2001/sw/BestPractices/OEP/>]
- [11] Cimino JJ, Zhu X. **The practical impact of ontologies on biomedical informatics.** *Yearb Med Inform.* 2006:124-35.

A Generic Reification Strategy for n -ary Relations in DL

Niels Grewe

University of Rostock, Rostock, Germany

ABSTRACT

The representation of relations with arity > 2 in description logic formalisms is usually performed by resorting to ‘reification’, i.e. the representation of relations as classes of entities instead of roles. We assess the various shortcomings of reification procedures and suggest a new strategy for performing reification that specifically addresses the issues of ontological adequacy and inter-modeller comparability. This proposal involves the creation of an ontology template or design pattern capable of expressing arbitrary relations.

1 INTRODUCTION

1.1 Expressiveness/Complexity Tradeoffs

When logic based formalisms, such as the various systems of description logic (DL), are sought for representing information about the world, they are usually designed to facilitate machine-reasoning procedures. This has the perceived benefit of allowing the machine to explicate those parts of knowledge about the topic domain that are only implicit in the original data. To that end, DL systems have to meet two crucial requirements:

- *Ontological adequacy*: The system has to accurately represent the information about the topic domain.
- *Pragmatic adequacy*: The computational complexity of the system needs to allow for an effective implementation.

One would thus be compelled to reject systems that either cannot represent crucial parts of the domain ontology or that have algorithms with unfavourable computational properties, like not terminating within reasonable time or space limits (or worse: not terminating at all).

There is, however, a direct link between the expressiveness of the formalism and its complexity properties: While ordinary first-order logic (FOL) is sufficiently expressive to account for a large subset of everyday as well as scientific knowledge, it is not, as Church [4] and Turing [16] have shown, decidable in the general case. Because of the general link between complexity and expressiveness, one main area of research in description logics is concerned with finding appealing tradeoffs between expressiveness and complexity [6].

One particularly fruitful way to reduce complexity is the reduction of the number of variables admissible when constructing sentences of the language in question. It has since been shown that the two-variable fragment of FOL (FO^2) is decidable without [13], and also with equality [10]. This makes FO^2 a useful basis for creating DL formalisms and, in fact, most DL languages restrict themselves to the two-variable subset of FOL.¹ Hence, only binary relations can be expressed natively in common DL systems.

¹ Though some very expressive description logics, such as \mathcal{DLR} [3] include support for n -ary relations, but they also make use of reification internally [2].

1.2 n -ary Relations in the Biomedical Domain

Unfortunately, not all aspects of reality can be represented by classes of entities (universals) and binary roles alone, and quite a few of those cases are relevant in the biomedical domain. It is thus relevant to assess whether the restriction to binary relations is a serious limitation for the use of DL formalisms in the organisation of knowledge in a life-science context or whether the limitation can be circumvented by suitable strategies.

One possible strategy would be to deny that there are actual relations with an arity > 2 , i.e. that all relations with arity > 2 can be analysed into more elementary binary relations. There are (of course) cases where such a reduction is possible. For example, the relation

- (1) Cell C has organelle O with function F .

can easily be made sense of by observing that the following relations suffice to express its content:

- (2) Cell C has organelle O .
- (3) Organelle O has function F .

Here, the O can be regarded as the linking node between the two relations, which are only ‘daisy-chained’ in (1) to aggregate the information expressed by (2) and (3).

Still, it is clearly naïve to assume that such a reduction is always possible. Take for example the relation

- (4) Disorder D has morphology M at site S .

Relations of this form are prominent in the description of disorders and cannot, as Spackman et al. [15] have shown, be analysed by the simple ‘daisy-chaining’ approach used in the decomposition of (1). For, if we were to take D as the linking node between two relations, we would lose the ability to determinately refer to morphology-site pairs associated with a disorder (as is needed in the case of disorders such as the Tetralogy of Fallot, cf. [15]). Also, we have little ontological ground to prefer either M nor S as the linking class, making a decision completely arbitrary.

Another important group of statements that cannot easily be analysed by means of binary relations are dispositional statements [12]. It might, for example, be necessary to assert that a patient’s hypersensitivity to erythromycin is realised through severe nausea when a certain dose of erythromycin is present in the patient [7]. These statements usually take the following form:

- (5) Disposition D has realisation R under condition C .

It is clear that this relation too cannot be analysed into binary components. This is due to the fact that one cannot ascribe the condition to either the disposition or the realisation: The intake of a certain dose of erythromycin is not the only condition under which severe nausea can occur: There are many other things that bring about nausea.

On the other hand, it does not help to say that erythromycin intolerance has as a condition the intake of a certain dose of erythromycin: To make sense of that statement, we have to say that said intake is a condition *for the realisation* of the disposition. Since different doses of erythromycin might have different effects on the patient, or even no effect, we cannot get rid of the ternarity: The conditions for the realisation of a dispositions only apply to the disposition-realisation pair and not to one of its elements.

2 NAÏVE REIFICATION

2.1 Strategy

Since there are interesting cases of relations with greater-than-binary arity, their representation in DL formalisms deserves some scrutiny: One common way of treating n -ary relations is *reification*. In this context, the term refers to the representation of an n -ary relation as a class in the DL system. The most straightforward approach to this is to introduce a class C_R for every n -ary relation R that shall be expressed. This class is then linked to the relata of R by n binary roles. Applied to example (4), this reification strategy would yield a graph similar to the one represented in figure 1. This naïve strategy is also suggested as a design pattern by the ODP Public Catalog [1].

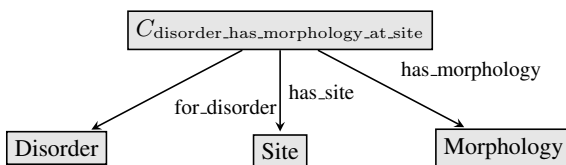


Figure 1. Naïve reification of a ternary relation

2.2 Problems

This approach has a number of well-known problems [2, 14]. On the technical side, the representation as a class makes it impossible to use role constructors such as role composition or transitive closure to state axioms about the relation. It is also difficult to enforce uniqueness-constraints that are usually implicit in the set-theoretical semantics of a DL model: Two instances of a ‘normal’ role assertion are the same instance iff they contain the same elements (i.e. $\langle d, s, m \rangle$ is the same tuple as $\langle d, s, m \rangle$).

This is not the case with reified relations: Two distinct instances of the relation-class can be connected to the same relata without being the same instance. This situation cannot be avoided by explicit axioms. But it can be shown that from every model in which multiple instances of C_R represent the same tuple, one can generate a model where such duplication does not occur [2].

Apart from these technical difficulties, reification also has quite a few implications for a modeller (cf. [14]): It results in an immense multiplication of classes and roles. The above examples of ternary relations would require the inclusion of one additional class and three additional roles per ternary relation. Also, additional axioms might be needed to enforce constraints on the reified structure.

This proliferation of classes and roles can severely impede the manageability of an ontology. This is even more the case when one observes that the introduced entities are purely of technical nature:

They are implementation details that the modeller, usually being a domain expert, not a DL researcher, will not (and should not need to) be familiar with.

It also introduces the problem that there are multiple ways to reify the same relation, especially with growing arity. One modeller might analyse, say, a 6-ary relation into two reified ternary relations, while another might reify it directly without referring to any intermediates. This can make it rather difficult to understand reifications created by a modeller with a different reification-‘style’ and can additionally pose a problem for ontology alignment.

Furthermore, the classes introduced by reification lack clear ontological status: What kind of entity is an instance of $C_{\text{disorder_has_morphology_at_site}}$? If it is merely a modelling artifact, why does it still express some bit of domain knowledge? If it is more than that, where is its proper place in the hierarchy of classes? Such questions need to be answered in order to make sense of reification in ontologies with strict formal requirements.

3 BEST-PRACTICE DRIVEN REIFICATION

3.1 Strategy

In view of these problems, Severi, Fiadeiro and Ekserdjian have, in recent work [14], drawn the conclusion that it is expedient to reduce the number of reifications employed. Taking hints from the concept of aggregation as it is discussed in the relational database community, they suggest a set of best practices designed to reduce the number of reifications needed and to facilitate the creation of ‘better’ reifications in general. These best practices include the following maxims:

1. *Separate ontology and information system*: Many relations can be reduced in their arity simply by relegating descriptive properties from concrete value domains (e.g. the concentration of a substance) into an information system associated with the ontology. This is possible whenever a functional dependency can be established from instances in the DL knowledge-base to the concrete value domain.
2. *Facilitate reuse*: Many n -ary relations that might be candidates for reification actually contain similar sub-relations that may be reused in different contexts. Many relations express, for example, information not about entities *simpliciter* but about entities located at specific sites. In these cases, it is beneficial not to reify the whole relation but to reify the reusable sub-relation for use as a component in further reifications.
3. *Follow domain ontology*: Modellers should be, after all, specialists in the application domain of an ontology. To the specialist, some reifications will suggest themselves as natural with regard to the topic domain and will correspond to *bona-fide* ontological classes in the ontology. One should prefer to use those classes instead of *ad-hoc* reifications.

We can now reformulate our reification of (4) using these maxims. To do that, we turn to maxim 2. and recognise that there is a close connection between a site and a morphology. It thus seems sensible to group them together and reify the relationship between them, so that it can be referred to as a simple property of the disorder, facilitating the reuse of the pair along the way. This line of reasoning corresponds to the ‘role group’-approach that has been adopted by Spackman et al. [15] for SNOMED CT.

Unfortunately, in SNOMED CT, role groups are anonymous classes that are only implicitly asserted in the grouping of properties. We thus turn to maxim 3. and ask whether there is an interpretation that might compel one to accept that the reified relation (the role group) corresponds to a bona-fide entity of the topic domain. This approach has also been taken, e.g. by Cornet and Schulz [5] who interpret the existence of a certain morphology M at site S as a sign for the disorder. We can thus generate a refined reification-graph for the ternary relation in question (Figure 2).

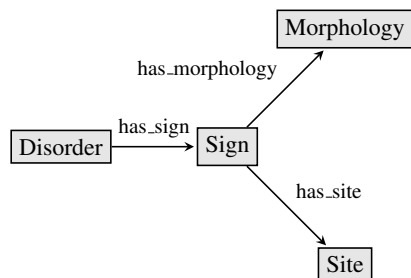


Figure 2. Best-practice driven reification of a ternary relation.

It is clear that the approach of adhering to a set of established, sensible guidelines yields far better results than the naïve approach. It allows for greater flexibility in the use of role and concept constructors. One could, for example, use the *has_sign* role to define the class HypertrophyDisorder by means of the binary roles used to construct the reification:

$$(6) \text{ HypertrophyDisorder} \equiv \text{Disorder} \sqcap \exists \text{has_sign} \circ \text{has_morphology} . \text{Hypertrophy}$$

The same content could still be expressed through the roles present in the original ‘naïve’ reification (by observing that *has_sign* is the inverse of *for_disorder*), the resulting expression is less natural and far more confusing to the modeller.

3.2 Problems

While the approach sketched by [14] is clearly beneficial and should, by all means, guide a modeller’s decisions about when and what to reify, it still leaves two problems:

1. It does not increase inter-modeller consistency. On the contrary: Especially maxim 3. will not assure greater convergence but rather entice greater differences, since the decision will not only depend on the skills and motivation of the modeller but also on his or her ontological intuitions and commitments which can sometimes give rise to heated debate.
2. In some cases, achieving a strictly binary ontological model of the n -ary relation in question might not be possible. (5) might be an example of such a case, if one wishes to advance the thesis that conditions only hold jointly for dispositions and their realisations.

4 TEMPLATE-BASED REIFICATION

4.1 Properties of relations

To address these problems it is expedient take a closer look at the general properties of relations which we need to take into account when talking about their possible representations. First of all, it is important to note that relations are not mere abstractions: Facts about the relationships of entities are facts about the world. I.e. if ‘ a is larger than b ’ holds, then the object denoted by ‘ a ’ is really larger than the one denoted by ‘ b ’. It is true that we might make up certain relations (like ‘ x is larger than y and smaller than z ’), but that does not vitiate the fact that they hold by virtue of the underlying constellation of entities.

This also elucidates the fact the relations always have other entities on which they depend. Without humans that ‘fill in the blanks’ there cannot be a relation like ‘loves’. Also, the number of entities that partake in a relation is fixed. A single instance of ‘loves’ does not connect 1, 2, 3, or n entities but exactly two of them, and in a fixed order at that: It makes a huge difference whether Apollo loves Daphne or Daphne loves Apollo (the former does hold, the latter doesn’t).

4.2 Constructing a template

We will now try to find a way to construct classes that can encode these features in a DL formalism. This carries some inherent difficulties: Since the aim of reification is to work around the limited expressiveness of DL formalisms, it has to respect these limitations as well. In this case, many problems can be avoided by adopting *SRIOQ* [8], on which OWL 2 is based, as the formalism of choice. It provides the $\text{Dis}(R, S)$ (disjointness) role assertion that can be employed to enforce the unique ordering of arguments. One important limitation remains though: *SRIOQ* only allows so called *simple roles* to appear in qualified number restrictions, a problem which will be discussed in due time.

The reification strategy proposed here is in general quite similar to the way n -ary trees are usually encoded as binary trees [9].² We encode the arguments of a relation as a binary tree where one child node connects the argument to the value it takes and the other takes the next argument of the relation (cf. Fig. 3). This encoding can

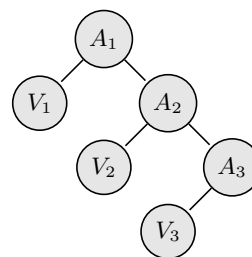


Figure 3. Binary tree encoding of an argument chain (‘ A_n ’ denotes arguments, ‘ V_n ’ denotes values)

be axiomatised as follows using the class *Argument* and the roles *succeeds* and *has_value* for the sibling (argument to argument) and child (argument to value) connections respectively:

² I am indebted to an anonymous reviewer for noticing the similarity.

- (7) $\text{Argument} \equiv \forall \text{succeeds}.\text{Argument} \sqcap \leq 1 \text{succeeds}$
 $\sqcap \leq 1 \text{succeeds}^- \sqcap = 1 \text{has_value}.\top$
(8) $\text{Argument0} \equiv \text{Argument} \sqcap \neg \exists \text{succeeds}.\text{Argument}$
 $\sqcap = 1 \text{has_first_argument}^-.\text{Relation}$
(9) $\text{ArgumentN} \equiv \text{Argument} \sqcap \geq 1 \text{succeeds}$
 $\sqcap \geq 1 \text{succeeds}^-$
(10) $\text{ArgumentLast} \equiv \text{Argument} \sqcap \neg \exists \text{succeeds}^-.\text{Argument}$
(11) $\text{Argument0} \sqcap \text{ArgumentN} \equiv \perp$
(12) $\text{Dis}(\text{succeeds}, \text{succeeds}^-)$
(13) $\text{Relation} \equiv = 1 \text{has_first_argument}.\text{Argument0}$

With axiom (7), we establish that the argument chain is free of branchings and fusions (because every argument has at most one predecessor and successor). Together with axiom (12) the required ordering properties are obtained: By requiring the *succeeds*-role and its inverse to be disjoint, *succeeds* will be asymmetric³ (and, because it is asymmetric, also irreflexive).

Axiom (7) also entails that every instance of *Argument* is connected to exactly one value, which corresponds to the intuition that relational facts are only asserted if some underlying constellation of entities is present.

Additionally, since there is a fixed number of arguments to every relation, it is necessary to ensure that both the beginning and the end of the argument chain can be designated when constructing the reification. This is possible by using the subclasses *Argument0* (8) and *ArgumentLast* (10), which specify that no argument shall succeed or precede the argument in question. Intermediary arguments should be classified as *ArgumentN* (9) to ensure that they are connected to the other arguments in the chain. Finally, (8), together with (13), also mandates that there is a one-to-one correspondence of relations and argument chains.⁴

Unfortunately, the means available in *SR_OI_Q* do not allow the number of arguments for a relation to be encoded directly in some abstract pattern. This would require the use of the transitive closure of *succeeds* in order to 'reach' all arguments of the relation. If that was possible, one could restrict the arity of a relation as follows:

- (14) $\text{TernaryRelation} \equiv$
 $\text{Relation} \sqcap = 3 \text{has_argument}.\text{Argument}$

Unfortunately, this violates the restrictions placed by *SR_OI_Q* on the use of roles in number restrictions, because *has_argument* would need to be defined by role composition and would hence not count as a *simple role*. Instead, the required number of arguments has to be enforced on the individual elements of the argument chain when modelling the reification. A simple algorithm can be followed to perform this task.

Given a n -ary predicate-letter R in a first order language, perform the following steps to generate a reified equivalent $R_R \sqsubseteq \text{Relation}$:

1. Take $i = 1$ to be the first argument position of R .

2. Let A_i be a new class with $A_i \sqsubseteq \text{Argument}$.
3. If i is the first argument position in R
 - a. assert $A_i \sqsubseteq \text{Argument0}$.
4. If i is not the first argument position in R
 - a. assert $A_i \sqsubseteq \text{ArgumentN}$.
 - b. assert $A_i \sqsubseteq = 1 \text{succeeds}.A_{i-1}$.
5. Let V be the class that designates the intended values of the argument at position i . Assert $A_i \sqsubseteq = 1 \text{has_value}.V$.
6. If there is an argument position of R that follows i ,
 - a. Increment i to signify the next argument position of R .
 - b. Continue with step 2.
7. If i is the last argument position in R
 - a. Assert $A_i \sqsubseteq \text{ArgumentLast}$.
8. Assert $R_R \sqsubseteq = 1 \text{has_first_argument}.A_1$.

If this procedure is followed, it is also possible to reuse the *has_value*, *succeeds*, and *has_first_argument* roles because for any given domain the intended range will be clearly specified by the restrictions imposed on the argument and relation classes. This also facilitates the creation of a hierarchy of reified relations.

4.3 Performance

To assess the performance of the reification strategy suggested here, a small test case based on the Ontology for General Medical Science (OGMS, [11]) was prepared by adding 20 subclasses of the classes for disease, disease course and disorder. Each of the classes was populated with 20 instances. The performance of the following five ontologies was compared by measuring the time the Hermit reasoner took to classify them (measurements being made using the Unix utility 'time'):

- BL Baseline profile, 60 classes with 20 individuals each added to OGMS.
- NA-20 The baseline profile, extended by 20 naively reified ternary relation-classes (400 instances asserted).
- R3-10 The baseline profile, extended with 10 reified ternary relation-classes using the template sketched in this paper (200 instances asserted).
- R3-20 The baseline profile, extended with 20 reified ternary relation-classes using the template sketched in this paper (400 instances asserted).
- R4-10 The baseline profile, extended with 10 reified quarternary relation-classes using the template sketched in this paper (400 instances asserted).

Ontology	User Time
BL	13.69s
NA-20	26.59s
R3-10	38.39s
R3-20	74.37s
R4-10	73.67s

Table 1. Performance comparison for different reifying ontologies

³ If $\langle x, y \rangle \in \text{succeeds}$, then $\langle y, x \rangle \in \text{succeeds}^-$. The disjointness axiom (12) thus entails $\langle y, x \rangle \notin \text{succeeds}$ iff $\langle x, y \rangle \in \text{succeeds}$.

⁴ Since any instance of *Argument0* suffices to uniquely identify a reified relational assertion, one could also argue that the class *Relation* is superfluous. It does, though, add some symmetry with our talk about relations: We don't say that a relation *is* its first argument, but rather that it *has* a first argument.

The measurements, as shown in table 1, indicate that the performance impact of going from naïve reification, expressible in computationally more favourable DL-dialects, to the approach suggested here is indeed significant. Considering only the ontologies using the new scheme, classification time seems to increase linearly with the number of reified relation-classes, the impact from

5 DISCUSSION

The template-based reification strategy sketched in this paper does not intend to replace the best practices outlined in section 3.1. But it might be useful as a supplement. Whenever an n -ary relation cannot be made sense of by identifying an ontological counterpart for it, or if that counterpart is disputed, this scheme allows the construction of a reification in a systematic way, so that it can easily be shared with other modellers. It also ensures that the reification is ontologically adequate: All instances of the relation will take the same number of arguments, none of the arguments can be asserted without the others etc.

These benefits come, however, at a price: The proliferation of classes is even greater in this framework than in the ‘naïve’ approach (but a fixed number of relations is sufficient). Also, in the form sketched here, it makes use of features that require the expressivity of *SRIOQ* and results in significantly degraded performance compared to the naïve approach. It should thus be carefully evaluated whether performance needs to be traded for additional modelling consistency.

Still, it shows that a careful consideration of the ontological properties of relations can lead to systematic, controlled strategies in the creation of ontologies, even if, in this case, these strategies are only necessary because of the complexity constraints placed on automated reasoning systems.

ACKNOWLEDGEMENTS

Thanks to Ludger Jansen and Johannes Röhl, Rostock, for fruitful discussions and three anonymous reviewers for their insightful and stimulating comments. This work is supported by the German Science Foundation (DFG) as part of the research project ‘Good Ontology Design’ (GoodOD).

REFERENCES

- [1]Mikel Egaña Aranguren et al. ‘Nary Relationship’. In: *Ontology Design Patterns Public Catalog* (2009) URL: http://www.gong.manchester.ac.uk/odp/html/Nary_Relationship.html (visited on 28/06/2010)
- [2]Diego Calvanese and Giuseppe De Giacomo. ‘Expressive Description Logics’. In: *The Description Logic Handbook. Theory, implementation, and applications*. Ed. by Franz Baader et al. Cambridge: Cambridge University Press, 2003. Chap. 5, 178–218.
- [3]Diego Calvanese, Giuseppe De Giacomo and Maurizio Lenzerini. ‘Conjunctive Query Containment in Description Logics with n -ary Relations’. In: *Proceedings of the 1997 Description Logic Workshop (DL’97)* (1997) 5–9.
- [4]Alonzo Church. ‘An unsolvable problem of elementary number theory’. In: *American Journal of Mathematics* 58 (1936) 345–363.
- [5]Ronald Cornet and Stefan Schulz. ‘Relationship Groups in SNOMED CT’ in: *Medical Informatics in a United and Healthy Europe*. Ed. by K.-P. Adlassnig et al. Amsterdam: IOS Press, 2009, pp. 223–227. DOI: 10.3233/978-1-60750-044-5-223.
- [6]Francesco M. Donini. ‘Complexity of Reasoning’. In: *The Description Logic Handbook. Theory, implementation, and applications*. Ed. by Franz Baader et al. Cambridge: Cambridge University Press, 2003. Chap. 2, 96–136.
- [7]William R. Hogan. ‘Towards an ontological theory of substance intolerance and hypersensitivity’. In: *Journal of Biomedical Informatics* (2010) DOI: 10.1016/j.jbi.2010.02.003.
- [8]Ian Horrocks, Oliver Kutz and Ulrike Sattler. ‘The Even More Irresistible *SRIOQ*’. In: *Proceedings of the 10th Int. Conf. on Principle of Knowledge Representation and Reasoning (KR 2006)* (2006) URL: <http://www.cs.man.ac.uk/~sattler/publications/sroiq-TR.pdf>.
- [9]Donald E. Knuth. ‘Binary Tree Representation of Trees’. In: *The Art of Computer Programming*. Vol. 1: Fundamental Algorithms. 1997, pp. 334–347.
- [10]Michael Mortimer. ‘On languages with two variables’. In: *Zeitschrift für mathematische Logik und Grundlagen der Mathematik* 21 (1975) 135–140.
- [11]Cornelius Rosse et al. *Ontology for General Medical Science*. URL: <http://purl.obolibrary.org/obo/ogms.owl> (visited on 21/08/2010)
- [12]Stefan Schulz and Ludger Jansen. ‘Molecular Interactions. On the Ambiguity of Ordinary Statements in Biomedical Literature’. In: *Applied Ontology* 4 (2009) pp. 21–34. DOI: 10.3233/AO-2009-0061.
- [13]Dana Scott. ‘A decision method for validity of sentences in two variables’. In: *Journal of Symbolic Logic* 27 (1962) p. 477.
- [14]Paula Severi, José Fiadeiro and David Ekserdjian. ‘Guiding Reification in OWL through Aggregation’. In: *Proceedings of the 2010 Description Logic Workshop (DL2010)* (2010) 416–427.
- [15]Kent A. Spackman et al. ‘Role Grouping as an Extension to the Description Logic of Ontolog motivated by Concept Modelling in SNOMED’ in: *Proceedings of the AMIA Symposium* (2002) 712–716.
- [16]Alan Turing. ‘On computable numbers, with an application to the Entscheidungsproblem’. In: *Proceedings of the London Mathematical Society*. 2nd ser. 42 (1937) 230–265.

In der Reihe IMISE-REPORTS sind bisher erschienen:

2002

- | | | |
|--------|---|---|
| 1/2002 | Barbara Heller, Markus Löffler | Telematics and Computer-Based Quality Management in a Communication Network for Malignant Lymphoma |
| 2/2002 | Barbara Heller, Katrin Kühn, Kristin Lippoldt | Report OntoBuilder |
| 3/2002 | Barbara Heller, Katrin Kühn, Kristin Lippoldt | Handbuch OntoBuilder |
| 4/2002 | Barbara Heller, Katrin Kühn, Kristin Lippoldt | Leitfaden für die Eingabe von Begriffen in den OntoBuilder |
| 5/2002 | Mitarbeiter des IMISE | Skriptenheft für Medizinstudenten
Medizinische Biometrie
Medizinische Statistik und Informatik
(Kursus zum Ökologischen Stoffgebiet) |

2003

- | | | |
|--------|---|--|
| 1/2003 | Birgit Brigl, Thomas Wendt, Alfred Winter | Ein UML-basiertes Meta-Modell zur Beschreibung von Krankenhausinformationssystemen |
| 2/2003 | Thomas Wendt | Modellierung von Architekturstilen mit dem 3LGM ² |
| 3/2003 | Birgit Brigl, Thomas Wendt, Alfred Winter | Requirements on tools for modeling hospital information systems |
| 4/2003 | Madlen Dörschmann | Evaluation der Fehlerhäufigkeit im Rahmen einer Klinischen Studie |
| 5/2003 | Mohammad Zaino | Statistische Analyse zur Aufdeckung von neurotoxischen Störungen infolge langjähriger beruflicher Schadstoffexposition |

2004

- | | | |
|--------|--|--|
| 1/2004 | Mitarbeiter des IMISE | Skriptenheft zum SPSS-Kurs
Kurs zur Auswertung medizinischer Daten unter Verwendung des Statistikprogramms SPSS |
| 2/2004 | Renate Abelius, Barbara Heller, Luisa Mantovani, Frank Meineke, Roman Mishchenko, Jan Ramsch | Standardisierung von Studienkurzprotokollen - Qualitätsgesicherte rechnerbasierte Erfassung, Verarbeitung und Speicherung |
| 3/2004 | Jan Ramsch, Renate Abelius, Barbara Heller, Luisa Mantovani, Frank Meineke, Roman Mishchenko | Therapieschemata - Qualitätsgesicherte vereinheitlichte rechnerbasierte Erfassung, Verarbeitung und Speicherung |
| 4/2004 | Jan Ramsch | Variabilität beim Einsatz von onkologischen Therapieschemata - Erkennung von Ausnahmen und resultierenden Therapieänderungen |
| 5/2004 | André Wunderlich (Diss.) | Prognostische Faktoren für chemotherapieinduzierte Toxizität in der Behandlung von Malignomen speziell bei aggressiven Non-Hodgkin-Lymphomen |

6/2004	Mitarbeiter des IMISE	Skriptenheft für Medizinstudenten Methodensammlung zur Auswertung klinischer und epidemiologischer Daten
7/2004	Grit Meyer (Diss.)	Charakterisierung der zellkinetischen Wirkungen bei exogener Applikation von Erythropoetin auf die Erythropoese des Menschen mit Hilfe eines mathematischen Kompartimentmodells
2005		
1/2005	Ingo Röder (Diss.)	Dynamic Modeling of Hematopoietic Stem Cell Organization – Design and Validation of the New Concept of Within-Tissue Plasticity
2/2005	Katrin Braesel (Dipl.)	Modellierung klonaler Kompetitionsprozesse hämatopoetischer Stammzellen mit Hilfe von Computersimulationen
3/2005	Dr. Barbara Heller (Habil)	Knowledge-Based Systems and Ontologies in Medicine
2006		
1/2006	Alexander Strübing, Ulrike Müller	Evaluation des 3LGM ² Baukastens Studienplan - Ergebnisse - Auswertung
2/2006	Marc Junger (Diss.)	Benutzermodellierung bei der Qualitätssicherung im onkologischen Studienmanagement
3/2006	Thomas Wendt (Diss.)	Modellierung und Bewertung von Integration in Krankenhausinformationssystemen
2007		
1/2007	Markus Kreuz (Dipl.)	Entwicklung und Implementierung eines Auswertungswerkzeuges für Matrix-CGH-Daten
2/2007	Mitarbeiter des IMISE	Skriptenheft für Studenten Methodensammlung zur Auswertung klinischer und epidemiologischer Daten
3/2007	Frank Meineke (Diss.)	Räumliche Modellierung und Simulation der Organisations- und Wachstumsprozesse biologischer Zellverbände am Beispiel der Dünndarmkrypte der Maus
2008		
1/2008	Daniel Müller-Briel (Dipl.)	Standardisierung klinischer Studienprotokolle unter Berücksichtigung der Therapieplanung
2010		
1/2010	A. Winter, L. Ißler, F. Jahn, A. Strübing, T. Wendt	Das Drei-Ebenen-Metamodell für die Modellierung und Beschreibung von Informationssystemen (3LGM ² V3)