IMISE-REPORTS
Herausgegeben von Professor Dr. Markus Löffler

H. Herre, R. Hoehndorf, F. Loebe (Eds.)

OBML 2011
Workshop Proceedings

Berlin, October 6-7, 2011

IMISE-REPORT Nr. 1/2011

UNIVERSITÄT LEIPZIG
Medizinische Fakultät

Impressum

Herausgeber: Universität Leipzig
Medizinische Fakultät
Institut für Medizinische Informatik, Statistik und Epidemiologie (IMISE)
Härtelstraße 16-18, 04107 Leipzig
Prof. Dr. Markus Löffler

Editoren: Heinrich Herre, Robert Hoehndorf, Frank Loebe
Redakteur: Frank Loebe

Kontakt: Telefon: (0341) 97-16100, Fax: (0341) 97-16109
Internet: http://www.imise.uni-leipzig.de

Redaktionsschluss: 25. September 2011

Druck: Inhalt: Universitätsklinikum Leipzig AöR, Bereich 2 - Abteilung Zentrale Vervielfältigung/Formularwesen
Einband: Buch- und Offsetdruckerei Herbert Kirsten

ISSN 1610-7233

Proceedings of the

3rd WORKSHOP OF THE

GI WORKGROUP
"ONTOLOGIES IN BIOMEDICINE AND LIFE SCIENCES" (OBML)

Berlin, Germany
October 6-7, 2011

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

Organizers

Heinrich Herre	(*coordinator*)	University of Leipzig
Robert Hoehndorf		University of Cambridge, UK
Frank Loebe		University of Leipzig

Local Organizer

Peter Robinson	Charité Berlin

Keynote Speakers

Dietrich Rebholz-Schuhmann	European Bioinformatics Institute, Cambridge, UK
Peter Robinson	Charité Berlin

Program Committee

Robert Hoehndorf	(*chair*)	University of Cambridge, UK
Sören Auer		University of Leipzig
Franz Baader		Technical University Dresden
Martin Boeker		University Medical Center Freiburg
Patryk Burek		University of Leipzig
Fred Freitas		Federal University of Pernambuco, Recife, Brazil
Georgios V. Gkoutos		University of Cambridge, UK
Giancarlo Guizzardi		Federal University of Espirito Santo, Brazil
Heinrich Herre		University of Leipzig
Josef Ingenerf		University of Lübeck
Ludger Jansen		University of Rostock
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
Roberto Poli		University of Trento, Italy
Dietrich Rebholz-Schuhmann		European Bioinformatics Institute, Cambridge, UK
Peter Robinson		Charité Berlin
Michael Schröder		Technical University Dresden
Stefan Schulz		Medical University of Graz, Austria
Luca Toldo		Merck KGaA

Additional Reviewer

Anika Oellrich	European Bioinformatics Institute, Cambridge, UK

Authors

Maurício Almeida	Federal University of Minas Gerais, Brazil
André Andrade	Federal University of Minas Gerais, Brazil
Martin Boeker	University Medical Center Freiburg
Cristine Bonfim	Joaquim Nabuco Foundation, Recife, Brazil
Roberta Fernandes	Federal University of Pernambuco, Recife, Brazil
Fred Freitas	Federal University of Pernambuco, Recife, Brazil
Georgios V. Gkoutos	University of Cambridge, UK
Niels Grewe	University of Rostock
Heinrich Herre	University of Leipzig
Robert Hoehndorf	University of Cambridge, UK
Ludger Jansen	University of Rostock
Oliver Kutz	University of Bremen
Frank Loebe	University of Leipzig
Zulma Medeiros	Federal University of Pernambuco, Recife, Brazil
Anika Oellrich	European Bioinformatics Institute, Cambridge, UK
Djamila Raufie	University Medical Center Freiburg
Dietrich Rebholz-Schuhmann	European Bioinformatics Institute, Cambridge, UK
Norbert Reinsch	Leibniz Institute for Farm Animal Biology, Dummerstorf
Johannes Röhl	University of Rostock
Filipe Santana	Federal University of Pernambuco, Recife, Brazil
Daniel Schober	University Medical Center Freiburg
Stefan Schulz	Medical University of Graz, Austria
Aleksandra Sojic	European School of Molecular Medicine, Milan, Italy
Frank Stumpf	University of Leipzig
Vojtech Svatek	University of Economics, Prague, Czech Republic
Silke Trißl	Leibniz Institute for Farm Animal Biology, Dummerstorf
Ilinca Tudose	University Medical Center Freiburg

Preliminary Program

as of September 25, 2011

THURSDAY,	Oct 6, 2011	
(12:00 – 13:00)	(Getting together / Registration / *COFFEE*)	
13:00 – 13:20	H. Herre	Welcome Remarks
13:20 – 14:15	P. Robinson	*Keynote* : Phenotype Ontologies
14:15 – 14:30	*COFFEE*	
Session 1	Bio-Ontologies and Phenotype Ontologies	Chair: Frank Loebe
14:30 – 14:50	A. Oellrich	Quantitative Comparison of Mapping Methods between Human and Mammalian Phenotype Ontology
14:50 – 15:10	S. Trißl	Developing an Animal Trait Ontology: Why Phenotype Ontologies Are Not Enough
15:10 – 15:30	R. Hoehndorf	Ontology-Based Cross-Species Integration and Analysis of Saccharomyces Cerevisiae Phenotypes
15:30 – 16:00	*COFFEE*	
Session 2	Formal Ontology and Methodology I	Chair: Robert Hoehndorf
16:00 – 16:20	J. Röhl	The Ontology of Biological Mechanisms
16:20 – 16:40	D. Schober	Representing Casualties for Epidemiological Data Processing
16:40 – 17:00	N. Grewe	Continuation-like Semantics for Modeling Structural Event Anomalies
17:00 – 17:20	*COFFEE*	
17:20 – 17:40	L. Jansen	The Ten Commandments of Ontological Engineering
17:40 – 18:30	(*Workgroup*)	*Open Discussion*
19:30 – ?	*DINNER*	

FRIDAY,	Oct 7, 2011	
09:00 – 09:20	*COFFEE*	
09:20 – 10:15	D. Rebholz-Schuhmann	*Keynote* : Semantic Interoperability between Literature and Data Resources: From Genes to Diseases
10:15 – 10:30	*COFFEE*	
Session 3	Representation and Tools	Chair: Peter Robinson
10:30 – 10:50	F. Loebe	Towards Improving Phenotype Representation in OWL
10:50 – 11:10	D. Raufie	Redesigning an Ontology Design Pattern for Realist Ontologies
11:10 – 11:30	D. Schober	OntoCheck: Verifying Ontology Naming Conventions in Protégé 4
11:30 – 12:00	*COFFEE*	
Session 4	Foundations and Methodology II	Chair: Heinrich Herre
12:00 – 12:20	A. Sojic / O. Kutz	Beyond the Tumour: Breast Cancer Phenotypes – Towards a Pluralistic Integration of Heterogeneous Representations –
12:20 – 12:40	A. Andrade	Information, Reality and Epistemology: An Ontological Take
12:40 – 14:00	*LUNCH*	
14:00 – 15:00	*(Workgroup)*	*Open Discussion*

Table of Contents

Quantitative comparison of mapping methods between Human and Mammalian Phenotype Ontology

Anika Oellrich[1], George Gkoutos[2], Robert Hoehndorf[2], Dietrich Rebholz-Schuhmann[1]

[1]European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, CB10 1SD
[2]Department of Genetics, University of Cambridge, Downing Street, Cambridge, CB2 3EH

ABSTRACT

Researchers use animal studies to better understand human diseases. In recent years, large-scale phenotype studies such as Phenoscape and EuroPhenome have been initiated to identify genetic causes of a species phenome. Species-specific phenotype ontologies are required to capture and report about all findings and to automatically infer results relevant to human diseases. The integration of the different phenotype ontologies into a coherent framework is necessary to achieve interoperability for cross-species research.

Here, we investigate the quality and completeness of two different methods to align the Human Phenotype Ontology and the Mammalian Phenotype Ontology. The first method combines lexical matching with inference over the ontologies taxonomic structures, while the second method uses a mapping algorithm based on the formal definitions from the ontologies. Neither method could map all concepts. Despite the formal definitions method provides mappings for more concepts than does the lexical matching method, it does not outperform the lexical matching in a biological use case. Our results suggest that combining both approaches will yield to better mappings in terms of completeness, specificity and application purposes.

1 INTRODUCTION

Large-scale mutagenesis projects aim to identify the phenotypes of organisms resulting from modifications to the organisms' genetic markup and thereby provide the tantalizing possibility for revealing valuable information about the molecular mechanisms underlying human disease (Rosenthal and Brown 2007). In particular, phenotype studies in mice have been demonstrated to provide insights into human disease mechanisms (Schofield et al. 2010), and large phenotype studies are underway with the aim to identify mouse phenotypes resulting from deactivating every single gene in the organism (Abbott 2010, Collins et al. 2007). To describe phenotypes within a species and to allow access to the scientific community for further analyses, phenotype ontologies were created to standardize the terminology used in describing phenotypes, e.g. (Smith et al. 2004, Robinson et al. 2008).

We are now facing the challenge to enable the translation of these species-specific standardized phenotypic information across various species. Two approaches are currently in use for aligning species-specific phenotype ontologies. In the first approach, lexical mappings between the labels of concepts in species-specific phenotype ontologies are used to identify related phenotypes in different species. One implementation of this approach is the Lexical OWL Ontology Matcher (LOOM) (Ghazvinian et al. 2009) which has been shown to perform well on aligning anatomical ontologies. The second approach towards integrating phenotypes across species relies on formal definitions of concepts in phenotype ontologies using the Phenotypic Attribute and Trait Ontology (PATO) (Gkoutos et al. 2005) and the Entity-Quality (EQ) syntax (Mungall et al. 2010). The EQ representation allows for the phenotypic definitions to be integrated across species following the application of automated reasoning over their combination with a cross-species anatomy ontology (Mungall et al. 2010, Washington et al. 2009). The second approach is implemented in the PhenomeBLAST software (Hoehndorf et al. 2011a) and both, software and the resulting mappings, are publicly available from `http://phenomeblast.googlecode.com`.

It is generally challenging to evaluate and quantify the quality and completeness of ontologies (Yao et al. 2011). The challenge is amplified by mappings that involve and bridge multiple ontologies due to the presence of potentially conflicting or implicit conceptualizations by different ontology developers. Furthermore, both the quality of an ontology or of a mapping between ontologies are expected to depend on the specific use-case; ontologies that perform well in one application may not necessarily perform well in other use cases.

Here, we perform a descriptive evaluation of mappings between the Human Phenotype Ontology (HP) (Robinson et al. 2008) and the Mammalian Phenotype Ontology (MP) (Smith et al. 2004). We compare the mappings directly and quantify their quality for predicting gene-disease associations based on phenotype data. We find that both methods do not generate a mapping for all ontology concepts and consequently allow for further improvement. Despite the fact that the formal definitions method generates approximately four times more mapped concepts than the lexical matching, it does not outperform the lexical matching in the biological use case. Given the differences in mappings, shown by a deviation when directly comparing the mappings to each other, and availability of mappings with each method, a combination of the results of both methods may lead to mappings which are more comprehensive and specific. The combination may therefore also improve methods that rely on phenotypes for the prioritization of disease gene candidates.

2 MATERIALS AND METHODS

2.1 Ontological resources

Mammalian Phenotype Ontology (MP): We downloaded an MP version from (OBO foundry) which was created on the 8th April 2011 and comprised 8,507 concepts. The formal definitions for MP were downloaded separately from the same source. The file provided 5,389 MP concepts with an associated formal definition.

Human Phenotype Ontology (HP): The HP version used for this study, was downloaded from (HP). It was created on the 7th April 2011 and contained 10,104 concepts. The formal definitions were downloaded separately from the same source and provided formal definitions for 4,860 concepts.

2.2 Databases containing gene-disease associations

We used two community-wide established resources containing manually verified gene and disease related data: the Mouse Genome Informatics (MGI) (Blake et al. 2011) and the Online Mendelian Inheritance in Man (OMIM) (Amberger et al. 2011) database.

The MGI database integrates genetic, genomic and phenotypic information about the laboratory mouse (Blake et al. 2011). For this study, three of the report files from the MGI database were downloaded (Jackson laboratory)

- MGI_GenoDisease.rpt, accessed on 9th March 2011,
- MGI_GenePheno.rpt, accessed on 9th March 2011, and
- HMD_Human5.rpt, also accessed on 9th March 2011.

MGI_GenoDisease.rpt contained associations between diseases and specific genotypes (one genotype corresponds to one mouse model) that can be linked to affected genes. MGI_GenePheno.rpt contained the information about genotypes and their observed phenotypes, which are described in MP. HMD_Human5.rpt covered the information about human-mouse orthologous genes.

The OMIM database collects information about human inheritable diseases, including genotype and phenotype information, and known gene-disease associations. It contains about 20,000 entries out of which around 13,000 describe genes and about 7,000 describe diseases. MorbidMap (downloaded on 1st March 2011) contains the up to date information about known links between human diseases and genes. The downloaded version for this study contained 2,717 diseases being linked to 2,266 genes, with 3,463 distinct gene-disease associations. Phenotypic information (HP annotations) for OMIM diseases are available from the HP web page (HP). The downloaded file comprised annotations for approximately 4,000 OMIM entries.

2.3 Mappings between species-specific phenotype ontologies

2.3.1 Mappings between ontologies Let O_1 and O_2 be two ontologies with a set of named concepts $C(O_1)$ and $C(O_2)$. A mapping between O_1 and O_2 is a set of axioms $Ax = \{\phi_1(x_1,y_1),...,\phi_n(x_n,y_n)\}$ such that $x_i \in C(O_1)$ and $y_j \in C(O_2)$.

Here, we focus on mappings where the axioms relating concepts from two ontologies take the form of sub-class and equivalent-class axioms between atomic concepts. In particular, given the two concepts $A \in O_1$ and $B \in O_2$, a mapping involving both A and B will be of the form

- `A SubClassOf: B`, or
- `B SubClassOf: A`, or
- `A EquivalentTo: B`.

2.3.2 Generating mappings through lexical matching In this study, we used the Lexical OWL Ontology Matcher (LOOM) (Ghazvinian et al. 2009) to generate the lexical matching of concepts between ontologies. LOOM was applied to HP and MP concept names and synonyms. Based on names and synonyms, LOOM extracted 495 HP-MP concept pairs in the form

```
HP:0002249   MP:0003292 .
```

We imported both ontologies into one single ontology, inserted the pairs extracted by LOOM as equivalence statements and reasoned over the ontology. We generate the mapping by extracting the equivalent and super concepts belonging to the other ontology. In most cases, one concept from one ontology was mapped to multiple concepts from the other ontology.

An example of the resulting mapping looks like

```
HP:0007062   MP:0000001 MP:0002106 MP:0004142
             MP:0004143 MP:0005369
```

Due to both ontologies differing in their structure, the mappings are not symmetrical. For example, HP:0008590 'Progressive childhood hearing loss' maps to MP:0006325 'Impaired hearing' but MP:0006325 maps to HP:0000365 'Hearing impairment' (only most specific concepts are given in this example).

The resulting mappings together with the ontology file can be downloaded from the project web page `http://code.google.com/p/ontmapcomp/`.

2.3.3 Mapping through automated reasoning PhenomeBLAST integrates the formal definitions that were created for classes from the HP and MP (Hoehndorf et al. 2010), including several other ontologies, such as Gene Ontology and UBERON. The ontologies are all converted into OWL EL to enable efficient automated reasoning (Hoehndorf et al. 2011b). PhenomeBLAST then uses the CB reasoner to classify the ontology (Kazakov 2009). To generate the mappings from MP to HP, PhenomeBLAST identifies all equivalent and superclasses of an MP class in HP, and *vice versa* for the direction of HP to MP. The mappings generated by the PhenomeBLAST software are available at `http://phenomeblast.googlecode.com` and for this study we downloaded the mappings provided (June 2011).

2.4 Direct comparison of mappings

The lexical matching method as well as the formal definitions method generate non-symmetrical mappings for each of the ontologies which results in four mappings in total (compare bottom rows in table 1). Due to the non-symmetry, the generated mappings had to be investigated independently. For the concepts being represented with either method, we compared the lists of mapped concepts with each other and determined how well the lists overlapped. The direct comparison was executed for both ontologies independently, HP to MP and MP to HP.

2.5 Impact of mapping methods on applications

To assess and quantify the quality of mappings, we additionally used the biological use case of disease candidate gene prediction to evaluate the performance of each method. For that purpose, we used the phenotypic descriptions of mouse models contained in MGI_GenePheno.rpt and the OMIM disease HP annotations. Due to the non-symmetry in mappings of either method, we investigated two different scenarios: in the first we "translated" the mouse model MP descriptions to HP using either methods' mapping, whilst for the second we "translated" the OMIM disease HP descriptions to MP. We identified the phenotypic similarity between all possible combinations of mouse models and diseases by calculating the phenotype similarity. The phenotype similarity is the cosine similarity between the vector representations of a disease and a mouse model. In the first scenario, both feature vectors are built using MP concepts and in the second, both feature vectors contain HP concepts.

The phenotype similarity score for each disease-model pair was used to rank the mouse models according to their phenotype similarity for each disease. Then, we compared the obtained gene-disease (each mouse model is associated with one gene) pairs to OMIM and recorded the ranks of the known gene-disease associations to evaluate the performance of each method. In the absence of true negative examples, we assume that *known* gene-disease associations constitute *positive* examples while *unknown* associations constitute *negative* examples. The true and false positive rates are calculated across all diseases and over all mouse models possessing a phenotypic representation compared to the in MorbidMap contained gene-disease associations. Both true and false positive rates are then used to draw the Receiver Operating Characteristics (ROC) curves (compare figure 3.3) for both scenarios of the biological use case.

3 RESULTS AND DISCUSSION

3.1 Generated mappings

Table 1 shows the number of mapped concepts available for each ontology and each method. For the formal definitions method, 80% of HP concepts and 50% of MP concepts can be mapped, whereas the lexical matching method provides a mapping for 27% and 12% respectively. Despite the formal definitions method producing a mapping for about four times more concepts than the lexical matching method does, the average amount of mapped concepts to one particular concept is lower. The lower number of mapped concepts for one particular concept suggests that the formal definitions method maps to more generalized concepts (which are higher in the taxonomy) of the other ontology.

Both methods are hampered by the definition of concepts in the ontology. The number of mapped concepts and the specificity of the mappings generated by the formal definitions method depends solely on the availability and quality of the formal definitions for both ontologies, which constitutes an advantage at the same time. E.g. a complex phenotypic expression in HP like *Tetralogy of Fallot* which would have no corresponding concept in MP, can still be mapped as long as it is formally defined. The lexical method is limited by the naming of the concepts which is demonstrated by the low number of concepts being mapped from each of the ontologies (four times less than the formal definition method). The number of mapped concepts could potentially be increased by using a less strict text matching algorithm but the method would still rely on the words being used for naming a concept or its synonyms. On average, the method allows for matching more specific concepts than does the formal definitions method indicated by the higher number of mapped concepts from one ontology to the other (see table 1). Given the complexity of some of the phenotypes contained in either ontology, it is still challenging to find appropriate formal definitions in which case the lexical method may align concepts, given that they exist in both ontologies.

3.2 Direct comparison of mappings

When comparing the mappings directly to each other, we identified five types of overlap, indicating a deviation in the mappings produced by both the methods. The five different types of overlap are illustrated in figure 3.2. The amount of concepts falling into each of the five overlap categories are illustrated in table 2. The table shows that only a low proportion of **exact** matches exists and most of the results fall into the **overlap** category.

The direct comparison of mappings produced by each method shows that for most of the concepts common to both methods, the mappings share at least some overlap (categories **exact**, **formal** $\subset$ **lexical**, **lexical** $\subset$ **formal** and **overlap**), even though the number of **exact** matches is low. Four of the categories, **formal** $\subset$ **lexical**, **lexical** $\subset$ **formal**, **overlap**, and **nothing** indicate a deviation in both mappings. The category **nothing** points to potential errors in the mappings produced by either method and present a good starting point for further investigations. Once the errors have been eliminated, the distribution of results over all other overlap categories will change.

Given that both methods generate mappings for concepts which are not contained in the other (compare table 1 and 2) and the fact that the results appear as subsets of each other for some concepts (see figure 3.2, categories b), c) and d)), it seems to be worthwhile to combine both the approaches and generate one mapping incorporating the results of both methods.

3.3 Impact of mapping methods on biological applications

Figure 3.3 shows the Receiver Operating Characteristic (ROC) curves for predicting gene-disease associations contained in OMIM's MorbidMap. The true and false positive rates are calculated across all diseases and over all mouse models possessing a phenotypic representation compared to the in MorbidMap

Table 1. Illustrates the numbers of concepts contained in each ontology but also incorporates the results of the mapping methods. The first bracket is the percentage calculated based on the total number of available concepts in each ontology (see section 2.1) and the second is the average number of concepts one particular concept is mapped to.

	HP			MP		
	HP	% total	avg # mapped	MP	% total	avg # mapped
# concepts	10104	100%	-	8507	100%	-
# with formal definition	4860	48.10%	-	5389	63.35%	-
# mapped with lexical	2740	27.12%	7.17	1046	12.30%	6.97
# mapped with formal definitions method	8184	80.10%	5.48	4446	52.26%	6.64

Table 2. Illustrates the amount of mappings falling into each of the overlap categories when both methods are compared. The mappings for HP to MP and MP to HP are compared independently due to non-symmetrical mappings.

	HP to MP	MP to HP
# exact	155	70
# lexical $\subset$ formal	755	287
# formal $\subset$ lexical	496	114
# overlap	952	215
# nothing	74	0
# concepts	2432	686

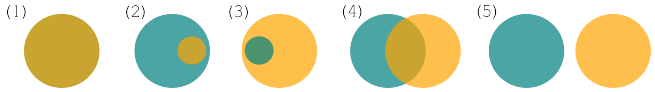


Fig. 1: Shows the different types of obtained overlap while directly comparing the mappings generated by both methods, regardless of the ontology the mapping is provided for. The amount of mapped concepts for the formal definitions method is represented with a yellow circle and the lexical matching is illustrated with a turquoise circle. We identified the following five categories: a) **exact** (both lexical matching and formal definitions method generated exactly the same list of mapped concepts), b) **formal** $\subset$ **lexical** (mapping generated by the formal definitions method is a subset of the list generated by lexical matching), c) **lexical** $\subset$ **formal** (mapping generated by lexical matching is a subset of the list generated by the formal definitions method), d) **overlap** (both lists contain additionally mapped concepts and share only a certain overlap), and e) **nothing** (despite both methods generate a list of mapped concepts for a specific concept, both lists have nothing in common).

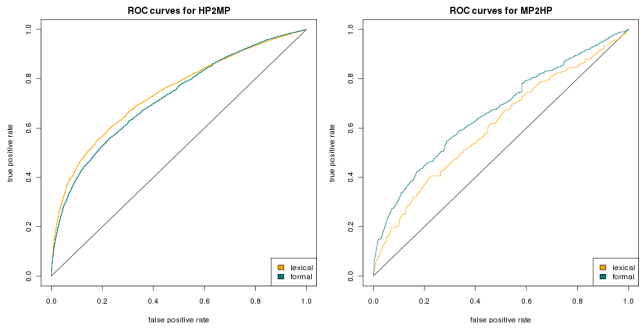


Fig. 2: Shows the Receiver Operating Characteristic (ROC) curves for both scenarios: the left panel illustrating the case where alleles are "translated" to HP and the right illustrating the case where diseases are "translated" to MP. In the first scenario the application of the lexical mappings (AUC: 0.74) seems to have better performance than the formal definitions mappings (AUC: 0.72), whereas in the second scenario the formal definitions mappings (AUC: 0.66) seem to yield better results in the biological use case than the lexical mappings (AUC: 0.61).

contained gene-disease associations. We assume that *known* gene-disease associations constitute *positive* examples while *unknown* associations constitute *negative* examples.

The left panel of figure 3.3 corresponds to the first scenario in which OMIM diseases are "translated" from HP to MP and the candidate gene prediction is performed by comparing sets of MP concepts. The results show that if the lexical mappings are used, the overall performance for this particular biological use case is better (AUC 0.74) than the mappings generated through automated reasoning (AUC 0.72). The results may be explained with the fact that the HP-based annotations of OMIM diseases use specific ontology concepts (concepts which are deeper in the hierarchy of an ontology). These specific terms (such as *Eosinophilia*) can often be accurately mapped through lexical matching, while a formal definition may not be available due to the complexity of the underlying phenomenon.

The right panel of figure 3.3 corresponds to the second scenario in which alleles are "translated" from MP to HP and the candidate gene prediction is performed by comparing sets of HP concepts. The results illustrate that in this particular use case, the application of the formal definitions mappings leads to a better performance (AUC 0.66) than the lexical mappings (AUC: 0.61). Mouse models are less frequently annotated with specific ontology classes that can accurately be mapped through lexical matching. Automated reasoning over the formal definitions provides a sufficient number of mappings for classes that are less specific, while lexical matching does not establish these mappings. Consequently, more information is retained when using ontology-based mappings and the prediction of known gene-disease associations performs better.

4 CONCLUSIONS

We have evaluated and compared two methods for aligning HP and MP. The first method is based on lexical matching, whereas the second method uses automated reasoning and formal definitions of phenotypes to perform the mapping. While automated reasoning over the formal definitions generates more mappings between both ontology than lexical matching, these mappings are, on average, less specific than the mappings established through lexical matching. As a result, the mappings perform differently when used for prioritizing disease gene candidates, depending on whether disease phenotypes (which use specific HP phenotypes) are translated into an MP-based representation, or whether MP-based descriptions of mouse genotypes are translated into an HP-based description.

In future research, we intend to extend our analysis of mapping methods and identify strategies to further combine both approaches. Our comparative evaluation can help to improve phenotype-based methods for predicting gene-disease associations and may further extend their capabilities for identifying new gene-disease associations.

REFERENCES

A. Abbott. Mouse megascience. *Nature*, 465:526, 2010.

Joanna Amberger, Carol Bocchini, and Ada Hamosh. A new face and new challenges for online mendelian inheritance in man (OMIM). *Hum Mutat*, 2011. ISSN 1098-1004.

Judith A. Blake et al. The Mouse Genome Database (MGD): premier model organism resource for mammalian genomics and genetics. *Nucleic Acids Res*, 39(Database issue):D842–D848, Jan 2011.

FS Collins, RH Finnell, J Rossant, and W Wurst. A new partner for the international knockout mouse consortium. *Cell*, 129(2):235, 2007.

Amir Ghazvinian et al. Creating mappings for ontologies in biomedicine: simple methods work. *AMIA Annu Symp Proc*, 2009:198–202, Nov 2009.

Georgios V. Gkoutos et al. Using ontologies to describe mouse phenotypes. *Genome biology*, 6(1), 2005. ISSN 1465-6914.

Robert Hoehndorf, Anika Oellrich, and Dietrich Rebholz-Schuhmann. Interoperability between phenotype and anatomy ontologies. *Bioinformatics*, 26(24):3112–8, Dec 2010.

Robert Hoehndorf, Paul N. Schofield, and Georgios V. Gkoutos. Phenomenet: a whole-phenome approach to disease gene discovery. 2011a.

Robert Hoehndorf et al. A common layer of interoperability for biomedical ontologies based on OWL EL. *Bioinformatics*, Feb 2011b.

HP. HP download page. URL `http://compbio.charite.de/svn/hpo/trunk/src/ontology`.

Jackson laboratory. Download page for Mouse Genome Informatics database report files. URL `ftp://ftp.informatics.jax.org/pub/reports/index.html`.

Y Kazakov. Consequence-driven reasoning for horn shiq ontologies. *Proc. of IJCAI-09*, Jan 2009.

Christopher Mungall et al. Integrating phenotype ontologies across multiple species. *Genome Biology*, 11(1):R2+, 2010.

OBO foundry. OBO foundry web page. URL `http://www.obofoundry.org/`.

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, November 2008. ISSN 1537-6605.

Nadia Rosenthal and Steve Brown. The mouse ascending: perspectives for human-disease models. *Nature Cell Biology*, 9: 993 – 999, 2007.

Paul N. Schofield et al. Phenotype ontologies for mouse and man: bridging the semantic gap. *Disease Models & Mechanisms*, 3 (5-6):281–289, May 2010.

Cynthia L. Smith, Carroll, and Janan T. Eppig. The Mammalian Phenotype Ontology as a tool for annotating, analyzing and comparing phenotypic information. *Genome Biology*, 6(1):R7, 2004.

Nicole L. Washington et al. Linking human diseases to animal models using ontology-based phenotype annotation. *PLoS Biol*, 7(11):e1000247, 11 2009.

Lixia Yao et al. Benchmarking Ontologies: Bigger or Better? *PLoS Comput Biol*, 7(1):e1001055, 01 2011.

Developing an Animal Trait Ontology – Why Phenotype Ontologies are not enough.

Silke Trißl* and Norbert Reinsch

Leibniz Institute for Farm Animal Biology, Wilhelm-Stahl-Allee 2, 18196 Dummerstorf, Germany.
Email: {trissl, reinsch}@fbn-dummerstorf.de

ABSTRACT

Describing features or traits of farm animals as exactly as possible is important. Current phenotype ontologies usually only allow to specify in which form an observed characteristic differs from the norm. But this is not sufficient, as even minor differences in milk production, meat quality or quantity, or behavior are commercially relevant.

To compare performance or behavior traits of animals we argue that they should be described using concepts of ontologies. We propose to use the EAV (Entity-Attribute-Value) approach. We extensively discuss which ontologies may be used for the three parts entity, attribute, and value.

1 INTRODUCTION

Ontologies provide an excellent opportunity to describe objects and observations in an area of interest. Concepts of an ontology allow for a standardized description. Thus, observations or objects described by different individuals may become comparable for humans, but are also accessible for computational analyses. In addition, an ontology provides relationships between concepts. Using these relationships we are able to infer knowledge by incorporating parent or ancestor concepts in an analysis (Resnik (1999)).

Our area of interest is the performance and behavior of farm animals. Our main focus lies on the commercially interesting species pig and cattle. An important research area in farm animal biology is to map observed features of an animal, such as meat quality and quantity, milk yield, or health status to genomic regions, called QTL (quantitative trait loci) (Geldermann (1975)). The knowledge which genomic regions may influence which features are used in breeding programs.

In recent years the behavior of animals, e.g., if they are calm, shy, or aggressive, plays a more and more important role. Researchers believe that behavior is correlated to health status or meat quality (Beattie *et al.* (2000)). In addition, the welfare of animals comes more and more into focus of authorities and customers (Vanhonacker *et al.* (2008)).

Currently, we find descriptions of features only in articles as non-structured free-text, because up to now there exist no standardized way of describing features of farm animals (Smith and Eppig (2009)). In our opinion this should be changed and thus, we propose to facilitate ontologies for describing performance and behavior features of farm animals. We intend to use ontologies to allow for a consistent description of farm animals in Phänomics, which is a project involving dozens of researchers on ten different research institutes.

Ontologies have been developed to describe anatomical features, phenotypes, and traits. The differentiation of the ontology types is not always trivial as the following example shows. The disease 'obesity' is associated with the phenotype 'Increased lipid weight'. The trait, the figure which is measured, is 'lipid weight' in kg. Lipid may be found at various anatomical parts, such as the 'abdomen'. This small example shows that a distinction between the content of different ontologies should exist, but this distinction may be very narrow and ontologies may even overlap.

In the following we present in Section 2 the preferred model to represent trait information for an animal. In Section 3 we describe existing ontologies that are relevant for the description of features of farm animals. Finally, in Section 4 we conclude the paper.

*to whom correspondence should be addressed

2 REPRESENTING TRAIT BY USING ONTOLOGIES

According to Collins English Dictionary (2010) a trait is "a characteristic feature or quality distinguishing a particular person or thing". A trait may be physical, such as the thickness of abdominal fat, or it may be behavioral, e.g., how active or how lethargic an animal is.

In general trait information is comprised of three parts, as Figure 1 shows. The *feature* is either the anatomical or behavioral feature that should be described. The *value* is the experimentally determined value for that feature. Each value is determined by an *assay*. There exists a representation that allows storing these triples, known as the Entity-Attribute-Value (EAV) approach (Nadkarni *et al.* (1999)). Gkoutos *et al.* (2005) propose this representation for describing mouse phenotypes. The entities are features, the attribute is the assay with which the value of this feature has been created, while the experimentally determined value is stored in value.

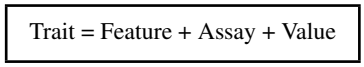


Fig. 1. Trait is comprised of three parts of information. The feature, which should be described, the value for this feature, and the assay with which the value for this feature has been measured.

To allow for comparability between different studies all three parts that comprise the description of trait should use controlled vocabulary. Advanced versions of controlled vocabularies are ontologies (Gruber (1993)). For the biomedical community the Open Biological and Biomedical Ontologies Foundry (Smith *et al.* (2007)) lists and provides numerous ontologies. In addition, several groups have developed their own ontology for their area of interest. In the following section we describe ontologies that may be interesting for the description of trait in farm animal biology.

3 EXISTING ONTOLOGIES

Developing an ontology from scratch is tedious and time consuming. Thus, we want to use existing ontologies to describe trait information of farm animals. In the following we discuss for all three components, i.e., feature, value, and assay, which ontologies may be employed.

3.1 Ontologies for *Feature*

We may apply two groups of ontologies, namely anatomy or phenotype and trait ontologies for the description of physical and behavioral features of farm animals. In the following we present selected ontologies of both groups.

3.1.1 Anatomy Ontology The OBO Foundry lists 40 anatomy ontologies (as of August 2011). The majority (25) of these ontologies provide concepts for species that are not relevant for farm animal biology, such as insects, amphibia, arachnida, fungi, or plants. Some (5) ontologies describe the anatomy at cellular level, in which we are currently also not interested. We discuss the remaining anatomy ontologies in the following.

SNOMED CT (Wang *et al.* (2002)) is an ontology that has been created for clinical applications. Its main focus lies on humans for which it is well accepted (Cornet and de Keizer (2008)). SNOMED CT also contains concepts used in veterinary medicine such as *8384002*:'foot and mouth disease' or *27528008*:'udder'. Although SNOMED CT contains concepts of interest for farm animals Zimmerman *et al.* (2005) show that only about 50 % of the concepts extracted from text documents dealing with farm animals could be mapped to concepts in SNOMED CT. This figure is lower compared to similar studies for humans, where about 70 % of concepts are mapped. The authors conclude that SNOMED CT should be enriched with additional concepts for veterinary medicine.

Concepts of the anatomy sub-ontology of SNOMED CT have been mapped to concepts of the *Foundational Model of Anatomy* (FMA) (Bodenreider and Zhang (2006)). Although FMA captures only the anatomy of humans (Rosse and Mejino (2003)) this mapping may be useful for comparative analyses between different species.

In farm animal biology the mouse is also used as model organism. To describe anatomical features of the mouse the *Mouse Adult Gross Anatomy* (MA) may be used (Hayamizu *et al.* (2005)). This ontology only provides concepts to describe features for the adult mouse. The *Mouse gross anatomy and development* (EMAP) is used to describe the developmental process of the mouse.

Not only mammals are interesting for farm animal biology, but also fish. The *Teleost Anatomy Ontology* (TAO) is an ontology to describe anatomical features for several species of teleost fish. Dahdul *et al.* (2010) describe challenges and difficulties of creating an ontology for diverse species. One of the problems is an ever growing ontology as the specificities for each species should be captured. In addition, ambiguity of terms within different research communities must be incorporated. On the other side, they also describe the opportunities such an ontology provides, e.g., the possibility to map anatomical features of different species to each other.

In OBO several other ontologies exist that provide concepts suitable to describe features in various species. Bard *et al.* (2008) describe the *Minimal Anatomy Terminology* (MAT). It is based on several anatomy ontologies, which have been merged. MAT allows the description of anatomical features of about 500 different species. On the downside, MAT only contains some 460 concepts. This number may not be sufficient to describe very specific parts in an animal of interest. Haendel *et al.* (2009) present *Uberon*, a multi-species anatomy ontology intended for metazoan. In contrast to MAT Uberon contains some 6,200 concepts for the description of anatomical features.

High-level ontologies contain general concepts, which may be mapped to species-specific ontologies. This mapping may alleviate the identification of similar anatomical structures in different species. One example of a high-level ontology for anatomy is the *Common Anatomy Reference Ontology* (CARO) (Haendel *et al.* (2008)). Two ontologies extend CARO, which are the *Anatomical Entity Ontology* (AEO)[1] and the newly developed *Vertebrate Anatomy Ontology* (VAO)[2]. Although these ontologies allow for a mapping between different ontologies the concepts may not be useful for the direct description of features in farm animals.

The presented ontologies contain concepts to describe anatomical features of farm animals. In our research we are also interested in describing the behavior of farm animals. Although some behavior patterns may be associated with specific body parts, the main focus of behavior does not lie on anatomy. Thus, we continue to phenotype and trait ontologies.

3.1.2 Phenotype and Trait Ontologies In literature exists no clear distinction between phenotype and trait. According to Collins English Dictionary (2010) the phenotype are "the physical and biochemical characteristics of an organism as determined by the interaction of its genetic constitution and the environment". This means the phenotype may be considered as the sum of all observable features. In contrast, trait, which is also known as phenotypic trait, is a single characteristic feature or quality.

The OBO Foundry lists seven phenotype and trait ontologies. Two of the ontologies, namely Units of Measurement (UO) and Phenotypic Quality (PATO), are discussed in Section 3.2. The three ontologies C. elegans Phenotype (WBPhenotype), Ascomycete Phenotype Ontology (APO), and Plant Trait Ontology (TO) are not relevant in farm animal biology. We discuss the remaining two ontologies, which are Human Phenotype Ontology (HP) and Mammalian Phenotype (MP) in the following.

The *Human Phenotype Ontology* (HP) focuses on concepts to describe hereditary diseases in humans (Robinson *et al.* (2008)). The main focus of HP lies on the distinction of normal versus abnormal phenotypes. For example, HP contains the concept *HP:0002813*:'Abnormality of the extremities'. This concept contrasts the Entity-Attribute-Value approach, in which the fact would be represented as (Entity: *MA:0000007*:'extremity', Attribute: 'visual', Value: *PATO:0000460*:'abnormal'). As HP only targets human diseases very few concepts may be of interest for farm animal biology.

The ontology for *Mammalian Phenotype* (MP) was developed to describe abnormal mammalian phenotypes primarily in mice and rats, but has been used for other mammals as well (Smith and Eppig (2009)). MP contains for example the concept

[1] Anatomical Entity Ontology – NCBO BioPortal – http://purl.bioontology.org/ontology/AEO

[2] Vertebrate Anatomy Ontology – NCBO BioPortal – http://purl.bioontology.org/ontology/VAO

MP:0000545:'abnormal limbs/digits/tail morphology', which may be considered similar to *HP:0002813*.

The broad classification normal versus abnormal, which is by now often used for phenotype description (Crusio (2002)), may not be sufficient for a researcher to describe trait information. Thus, further ontologies, some of which are still under development and not yet members of the OBO Foundry, should be considered to describe trait information in farm animals.

Hughes *et al.* (2008) detected the need for an animal trait ontology to describe performance and behavior features. They developed the *Animal Trait Ontology* (ATO), which provides concepts for the most relevant farm animal species *Bos taurus* (cattle), *Sus scrofa* (pig), and *Gallus gallus* (chicken). The ontology contains three sub-ontologies, one for each species. We believe this approach may not be purposeful as for each species a new and individual sub-ontology has to be developed. The redevelopment of ATO resulted in the *Animal Trait Ontology for Livestock* (ATOL)[3] where the distinction between species is abrogated. A concept may be marked as being applicable to a certain species, e.g., *VT1000155*:'milk yield' to cattle, sheep, and goat, but not for chicken and trout. ATOL focuses on performance and animal welfare and provides concepts such as *PH:0000100*:'meat ratio of omega-6/omega-3' or *PH:0000813*:'aggressive behavior'.

The *Vertebrate Trait Ontology* (VT)[4] developed at the Medical College of Wisconsin has recently been made available and is intended for morphological, physiological and developmental traits in vertebrates. It takes over some concepts from another Vertebrate Trait Ontology (VTO)[5] developed by the NAGRP Bioinformatics Team. In contrast to VT the VTO contains a small sub-ontology to capture behavior.

The extensive behavior sub-ontology of ATOL may be complemented by concepts from the *Neuro Behavior Ontology* (NBO)[6], which targets the description of behavior in general. The ontology contains the concept *NBO:0000095*:'fear towards living things' as successor concept of *NBO:0000018*:'fear/anxiety related behavior'. To describe the reaction of an animal during human interaction these concepts may be used. For behavior we may also consider concepts of the *Mammalian Behavior Ontology* (MBO) (Beck *et al.* (2009)), which unfortunately does not seem to have left the status of a draft yet.

In contrast to HP and MP the remaining presented phenotype and trait ontologies provide concepts without judgment. Thus, we may use these ontologies to describe features of performance and behavior traits using the EAV approach. Still, the question remains, if any of the presented ontologies is sufficient to describe anatomical and behavioral features of farm animals. We may be able to collaborate with a group developing a Trait Ontology by contributing concepts, which we consider important. Another option is to develop an own ontology that extensively borrows concepts and relationships from some of the above presented ontologies.

3.2 Ontologies for *Value*

One part of trait information is the determined value for a feature. This value may be numerical, e.g., the abdominal fat has a thickness of 5 cm, or categorical, e.g., the fur color. In both cases ontologies may be useful. The OBO Foundry lists two ontologies of interest.

The *Phenotypic Quality Ontology* (PATO) (Gkoutos *et al.* (2005)) contains concepts for categorical values. For example, to describe the fur color of animals the concepts *PATO:0001245*:'dark brown', *PATO:0001252*:'light grey', or *PATO:0000323*:'white' may be used. The selection of color may be extended, as 'beige' or 'pink' are also valid colors for animals, but not present in PATO. Although PATO is quite extensive with about 2,200 concepts, it does not provide a terminology for units of values. In this case the *Units of Measurement Ontology* (UO)[7] may be used. It contains for example the concepts *UO:0000016*:'millimeter' and *UO:0000009*:'kilogram'.

Concluding, for the description of values for trait the existing ontologies PATO and UO are a good foundation. As assays and experiments are conducted we may find it necessary to propose extensions for those two ontologies.

3.3 Ontologies for *Assay*

The OBO Foundry lists eight ontologies for the description of assays. Most of these are concerned with the description of molecular biology experiments and assays. Only one listed ontology may be applicable for assays to determine physiological and behavioral features of farm animals.

The *Ontology for Biomedical Investigations* (OBI)[8] provides terms to describe clinical and biological investigations. For example, concepts in OBI are *obo:OBI_0000418*:'measuring glucose concentration in blood serum' and *obo:OBI_0000694*:'animal feeding'. Although OBI contains those two concepts, its main focus lies on the description of molecular biology experiments. Thus, to describe performance and behavior experiments we need to add several concepts to the ontology.

4 CONCLUSION AND FUTURE DIRECTIONS

In this work we argue that describing features of a farm animal as exactly as possible is important. Current phenotype ontologies usually only allow to specify in which form an observed characteristic differs from the norm. This is not sufficient, as even minor differences in milk production, meat quality or quantity, or behavior are commercially relevant.

To describe a trait of a farm animal we have to provide the feature, its measured value, and the assay with which it has been measured. We propose to employ the Entity-Attribute-Value (EAV) approach for the description. For each part of the EAV triple we may employ concepts of an ontology.

The NCBO BioPortal[9] lists in total 287 ontologies as of August 2011. In this work we highlight 21 ontologies, 17 of which are

[3] pers. communication with Pierre-Yves Le Bail, August 2011

[4] Vertebrate Trait Ontology – NBCO BioPortal – http://purl.bioontology.org/ontology/VT

[5] AmiGO: Tree Browser – http://www.animalgenome.org/cgi-bin/amion/browse.cgi

[6] Neuro Behavior Ontology – NCBO BioPortal – http://purl.bioontology.org/ontology/NBO

[7] Units of Measurement Ontology – NCBO BioPortal – http://purl.bioontology.org/ontology/UO

[8] Ontology for Biomedical Investigations – NCBO BioPortal – http://purl.bioontology.org/ontology/OBI

[9] Welcome to the NCBO BioPortal — NCBO BioPortal – http://bioportal.bioontology.org/

Table 1. The table summarizes the presented ontologies. We show for each ontology present at the NCBO BioPortal the number of concepts as of August 2011.

Name	Abbreviation	# concepts
SNOMED Clinical Terms	SNOMED CT	391,173
Foundational Model of Anatomy	FMA	71,586
Mouse Adult Gross Anatomy	MA	2,968
Mouse Gross Anatomy and Development	EMAP	13,731
Teleost Anatomy Ontology	TAO	3,039
Minimal Anatomical Terminology	MAT	461
Uber Anatomy Ontology	Uberon	6,241
Common Anatomy Reference Ontology	CARO	48
Anatomical Entity Ontology	AEO	137
Vertebrate Anatomy Ontology	VAO	189
Human Phenotype Ontology	HP	10,209
Mammalian Phenotype	MP	8,668
Vertebrate Trait Ontology	VT	3,056
Neuro Behavior Ontology	NBO	730
Phenotypic Quality Ontology	PATO	2,228
Units of Measurement	UO	2,507
Ontology for Biomedical Investigations	OBI	3,372

listed at the NCBO BioPortal. Table 1 shows the number of concepts present in each ontology.

For the entity we propose to use a mixture of various anatomy ontologies to cover anatomical features, the Animal Trait Ontology for Livestock (ATOL) and Neuro Behavior Ontology (NBO) to describe performance and behavior features. We would like to present a single ontology to biologists. Thus, the first step is to extract relevant sub-ontologies and merge these under a single root. This approach is desired and supported by the OBO Foundry (Ghazvinian *et al.* (2011)).

The experimentally determined values can already be described using the Phenotypic Quality Ontology and Units of Measurement. We may require amendments to both ontologies. The biggest problem currently is to find suitable ontologies that allow describing assays. Again, we would prefer to provide biologists within the Phänomics project with a single ontology.

The selection of ontologies is one aspect. Another important point is the usage of the ontology. Features of an animal should be described according to the EAV approach, but understanding this concept may be difficult for a researcher. Thus, good software tools are required that are intuitive and easy to use, like Phenex (Balhoff *et al.* (2010)).

In addition, trait ontologies pose new challenges for knowledge inference. While in a phenotype ontology 'increased abdominal fat' may be child of 'increased body weight', in the EAV approach we know that a measurement for abdominal fat has been made, but values and assays of the measurement, i.e., numbers and methods, also must be made comparable.

ACKNOWLEDGEMENT

This work is supported by the BMBF supported project Phänomics (grant no. 035539G).

REFERENCES

Balhoff, J. P., Dahdul, W. M., Kothari, C. R., Lapp, H., *et al.* (2010). Phenex: Ontological Annotation of Phenotypic Diversity. *PLoS One*, **5**(5), e10500.

Bard, J., Malone, J., Rayner, T., and Parkinson, H. (2008). Minimal Anatomy Terminology (MAT): a species-independent terminology for anatomical mapping and retrieval. In *AEEI Conference*. Toronto, Canada.

Beattie, V. E., O'Connell, N. E., and Moss, B. W. (2000). Influence of environmental enrichment on the behaviour, performance and meat quality of domestic pigs. *Livestock Production Science*, **65**(1-2), 71 – 79.

Beck, T., Hancock, J. M., and Mallon, A.-M. (2009). Developing a mammalian behaviour ontology. Available from Nature Precedings <http://dx.doi.org/10.1038/npre.2009.3565.1>.

Bodenreider, O. and Zhang, S. (2006). Comparing the Representation of Anatomy in the FMA and SNOMED CT. *AMIA Annu Symp Proc*, pages 46–50.

Collins English Dictionary (2010). *10th edition*. Harpercollins, UK.

Cornet, R. and de Keizer, N. (2008). Forty years of SNOMED: a literature review. *BMC Med Inform Decis Mak*, **8 Suppl 1**, S2.

Crusio, W. E. (2002). 'My mouse has no phenotype'. *Genes Brain Behav*, **1**(2), 71.

Dahdul, W. M., Lundberg, J. G., Midford, P. E., Balhoff, J. P., *et al.* (2010). The Teleost Anatomy Ontology: Anatomical Representation for the Genomics Age. *Syst Biol*, **59**(4), 369–383.

Geldermann, H. (1975). Investigations on inheritance of quantitative characters in animals by gene markers. *Theoretical and Applied Genetics*, **46**, 319–330. 10.1007/BF00281673.

Ghazvinian, A., Noy, N. F., and Musen, M. A. (2011). How orthogonal are the OBO Foundry ontologies? *J Biomed Semantics*, **2 Suppl 2**, S2.

Gkoutos, G. V., Green, E. C. J., Mallon, A.-M., Hancock, J. M., and Davidson, D. (2005). Using ontologies to describe mouse phenotypes. *Genome Biol*, **6**(1), R8.

Gruber, T. R. (1993). Towards Principles for the Design of Ontologies Used for Knowledge Sharing. In N. Guarino and R. Poli, editors, *Formal Ontology in Conceptual Analysis and Knowledge Representation*, Deventer, The Netherlands. Kluwer Academic Publishers.

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>.

Haendel, M. A., Neuhaus, F., Osumi-Sutherland, D., Mabee, P. M., *et al.* (2008). CARO – The Common Anatomy Reference Ontology. In A. Dress, M. Vingron, G. Myers, R. Giegerich, *et al.*, editors, *Anatomy Ontologies for Bioinformatics*, volume 6 of *Computational Biology*, pages 327–349. Springer London.

Hayamizu, T. F., Mangan, M., Corradi, J. P., Kadin, J. A., and Ringwald, M. (2005). The Adult Mouse Anatomical Dictionary: a tool for annotating and integrating data. *Genome Biol*, **6**(3), R29.

Hughes, L. M., Bao, J., Hu, Z.-L., Honavar, V., and Reecy, J. M. (2008). Animal trait ontology: The importance and usefulness of a unified trait vocabulary for animal species. *J Anim Sci*, **86**(6), 1485–1491.

Nadkarni, P. M., Marenco, L., Chen, R., Skoufos, E., *et al.* (1999). Organization of heterogeneous scientific data using the EAV/CR representation. *J Am Med Inform Assoc*, **6**(6), 478–493.

Resnik, P. (1999). Semantic Similarity in a Taxonomy: An Information-Based Measure and its Application to Problems of Ambiguity in Natural Language. *J. Artif. Intell. Res. (JAIR)*, **11**, 95–130.

Robinson, P. N., Köhler, S., Bauer, S., Seelow, D., *et al.* (2008). The Human Phenotype Ontology: A Tool for Annotating and Analyzing Human Hereditary Disease. *Am J Hum Genet*, **83**(5), 610–615.

Rosse, C. and Mejino, J. L. V. (2003). A reference ontology for biomedical informatics: the Foundational Model of Anatomy. *J Biomed Inform*, **36**(6), 478–500.

Smith, B., Ashburner, M., Rosse, C., Bard, J., *et al.* (2007). The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration. *Nat Biotechnol*, **25**(11), 1251–1255.

Smith, C. L. and Eppig, J. T. (2009). The Mammalian Phenotype Ontology: enabling robust annotation and comparative analysis. *Wiley Interdiscip Rev Syst Biol Med*, **1**(3), 390–399.

Vanhonacker, F., Verbeke, W., Poucke, E. V., and Tuyttens, F. A. (2008). Do citizens and farmers interpret the concept of farm animal welfare differently? *Livestock Science*, **116**(1-3), 126 – 136.

Wang, A. Y., Sable, J. H., and Spackman, K. A. (2002). The SNOMED Clinical Terms Development Process: Refinement and Analysis of Content. *Proc AMIA Symp*, pages 845–849.

Zimmerman, K. L., Wilcke, J. R., Robertson, J. L., Feldman, B. F., *et al.* (2005). SNOMED representation of explanatory knowledge in veterinary clinical pathology. *Vet Clin Pathol*, **34**(1), 7–16.

Ontology-based cross-species integration and analysis of *Saccharomyces cerevisiae* phenotypes

Georgios V. Gkoutos and Robert Hoehndorf

Department of Genetics, University of Cambridge, Downing Street, Cambridge, Cambridge CB2 3EH, UK

ABSTRACT

Ontologies are widely used in the biomedical community for annotation and integration of databases. Formal definitions can relate classes from different ontologies and thereby integrate data across different levels of granularity, domains and species. We have applied this methodology to the Ascomycete Phenotype Ontology (APO), enabling the reuse of various orthogonal ontologies and we have converted the phenotype associated data found in the SGD following our proposed patterns. We have integrated the resulting data to a cross-species phenotype network termed PhenomeNET and we make both the cross-species integration of yeast phenotypes and a similarity-based comparison of yeast phenotypes across species available in the PhenomeBrowser.

1 INTRODUCTION

Yeast phenotypes have been proven useful for investigating and revealing various aspects of cellular physiology and mechanisms. The study of these phenotypes has direct implications for understanding mammalian physiology in the context of pharmacodynamics and pharmacokinetics studies, in understanding signalling and regulatory networks, in studies that focus on the identification of response regulators, activators and inhibitors, and in chemical genetics [18, 17, 2, 30]. It is therefore essential that efficient ways are set in place to collect and analyse yeast phenotype data as well as compare them with other organism phenotypes held in a variety of resources.

Over the last years, a plethora of phenotype ontologies has been proposed [26, 22, 27, 29, 24, 20, 6]. These ontologies are developed by a variety of biomedical communities and aim to support the annotation of phenotypic observations derived either from the literature or from experimental studies, including large scale phenotype studies [3, 23]. To unify the species-specific efforts in representing phenotypes, to enable the integration of phenotype information across species, and to enhance the formally represented genotype-to-phenotype knowledge, the species and domain independent Entity-Quality (EQ) method for decomposing phenotypes was developed based on the Phenotype And Trait Ontology (PATO) [9]. According to the EQ method, a phenotype can be decomposed into an entity that is affected by a phenotype and a quality that specifies *how* the entity is affected. The EQ method has been successfully applied both for the direct annotation of species-specific phenotypes and for defining classes in species-specific phenotype ontologies to enable cross-species phenotype integration [10, 19, 11, 28].

The *Saccharomyces* Genome Database (SGD)[4] collects and curates yeast-related phenotype data using the yeast-specific Ascomycete Phenotype Ontology (APO) [7]. Here, we report our efforts to apply the EQ-based method to the APO and enable the reuse of biomedical reference ontologies to describe yeast-related phenotype information as well as integrate it with other species. We apply the results of our analysis to the cross-species phenotype network PhenomeNET [15] and make both the cross-species integration of yeast phenotypes and a similarity-based comparison of yeast phenotypes across species available in the PhenomeBrowser [14].

2 MATERIALS AND METHODS

2.1 *Saccharomyces* Genome Database

The *Saccharomyces* Genome Database (SGD) is a freely available collection of genetic and molecular information about *Saccharomyces cerevisiae*. The SGD contains, amongst others, sequence information for yeast genes and proteins as well as tools for their analyses and comparison, descriptions of their biological roles and molecular functions, the subcellular location at which proteins are active, literature information and links to external resources [4].

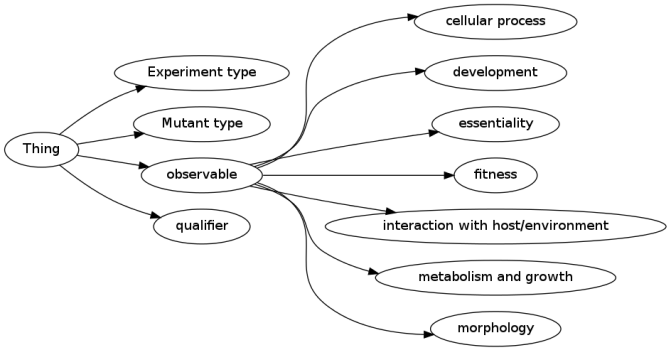
In particular, SGD contains information about phenotypes that arise from curation of either the published scientific literature of traditional bench experiments or from the results of a number of large-scale studies [7]. Such information can be useful for revealing new molecular functional information of genes and SGD curators currently focus on its integration with the available genetic information [4]. The phenotype information recorded includes developmental, metabolism and growth related, processual and morphological manifestations at the cellular level [7].

2.2 Annotating phenotypes using the Ascomycete Phenotype Ontology

The curation of yeast phenotype information is based on a combination of multiple controlled vocabularies which are available from the OBO Foundry ontology repository [25]. One of these vocabularies is the Ascomycete Phenotype Ontology (APO) that, as of 30/06/2011, contains 269 terms organised in four hierarchies [7]. Sub-classes of *Experiment type* provide a classification of genetic interactions and types of experiments (assays) performed on yeast. The class *Mutant type* has sub-classes that provide a classification of types of mutations in yeast that may cause a specific phenotype. Finally, the *observable* and *qualifier* classes are used to record the actual phenotypic observation [7]. The top-level classes of the APO are shown in Figure 1.

According to APO, the *observable* class corresponds to the feature or the trait of a phenotype. For example, traits that can be sub-classes of the *observable* class include the *shape* or *size* of a cell or the *rate* of a growth. These sub-classes are distinguished based on the entity that is affected in a phenotype manifestation

Fig. 1. Top-level of the Ascomycete Phenotype Ontology



and based on the *trait* that is affected. For example, classification based on the entity yields *cellular process*, *cell metabolism* and *cellular growth*, while the classification based on traits results in sub-classes such as *cell morphology*. The APO's *qualifier* class, on the other hand, provides a set of possible comparative values for these traits. For example, *increased*, *arrested* and *abnormal* are included as sub-classes of APO's *qualifier* class. In order to annotate a phenotype corresponding to the observation of *abnormal cell shape*, the APO class *cell shape* (APO:0000051) (a sub-class of *observable*) is combined with the APO class *abnormal* (APO:0000002) (a sub-class of *qualifier*). APO terms can further be used in conjunction with further ontologies, in particular the Chemical Entities of Biological Interest (ChEBI) ontology [5] to extent their ability to describe phenotypes.

3 RESULTS

To formally decompose APO's phenotype classes based on the EQ method and enable the integration of yeast phenotype annotations with phenotype annotations from other species, we have used the PATO [9] and the Gene Ontology (GO) [1] as well as ChEBI [5]. We apply different definition patterns for the different sub-classes of APO's *observable*.

3.1 Morphological traits

APO morphological characteristics are applicable to the morphology of either cellular or sub-cellular structures. We have used the class *Morphology* (PATO:0000051) and its subclasses, and we link them to the appropriate anatomical localisation provided by GO's cellular component branch. For example, to define the APO term *Cell wall morphology* (APO:0000053), the GO cellular anatomical term *Cell wall* (GO:0005618) is linked to the *Morphology* (PATO:0000051) term from the PATO ontology.

We implement this EQ-based definition in the OBO Flatfile Format [16] following the syntactic patterns associated with EQ [21]. In the OBO Flatfile Format, the definition can be expressed as follows:

```
[Term]
id: APO:0000053 ! cell wall morphology
intersection_of: PATO:0000051 ! morphology
intersection_of: inheres_in GO:0005618
```

Formally, we use the conversion approach used in the PhenomeBLAST software [14] to represent this syntactic description of a phenotype in OWL. PhenomeBLAST applies a simplified form of the phene-patterns [13], and the *Cell wall morphology* phenotype would be represented as a phenotype of entities that have a cell wall as part in which a quality of the type *Morphology* inheres:

```
APO:0000053 EquivalentTo: phenotype-of some
  (has-part some (GO:0005618 and
   has-quality some PATO:0000051))
```

In some cases, the APO terms are related to temporal stages, i.e., the phenotypes are observed only while the yeast cell is in a certain stage. For example, stages of the cell cycle are used in classes such as *Critical cell size at G2/M (cryptic G2/M cell size checkpoint)* (APO:0000142). To define a class involving reference to a temporal stage, we use the **during** relation and a class from the GO. In the OBO Flatfile Format, the class *Critical cell size at G2/M (cryptic G2/M cell size checkpoint)* is defined as follows:

```
[Term]
id: APO:0000142
intersection_of: PATO:0000117 ! size
intersection_of: inheres_in GO:0005623
intersection_of: during GO:0031576
```

Formally, this phenotype is translated into the OWL definition:

```
APO:0000142 EquivalentTo: phenotype-of some
  (has-part some (GO:0005623 and
   has-quality some PATO:0000117 and
   during some GO:0031576))
```

3.2 Developmental, metabolic and physiological phenotypes

The APO contains the classes *Cellular process*, *Development*, *Metabolism and growth* as well as *Interaction with host/environment*. We assume that each of these classes represents a phenotype that is based on a process. In particular, we use GO's classification of processes to define the APO class *Cellular process* (APO:0000066) as a phenotype of a *Cellular process* (GO:0009987), *Development* (APO:0000023) as a phenotype of a *Cellular developmental process* (GO:0048869) and *Metabolism and growth* (APO:0000094) as a phenotype of either *Cellular metabolic process* (GO:0044237) or *Cellular growth* (GO:0016049). To obtain additional inferences based on the parthood relations in the GO, we use definition patterns that include the **part-of** relation. For example, we formally define *Cellular process* as:

```
APO:0000066 EquivalentTo: phenotype-of some
  (has-part some (part-of some
    GO:0009987 and has-quality some
    PATO:0000001))
```

This definition pattern uses the **has-part** relation to relate an organism (the range of **phenotype-of**) to a process. We do not use the **participates-in** relation for this purpose, since explicitly distinguishing between processes and material objects will currently lead to contradictions in phenotype ontologies and the GO [12]. In the future, we intend to explicitly incorporate more expressive phenotype definition patterns that enable interoperability between ontologies of both anatomy and physiology [13].

To define APO classes that describe phenotypes associated with biological processes or molecular functions, we linked the appropriate GO classes with terms from PATO. The classification

of biological processes or molecular functions in the GO provide the entity affected by a phenotype while PATO characterizes *how* these entities are affected.

As a consequence of defining the sub-classes of *observable* in APO based on the GO using the **part-of** relation, we can infer a new and updated taxonomic structure of APO in which *Development* and *Metabolism and growth* are sub-classes of *Cellular process*. This inference is obtained through inference over GO's classification of processes and the definition patterns we provide.

3.3 Dispositional phenotypes

A common kind of phenotypes in yeast include dispositions to interact with other substances in a particular way. For example, the APO class *Metal resistant* (APO:0000090) is used to describe yeast's disposition to interact with metal.

In the EQ-based decomposition of the class *Metal resistant*, we use GO's process class *Response to metal ion* (GO:0010038) and combine it with the PATO class *Sensitivity of a process* (PATO:0001457):

```
[Term]
id:  APO:0000090
intersection_of: PATO:0001457
intersection_of: inheres_in GO:0010038
```

Similar to processual phenotypes, we do not yet use the **has-disposition** or **has-function** relation in formalizing this phenotype because formally distinguishing between functions and processes will lead to a large number of unsatisfiable class in phenotype ontologies and the GO. Consequently, we formally define *Metal resistant* as:

```
APO:0000090 EquivalentTo: phenotype-of some
  (has-part some (GO:0010038 and
   has-quality some PATO:0001457))
```

In the future, we intend to formalize dispositional phenotypes using the **has-disposition** or **has-function** relation.

3.4 Interoperability with chemistry ontology

Relational classes from the PATO ontology can also be used to characterize qualities of more than one entity. We use the **towards** relation to specify the second argument of a relational quality. For example, we define the APO term *Resistance to chemicals* (APO:0000087) by linking the class *Chemical compound* (CHEBI:37577) to the PATO class *Sensitivity of a process* (PATO:0001457) and the process class *Response to chemical stimulus* (GO:0042221):

```
[Term]
id: APO:0000087
intersection_of: PATO:0001457
intersection_of: inheres_in GO:0042221
intersection_of: towards CHEBI:37577
```

Formally, we express this statement as

```
APO:0000087 EquivalentTo: phenotype-of some
  (GO:0042221 and
   has-quality some (PATO:0001457 and
    towards some CHEBI:37577))
```

3.5 Phenotypic qualifiers

To relate APO's qualifier-classes to the PATO ontology, we created a statement of equivalency between PATO's qualifier classes and APO's qualifier classes. For example, for the APO term *arrested* (APO:0000250), we created an equivalent-class statement to the PATO term *arrested* (PATO:0000297).

Since PATO formally distinguishes between qualities that inhere in objects and qualities that inhere in processes such statements also allowed for reasoners to automatically check the consistency of the combination of qualifiers with anatomical or processual terms created by curators for annotation purposes.

3.6 Formalizing yeast phenotype annotations

The SGD makes phenotype annotations for specific genotypes and genetic interactions available. These annotations consist of a genotype identifier (such as S000029075) and either a pair or a triple of classes which describe the phenotype that is associated with the genotype. If the phenotype annotation consists of a pair of classes, a class from the APO's *observable* branch is combined with a class from the APO's *qualifier* branch. For example, the genotype S000029075, a conditional mutation of the *CDC29* gene, has three phenotype annotations in the SGD:

- heat sensitivity (APO:0000147): increased (APO:0000004)
- budding (APO:0000024): absent (APO:0000005)
- cell cycle progression (APO:0000253): arrested (APO:0000250)

To formalize these phenotypes, we first identify the entity and the quality that is affected in a phenotype. For example, *Heat sensitivity* (APO:0000147) is defined as a phenotype of a *Response to heat* (GO:0009408) process and is based on the PATO quality *Sensitivity of a process* (PATO:0001457). Based on this information, we create an OWL class expression. Since the qualifier that is applied to *Heat sensitivity* (APO:0000147) is *Increased* (APO:0000004) and the quality *Sensitivity of a process* (PATO:0001457), we construct an anonymous *Increased sensitivity of a process* class using the **increased-in-magnitude-relative-to** (similarly to PATO's definition of the *Increased sensitivity of a process* class) (PATO:0001551)and formalize *Heat sensitivity: increased* as:

```
phenotype-of some (has-part some
  GO:0009408 and has-quality some
  (PATO:0001457 and
   increased-in-magnitude-relative-to some
   normal))
```

Based on this information, the phenotype description will be inferred to be a sub-class of APO's *Heat sensitivity*, it will inter-operate with phenotypes that are based on PATO's *Increased sensitivity of a process* class (because they share the same definition) and through inference over the GO we can obtain basic interoperability across multiple species' phenotype descriptions.

We formalize the phenotype "cell cycle progression: arrested" using the PATO term *Arrested* (PATO:0000297) and the GO process class *Cell cycle process* (GO:0022402):

```
phenotype-of some (has-part some
  GO:0022402 and has-quality some
  PATO:0000297)
```

We formalize the remaining phenotype description of S000029075 in a similar way and combine the individual phenotype classes using class intersection.

Phenotype descriptions based on a triple consist of an entity, a qualifier and a second entity that is used to define the respective phenotype class. For example, S000000649 is annotated with *Ionic stress resistance: decreased* and the additional class *Sodium chloride* (CHEBI:26710). The intended meaning of this phenotype description is that the resistance of the yeast cell to respond to sodium chloride is decreased within the specific experiment that was performed. To formalize this phenotype, we combine the PATO class *Sensitivity of a process* (PATO:0001457), the GO class *Response to chemical stimulus* (GO:0042221) and the ChEBI class *Sodium chloride* (CHEBI:26710):

```
phenotype-of some (has-part some
  GO:0042221 and has-quality some
  (PATO:0001457 and towards some
   CHEBI:26710))
```

3.7 Cross-species phenotype integration

Many of the definitions we propose do not make full use of established phenotype definition patterns that enable interoperability with ontologies of functions and processes [13]. However, our prime motivation in defining yeast phenotypes was to enable cross-species phenotype integration and comparison using the PhenomeBLAST and PhenomeNET methods. We have formally integrated the APO and the definitions of the APO that we created with the ontology underlying PhenomeBLAST (the software and ontology are available from http://phenomeblast.googlecode.com), and we can represent yeast phenotypes using the phenotype ontologies that were created for other species. For example, the phenotypes of S000029048 (annotated with the single phenotype *Autophagy: absent*) expressed using the Mammalian Phenotype Ontology (MP) are *Abnormal metabolism*, *Homeostasis/metabolism phenotype* and *Mammalian phenotype*. Using the Worm Phenotype Ontology (WPO), which targets an organism that is more similar to yeast than mammals, we obtain as phenotypes abnormalities of *Autophagy*, *Intracellular transport*, *Small molecule transport* and *Cellular processes*.

4 CONCLUSION

In the future, we intend to evaluate and quantify the potential of yeast phenotype annotations to predict orthologous genes and genes involved in metabolic diseases based on comparisons of phenotypes. Furthermore, as cross-species phenotype integration progresses, we intend to update the definitions to accurately reflect more complex relations.

In the post-genomic era, the analysis and integration of phenotype data have been demonstrated as useful tools assigning genotype to phenotype correlations, providing insights in the nature of human disease and ultimately discovering novel therapeutic approaches. The challenge now remains to provide mechanisms and methods that allow such integration and analysis on a large scale that takes into account the vast amount of phenotypic information collected around the world for various species in a single framework. One such framework has been proposed based on the use of PATO and a variety of external ontologies [8] and has been successfully demonstrated to work for achieving such integration [10, 19, 11, 21].

Here we demonstrated how yeast phenotype information could be defined based on this framework and we have successfully included yeast phenotype data in a cross species phenotype data network. As such yeast phenotype data can be integrate and manalysed with data from other species and increases their potential for discovering new genotype to phenotype correlations.

REFERENCES

[1]Michael Ashburner, Catherine A. Ball, Judith A. Blake, David Botstein, Heather Butler, Michael J. Cherry, Allan P. Davis, Kara Dolinski, Selina S. Dwight, Janan T. Eppig, Midori A. Harris, David P. Hill, Laurie I. Tarver, Andrew Kasarskis, Suzanna Lewis, John C. Matese, Joel E. Richardson, Martin Ringwald, Gerald M. Rubin, and Gavin Sherlock. Gene ontology: tool for the unification of biology. *Nature Genetics*, 25(1), May 2000.

[2]Graham Bell. Experimental genomics of fitness in yeast. *Proceedings. Biological sciences / The Royal Society*, 277(1687):1459–1467, May 2010.

[3]S. D. Brown, P. Chambon, M. H. de Angelis, and Eumorphia Consortium. EMPReSS: standardized phenotype screens for functional annotation of the mouse genome. *Nat Genet*, 37(11):1155, 2005.

[4]J. M. Cherry, C. Adler, C. Ball, S. A. Chervitz, S. S. Dwight, E. T. Hester, Y. Jia, G. Juvik, T. Roe, M. Schroeder, S. Weng, and D. Botstein. SGD: Saccharomyces genome database. *Nucleic acids research*, 26(1):73–79, January 1998.

[5]K. Degtyarenko, P. Matos, M. Ennis, J. Hastings, M. Zbinden, A. McNaught, R. Alcantara, M. Darsow, M. Guedj, and M. Ashburner. ChEBI: a database and ontology for chemical entities of biological interest. *Nucleic Acids Research*, 2007.

[6]R. Drysdale. Phenotypic data in FlyBase. *Brief Bioinform*, 2(1):68–80, 2001.

[7]Stacia R. Engel, Rama Balakrishnan, Gail Binkley, Karen R. Christie, Maria C. Costanzo, Selina S. Dwight, Dianna G. Fisk, Jodi E. Hirschman, Benjamin C. Hitz, Eurie L. Hong, Cynthia J. Krieger, Michael S. Livstone, Stuart R. Miyasato, Robert Nash, Rose Oughtred, Julie Park, Marek S. Skrzypek, Shuai Weng, Edith D. Wong, Kara Dolinski, David Botstein, and J. Michael Cherry. Saccharomyces genome database provides mutant phenotype data. *Nucleic acids research*, 38(Database issue), January 2010.

[8]G. V. Gkoutos, E. C. J. Green, A. M. Mallon, J. M. Hancock, and D. Davidson. Building mouse phenotype ontologies. In Russ B. Altman, Keith A. Dunker, Lawrence Hunter, Tiffany A. Jung, and Teri E. Klein, editors, *Proceedings of the 9th Pacific Symposium on Biocomputing (PSB 2004), Hawaii, USA, Jan 6-10*, London, 2004. World Scientific.

[9]Georgios V. Gkoutos, Eain C. Green, Ann-Marie M. Mallon, John M. Hancock, and Duncan Davidson. Using ontologies to describe mouse phenotypes. *Genome biology*, 6(1), 2005.

[10]Georgios V. Gkoutos, Chris Mungall, Sandra Dolken, Michael Ashburner, Suzanna Lewis, John Hancock, Paul Schofield, Sebastian Kohler, and Peter N. Robinson. Entity/quality-based logical definitions for the human skeletal phenome using PATO. *Annual International Conference of the IEEE Engineering in Medicine and Biology Society.*, 1:7069–7072, 2009.

[11]John Hancock, Ann-Marie Mallon, Tim Beck, Georgios Gkoutos, Chris Mungall, and Paul Schofield. Mouse, man, and meaning: bridging the semantics of mouse phenotype and human disease. *Mammalian Genome*, 20(8):457–461, August 2009.

[12]Robert Hoehndorf, Michel Dumontier, Anika Oellrich, Dietrich Rebholz-Schuhmann, Paul N. Schofield, and Georgios V. Gkoutos. Interoperability between biomedical ontologies through relation expansion, upper-level ontologies and automatic reasoning. *PLOS ONE*, 6(7):e22006, July 2011.

[13]Robert Hoehndorf, Anika Oellrich, and Dietrich Rebholz-Schuhmann. Interoperability between phenotype and anatomy ontologies. *Bioinformatics*, 26(24):3112 – 3118, 10 2010.

[14]Robert Hoehndorf, Paul N. Schofield, and Georgios V. Gkoutos. Phenomebrowser. http://phenomebrowser.net, 2011.

[15]Robert Hoehndorf, Paul N. Schofield, and Georgios V. Gkoutos. Phenomenet: a whole-phenome approach to disease gene discovery. *Nucleic Acids Research*, 2011.

[16]Ian Horrocks. OBO flat file format syntax and semantics and mapping to OWL Web Ontology Language. Technical report, University of Manchester, March 2007. http://www.cs.man.ac.uk/~horrocks/obo/.

[17]Youn-Sig Kwak, Sangjo Han, Linda S Thomashow, Jennifer T Rice, Timothy C Paulitz, Dongsup Kim, and David M Weller. A saccharomyces cerevisiae genome-wide mutant screen for sensitivity to 2,4-diacetylphloroglucinol, an antibiotic

produced by pseudomonas fluorescens. *Appl Environ Microbiol*, 2010.

[18]Alex Lan, Ilan Y Smoly, Guy Rapaport, Susan Lindquist, Ernest Fraenkel, and Esti Yeger-Lotem. Responsenet: revealing signaling and regulatory networks linking genetic and transcriptomic screening data. *Nucleic Acids Res*, 2011.

[19]Paula M. Mabee, Michael Ashburner, Quentin Cronk, Georgios V. Gkoutos, Melissa Haendel, Erik Segerdell, Chris Mungall, and Monte Westerfield. Phenotype ontologies: the bridge between genomics and evolution. *Trends in ecology & evolution (Personal edition)*, 22(7):345–350, July 2007.

[20]Hiroshi Masuya, Yuko Makita, Norio Kobayashi, Koro Nishikata, Yuko Yoshida, Yoshiki Mochizuki, Koji Doi, Terue Takatsuki, Kazunori Waki, Nobuhiko Tanaka, Manabu Ishii, Akihiro Matsushima, Satoshi Takahashi, Atsushi Hijikata, Kouji Kozaki, Teiichi Furuichi, Hideya Kawaji, Shigeharu Wakana, Yukio Nakamura, Atsushi Yoshiki, Takehide Murata, Kaoru Fukami-Kobayashi, Sujatha Mohan, Osamu Ohara, Yoshihide Hayashizaki, Riichiro Mizoguchi, Yuichi Obata, and Tetsuro Toyoda. The RIKEN integrated database of mammals. *Nucleic Acids Research*, 39(suppl 1):D861–D870, January 2011.

[21]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.

[22]P. N. Robinson, S. Koehler, S. Bauer, D. Seelow, D. Horn, and S. Mundlos. The human phenotype ontology: a tool for annotating and analyzing human hereditary disease. *American journal of human genetics*, 83(5):610–615, 2008.

[23]Nadia Rosenthal and Steve Brown. The mouse ascending: perspectives for human-disease models. *Nature Cell Biology*, 9:993 – 999, 2007.

[24]Gary Schindelman, Jolene Fernandes, Carol Bastiani, Karen Yook, and Paul Sternberg. Worm phenotype ontology: integrating phenotype data within and beyond the c. elegans community. *BMC Bioinformatics*, 12(1):32, 2011.

[25]Barry Smith, Michael Ashburner, Cornelius Rosse, Jonathan Bard, William Bug, Werner Ceusters, Louis J. Goldberg, Karen Eilbeck, Amelia Ireland, Christopher J. Mungall, Neocles Leontis, Philippe R. Serra, Alan Ruttenberg, Susanna A. Sansone, Richard H. Scheuermann, Nigam Shah, Patricia L. Whetzel, and Suzanna Lewis. The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration. *Nat Biotech*, 25(11):1251–1255, 2007.

[26]Cynthia L. Smith, Carroll-Ann W. Goldsmith, and Janan T. Eppig. The mammalian phenotype ontology as a tool for annotating, analyzing and comparing phenotypic information. *Genome Biology*, 6(1):R7, 2004.

[27]Judy Sprague, Leyla Bayraktaroglu, Yvonne Bradford, Tom Conlin, Nathan Dunn, David Fashena, Ken Frazer, Melissa Haendel, Douglas G. Howe, Jonathan Knight, Prita Mani, Sierra A. Moxon, Christian Pich, Sridhar Ramachandran, Kevin Schaper, Erik Segerdell, Xiang Shao, Amy Singer, Peiran Song, Brock Sprunger, Ceri E. Van Slyke, and Monte Westerfield. The zebrafish information network: the zebrafish model organism database provides expanded support for genotypes and phenotypes. *Nucl. Acids Res.*, pages gkm956+, November 2007.

[28]Nicole L. Washington, Melissa A. Haendel, Christopher J. Mungall, Michael Ashburner, Monte Westerfield, and Suzanna E. Lewis. Linking human diseases to animal models using ontology-based phenotype annotation. *PLoS Biol*, 7(11):e1000247, 11 2009.

[29]Yukiko Yamazaki and Pankaj Jaiswal. Biological ontologies in rice databases. an introduction to the activities in gramene and oryzabase. *Plant Cell Physiol*, 46(1), 2005.

[30]Zhun Yan, Nicolas M Berbenetz, Guri Giaever, and Corey Nislow. Precise gene-dose alleles for chemical genetics. *Genetics*, 182(2):623–6, 2009.

The ontology of biological mechanisms

*Johannes Röhl**

University of Rostock, Rostock

ABSTRACT

Motivation: The concept of a mechanism has become a standard proposal for explanations in biology. In systems biology it is a desideratum to match established or hypothetical mechanisms with mathematical models of biosystems. Therefore a formal ontological description of mechanisms is desirable. Taking some hints from an "ontology of devices" I suggest as a general approach for this task the introduction of functional kinds and functional parts by which the particular relations between a mechanism and its components can be captured.

1 INTRODUCTION

The concept of a mechanism has in recent years become a standard philosophical proposal for explanations in biology and other sciences of complex systems where the traditional approach, subsumption under universal laws, has not been fruitful. This is in agreement with the practice of these sciences where the postulation of mechanisms on several levels (organismic, cellular, molecular, biochemical) is a common research practice: A stable behavior of some biological system is explained by the description of a (postulated) mechanism that is causally responsible for this behavior. Familiar examples include photosynthesis or proteinbiosynthesis; diseases are associated with mechanisms as well as the action of drugs (cf. Campaner 2011). The discovery of such a postulated mechanism is of course non-trivial and often a seminal scientific achievement. Whereas mechanistic explanations in biology have usually been mostly qualitative, Systems Biology is working with powerful mathematical tools and striving for quantitative results. Scientists working in the field have expressed the hope to get a better picture if the calculational approach could be aligned with mechanistic approaches (Boogerd et al. 2007, 326). The following considerations hope to give some first steps in this direction by an ontological analysis of mechanisms. Though the work presented here is conceptual groundwork, there is a rich field of possible applications in biomedical knowledge representation and knowledge eliciting. Ontology should answer the question, into which fundamental category something falls. To find out what a mechanism is, I will start with some definitions from the literature, discuss options for categories for the components of mechanisms and the mechanism as a whole, and suggest with which relations they could be tied together to constitute a mechanism. It should be noted beforehand, however, that the suggestions in this paper are not meant to necessarily imply a universal mechanistic and reductionist world view. It will be made clear in the following that mechanism in the sense used here is dependent on *functional* concepts and thus not reductionist in the sense criticized by Robert Rosen (Rosen 1991, 1998) and others. Furthermore, the claim that organisms contain (many) mechanisms is not meant to imply that all organisms are to be identified with mechanisms: An ontological analysis of mechanisms does not imply a universal mechanistic ontology.

2 WHAT ARE MECHANISMS?

2.1 Common definitions of a mechanism

Here are three different definitions from the literature on mechanisms:

"A mechanism for a behavior is a complex system that produces that behavior by the interaction of a number of parts, where the interactions between parts can be characterized by direct, change-relating generalizations." (Glennan 2002, 344)

"Mechanisms are entities and activities organized such that they are productive of regular changes from start or set-up to finish or terminating conditions."(Machamer et al. 2000, 2)

"A mechanism is a structure performing a function in virtue of its component parts, component operations, and their organization. The orchestrated functioning of the mechanism is responsible for one or more phenomena" (Bechtel/Abrahamsen 2005, 423)

All three conceptions identify several features which are important for the characterization of mechanisms: Mechanisms are (1) for a specific behavior, which can be characterized functionally by a specific input and output; they are (2) not stating mere input-output-correlations, but show the "inner workings" producing the output, they exhibit (3) a kind of continuity and lead without gaps from initial conditions to final states, and they are (4) complex and may, in general, have submechanisms at a lower level of granularity.

A problem for the first two definitions are cyclical mechanisms that cannot be simply characterized by initial or final states (cf. Bechtel 2006). But one can think of the

function performed by a mechanism in a more general way: A mechanism for cyclically maintaining some balance of concentrations, for example, will take as input any state of the system to be controlled, and if its value of the concentration ratio is outside some tolerance interval, the mechanism will produce as output a state with a value within the interval.

The key features seem to be that mechanisms are on the one hand functionally characterized, and on the other hand they are complex systems and the consecutive actions of their components together realize the function of the whole.

Is there a distinction from biochemical "pathways"? The usage in the scientific community seems to be not entirely clear. One difference seems to be that pathways are mostly chains of chemical reactions and transport phenomena involving free floating biomolecules, whereas mechanisms are more clearly localized and involve stable material structures, as in e.g. neurological mechanisms.

2.2 Components of a mechanism

According to the definitions above mechanisms have two types of components: "entities" and "interactions or "activities". What is meant with „entities" in the definitions above are stable material parts. Working with the top level ontology BFO, these can be classified as independent continuants. In BFO (cf. Spear 2006) independent continuant entities are what we would think of as "things", i.e., entities that can change in time, but have no temporal parts. In contradistinction, occurrents are entities like events or processes that take place in time and have temporal parts (phases or stages). Occurrents are dependent entities, they always need at least one independent continuant as participant. These components of a mechanism are identified functionally according to their contribution for the behavior characteristic for that mechanism as a whole. So they are not simply parts of the mechanism, but the recognizing of some parts as components is dependent on context. E.g, the „pumped" H^+-ions (protons) or the electrons transported by the electron transport chain in photosynthesis are components of the respective mechanisms, whereas the protons and electrons of the atoms of the molecules that do not change during the action of the mechanism are not involved and therefore not components, although parts. What about the other type of components?

2.3 Activities and interactions

Obviously the material components of a mechanism must be dynamically connected. There is a dispute how this should be captured by the ontological analysis. One option is to understand the interactions as relations between the continuants as proposed by Stuart Glennan: An "interaction" is taken to be a "correlative property change": "an occasion on which a change in a property of one part brings about a change in a property of another part" (Glennan 2002, 344f.). On the other hand one can argue that to represent the dynamics of a mechanism it is not enough to use only the material components and their relations, but opt for the acceptance of "activities" as a separate class of entities (cf. Machamer/Darden/Craver 2000). There are considerable advantages to an approach that accepts not only continuants and their relations, but also another type of entities, namely processes or occurrents. The continuants involved in the interaction are then connected to these occurrents by a relation like "participates_in" that express this involvement of one or more continuants in a process. And now one can speak about the location, frequency and other features of these interactions or about their preceding or following each other and define corresponding relations. (Cf. Schulz/Jansen 2009 for detailed arguments for the usefulness of admitting occurrents as a distinct category for representing molecular interactions). Note that in the Gene Ontology the term "activity" is used to describe molecular *functions*, like "catalytic activity". In the GO this means that the specific component has the 'function' to initiate or take part in this activity. I will from now call all interactions "processes" (and ignore the distinction between processes that involve only one participant and are therefore not interactions in the strict sense). Before discussing the relations between the material components and processes of a mechanism, I will turn to the question how the mechanism as a whole should be ontologically classified.

2.4 The device ontology

Galton and Mizoguchi have suggested a very general framework to analyze the mutual dependence of continuants, functions and processes (Galton/Mizoguchi 2009, 86f.). Starting from the analysis of engineering devices in their „ontology of devices" not only artifacts, but almost any entity can be considered as a „device" in the following general sense: A device is an object that is essentially characterized (functionally) by the input/output it can produce (apart from that it is a "black box"). But to generate that output the object has to undergo or "enact" a process: "an object […] is characterized in terms of the processes it enacts. These are what we call the external processes or behaviour of the object." (Galton/Mizoguchi 2009, 94) We thus have a tightly interrelated triple of <object, function, process>: The object „enacts" its "external processes", which produce „the generation of output from input." (ibid.) E.g., in photosynthesis, the object photosystem I has the function to transform the input light energy into the output chemical energy. Looking "inside the black box" we find „subdevices" with processes that are internal to the higher-level device and causally responsible for its external processes. And on any level the identity of an object depends on its function or disposition for external processes: „an object is a unity which is what enacts its

external processes" (Galton/Mizoguchi 2009 98). Such a nested structure of objects essentially characterized by their functions or their ability to enact specific processes corresponds very well to the mechanistic approach in biology as described so far. (The regress of structures (devices within devices) has to be stopped at some basic physical level, but this is not important for the purpose here.)

2.5 The mechanism as a whole

There are conflicting intuitions into which category a mechanism as a whole should belong. Glennan opts for continuant; a mechanisms is a „thing" like a clock or a cell, because of its endurance and stability compared to a process or chain of events (Glennan 2002, 345). Similarly, Bechtel: "it consist of an arrangement of parts and has at least some enduring identity" (Bechtel 2007, 275). On the other hand the mechanism could be a process: Although it does not follow from the fact that a mechanism has processes as „components" that it has to be a process itself, because not only a process can contain processes (cf. the "device ontology"), it seems clear that the unfolding activity of a mechanism has temporal stages characterized by the processes its components enact. And in our top level ontology the only entities with such a temporal structure are occurrents. But what if the mechanism is not active? It seems that the mechanism is still there without any (external) process going on (like the case of a stopped clock that needs to be wound up). So the mechanism cannot be categorized as a process.

If we take the mechanism as a generalized „device" in the sense of Galton and Mizoguchi, we might be able to do justice to both intuitions. We take the mechanism to be an independent continuant. But it is functionally characterized, that means its very identity depends on its function for its specific process (like conversion of light energy to chemical energy in photosynthesis.) The mutual dependency of continuant, specific function and process can be expressed by the ascription of a specific function as essential for the mechanism. Mechanisms are then continuants and classified as functional kinds: The essence of their being is their function. The concept of a function has been widely and controversially discussed in the philosophy of biology and elsewhere (cf. e.g. Krohs/Kroes 2009 for recent contributions). An elaborate ontology of functions has been proposed by (Burek et al. 2006). Both for lack of space and because I do not want to presuppose particular accounts of functions and functional roles, I will rely only on very general features of the concept of a function (apparently shared by Burek et al., although their account contains more detail.)

In the top level ontology BFO functions are "realizable dependent continuants" (cf. Arp/Smith 2005, 2f.). That means a function is dependent (like properties) on the independent continuant that is the bearer of the function. "Realizable" means that the instances of a function type are connected to processes, their realizations. These are processes with the bearer of the function as a participant. A few remarks are in order here: Not all processes involving the bearer of the function are realizations of (one of) its functions, but every realization of the function involves its bearer as participant. And functions do not have to be (always or ever) realized, as e.g. in the case of a safety mechanism the function of which will only be realized if certain conditions obtain (and they may never obtain). I will ignore the complications of this modal feature for now. Dispositions are very similar in this respect and are discussed at some length in (Roehl/Jansen 2011).

To illustrate this with an example from the domain of artifacts: One function of a hammer is to drive in nails, a process token of hammering in one particular nail is a realization token of this function of the hammer, and the hammer is of course a participant in this process. Therefore a necessary condition for being a hammer is that it can (and usually does) participate in nailing processes which are realizations of its function. For a full definition one would have to introduce the intention that the hammer has been designed for its particular function. A nail may be driven in with a stone or a heavy wrench, so these could perform the "hammering role" to some extent. But unlike the hammer hammering is not their essential function. This is different with many biological functions as these are not designed by intentions, but rather by evolution.

Along these lines one can define the class of functional kinds as below. I 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) and **has_participant** (which holds between particular processes and particular continuants), and **part_of**. I also adopt their 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. Not to proliferate relations I will stick to the OBO and BioTop relations when possible.

M *is_a* functional_kind := ∃F(F *is_a* function ∧ M *has_function* F ∧ ∀x∀f (f i**nstance_of** F ∧ x **has_function** f → x **instance_of** M) (1)

The specific function has to be assigned by way of definition to the specific kind in question. The classification of a mechanism as something which essentially has a specific function is necessary, but not enough. For a mechanism like photosynthesis to be what it is, the inner structure of the mechanism is also important, the material parts and their organization as well as the processes going on among them. The explanatory value of a mechanism rest on this: The function of the whole is explained by the

interplay of the functions of the parts. I will now try to capture these relations between a mechanism and its components as well as the relations between the components.

3 FORMAL RELATIONS

3.1 Functional kinds and functional parts

It seems clear that the material components are parts of a mechanism, but the usual part-whole relation is too weak to capture the more specific relation of a mechanism to its components. More sophisticated subrelations taking into account granularity levels (cf. Jansen/Schulz 2011 and the BioTop ontology, Beisswanger et al. 2008) do not help much either. A mechanism is clearly not a (homogeneous) collection of "grains". It is also doubtful whether the notion of a compound with components helps here, because of the "structure blindness" of the compound/component relation. Furthermore, the working parts can belong to different granularity levels: Pumped H^+ ions are (almost) elementary particles whereas the thylakoid membrane is on the level of cellular components. (This "level crossing" seems more the rule than the exception, cf. Richardson/Stephan 2007, 131f.). Mainly the functional role matters. So my suggestion is to introduce a relation "functional_part_of": a component must be a proper part of the whole and it must have a specified function. Similar suggestions have been made by Galton/Mizoguchi (2009, 96f.) and, Vieu/Aurnague (2007), but the latter do not refer to processes explicitly and are more focussed on properties of different parthood relations. But the main point of difference here is that the functional parts for biological mechanisms are not automatically given as such by some previously available description, like a door handle as a functional part of a door. Rather, the explanatory achievement of a mechanism is that the components, some biological entities like molecules, membranes etc. have certain functions that contribute to the function of the whole.

x **func_part_of** z := x **part_of** z ∧ ∃f (f **instance_of** Function ∧ x **has_function** f) (2)

This seems still rather weak as the ascription of functions is to some extent context-dependent. It is as part of the mechanism m_{WHOLE} that m has function x, not necessarily *per se*. For the relation of a functional part m to the mechanism as a whole m_{WHOLE}, we want the following conditions to hold:
(1) m_{WHOLE} is a mechanism, i.e. it is itself an instance of a functional kind with a specific function and realization. (2) m is a part of m_{WHOLE}. (3) The function of the part m contributes to the function of the whole mechanism m_{WHOLE}.
This means that the (internal) realization process p_{int} of the function of the part must contribute to the (external) process p_{ext} enacted by m_{WHOLE}. Galton/Mizoguchi (2009, 96) postulate a relation "p_{int} contributes_to p_{ext}" without specifying details. My suggestion is to use a parthood relation between the internal processes of the functional parts and the external process enacted by the whole. The BioTop ontology contains the relation "has_processual_part" for expressing parthood between processes that seems useful here. Because the same type of component may be a functional part of different types of mechanism (H^+ ions or the cell membrane figure in many mechanisms), but mechanisms have their functional parts as necessary components, I will use the inverse relation **has_part**. Let us first define the contribution relation between functions f, f* indicating that the realization process of f is a part of the realization of f*:

f **contributes_to** f*: ∃r ∃p (f **has_realization** r ∧ f* **has_realization** p ∧ p **has_processual_part** r) (3)

Not that this does not imply that every processual part of the encompassing process is a realization of a contributing function: There may well be "side effects" as the generation of heat, noise etc. (A referee brought this to my attention.) But all realizations of the contributing functions must still be processual parts of the larger process. Still, it is to be hoped that this notion of contribution can be refined in future work.
If an internal process p_{int} is a processual part of the realization process p_{ext} of m_{WHOLE}., than the participant m of p_{int} participate as well in p_{ext} (contra Galton/Mizoguchi). It follows immediately from the definition of functions that m and m_{WHOLE}. participate in the realization processes of their respective functions, so I will not mention that explicitly in the following:

x **has_func_part** z := x **has_part** z ∧
∃f ∃f* ∃r ∃P (z **has_function** f ∧ x **has_function** f* ∧ f* **has_realization** r ∧ r **instance_of** P ∧ P *is_a* process ∧ f **contributes_to** f* ∧ r **has_participant** z) (4)

What is a mechanism? It is (1) a complex continuant ("structured biological entity" from BioTop could be an appropriate class) that has (2) a specified biological function that is essential for it, and that (3) necessarily has a substructure of functional parts with functions that contribute to the function of the whole. A list of all the functional parts will give a further ("internal") specification of the mechanism in addition to the definition by its (external) function.

Most of what has been said so far is not restricted to biological entities. One can distinguish subclasses of "artefactual" and "biological mechanisms" by demanding that the latter are biological entities. The question then arises whether whole organs or even whole organisms

should be considered as mechanisms. In principle they could, but usually mechanisms are supposed to explain a particular function or phenomenon on a sub-organismic level. One could therefore include clauses to exclude whole organs from being mechanisms. It seems straightforward to add such restrictions by defining e.g. "molecular mechanism" as something that has necessarily some biomolecule among its parts, or to demand that the mechanism is part of some organ.

3.2 Relations between material components and processes of the mechanism

So far I focused on the relation of the mechanism as a whole to its components. The continuant components and the processes of a mechanism stand in the relation of participation to each other (cf. Smith et al. 2005): p **has_participant** x. This simply follows from the conception of a function and can usually be generalized to the type level, as many types of processes essentially involve specific types of participant: Electron transport trivially has at least one electron as participant. (Complications arising from the fact that often a collective of molecules participates in a subprocess will be ignored for now. This matters if a certain concentration is necessary for a process to happen (Schulz/Jansen 2009).) With the functional parts analysis this yields the statement that the functional parts of a mechanism are participants of the processes that are the realizations of these functions:

(X *func_part_of* Z ∧ X *has_function* F ∧ R *is_a* Process) → (F *has_realization* R → R *has_participant* X) (5)

A converse statement is usually not true as the tokens of the participants do not always participate in (again obvious for electrons), but could be formulated using a relation relativized to time (cf. Smith et al. 2005): X *sometimes_participates_in* P. The more interesting connections, though, arise from the fact that several components are changed by participating in one internal process and by the connections of the internal processes themselves. Although the temporal order of the internal processes is important, a generalization is difficult, because there may also be cyclical elements or many processes running parallel. Also in many cases we do not have knowledge of the temporal order. More possible relations are suggested by Grenon/Smith (2004, 290 f.), such as relations that express more specific relations than "has_participant" between continuants and processes like: "x initiates p", "x terminates p". For sake of parsimony it seems reasonable to restrict oneself to as few as possible, but this has to be decided in the modelling of the specific mechanism of interest.

4 CONCLUSION

It has been argued that an ontological analysis of biological mechanisms needs both continuants as their material parts and occurents that represent the changes of these components. The parts of the mechanism as well as the mechanism as a whole are functionally identified and therefore closely linked to the processes they enact. To make this explicit a conception of functional parts and a relation of contribution between functions have been introduced. This is a step towards a causal analysis of complex biological systems. The contributions of the functions of parts to the whole could be systematically collected on the basis of ontologies that contain the functions of important components of biological mechanisms like the Gene Ontology.

5 OUTLOOK

My goal was to bring together the rather general account of mechanisms proposed by philosophers of biology and more rigorous considerations from formal biomedical ontology. Admittedly, many details are still in need of refinement and the analysis could be made more precise in several ways: First of all, the context-dependence of functional parts would have to be clarified further. One could also capture differences like the distinction between „active" and „passive" components and between changes of already existing and the generation of new components. Also connections with topological and geometrical relations should be explored, such as the specification of the spatial region the mechanism as a whole or salient components of it occupy.

ACKNOWLEDGEMENTS

Research for this paper was funded by the German Research Foundation (DFG) within the research project "Good Ontology Design" (GoodOD). Thanks to Ludger Jansen, Niels Grewe, conference participants in Vienna and Rostock, and three anonymous referees for helpful comments.

REFERENCES

Arp., R. and Smith, B. (2008): Function, Role, and Disposition in Basic Formal Ontology, *Proceedings of Bio-Ontologies Workshop (ISMB2008)*, 45–48.

Bechtel, W. (2006): *Discovering Cell Mechanicsm*, Cambridge.

Bechtel, W. (2007): Biological mechanisms: organized to maintain autonomy, in Boogerd et al. (eds.): *Systems Biology. Philosophical Foundations*, Amsterdam 2007.

Bechtel, W./Abrahamsen, A. (2005): Explanation. A mechanist alternative, *Studies in the History and Philosophy of Biological and Biomedical Sciences*, 36, 421-441

Beisswanger, E./Schulz, S./Stenzhorn, H./ Hahn, U. (2008): BioTop An Upper Domain Ontology for the Life Sciences, *Applied Ontology* 3, 205-212.

BioTop. A top domain for the life sciences [http://purl.org/biotop/] (release July 1, 2010)

Boogerd, F.C/Bruggeman, F.J./Hofmeyr, J.S./Westerhoff, H.V. (eds.) (2007): *Systems Biology. Philosophical Foundations*, Amsterdam.

Boogerd, F.C et al. (2007): Afterthoughts as foundations for systems biology, in: Boogerd et al. (eds.), 321-336.

Burek, P./Hoehndorf, R./Loebe, F./Visagie, J./Herre, H./Kelso, J. (2006): A top-level ontology of functions and its application in the open biomedical ontologies, *Bioinformatics* 22, No.14, e66-e73.

Campana, R. (2011): Understanding mechanisms in the health sciences, *Theor. Med. Bioeth*. 32:5-17

Galton, A.; Mizoguchi, R. (2009): The water falls, but the waterfall does not fall: New perspectives on objects, processes and events, *Applied Ontology* 4, 71-107.

Glennan, S. (2002): Rethinking mechanistic explanation, *Philosophy of Science* 69, 342-353.

Grenon, P./Smith, B. (2004): "The Cornucopia of Formal-Ontological Relations, Dialectica 58, NR. 3, 279-96.

Jansen, L. (2007) Tendencies and other Realizables in Medical Information Sciences, *The Monist* **90/4**, 534-555.

Jansen, L./Schulz, S. (2011): Grains, components and mixtures in biomedical ontologies, to appear in *Journal of Biomedical Semantics*

Johansson, I./Smith, B./Munn, K./Tsikolia, N./Elsner, K./Ernst, D./Siebert, D. (2005): Functional Anatomy: A Taxonomic Proposal, *Acta Biotheoretica* 53: 153-166.

Krohs, U./Kroes, P. (eds.) (2009): Functions in biological and artificial worlds: comparative philosophical perspectives, Cambridge, Mass.

Machamer/Darden, L./Craver, C. (2000): Thinking about mechanisms, *Philosophy of Science* 67.

Richardson, R.C./Stephan, A. (2007): Mechanism and mechanical explanation in systems biology, in: Boogerd et al.

Roehl, J./Jansen, L. (2011): Representing dispositions, to appear in *Journal of Biomedical Semantics*

Rosen, R. (1991): *Life itself*, New York.

Rosen, R. (1998): *Essays on Life itself* , New York.

Schulz, S. and Jansen, L. (2009): Molecular Interactions: On the ambiguity of ordinary statements in biomedical literature, *Applied Ontology* **4**, 21-34.

Smith et al. (2005): Relations in Biomedical Ontologies. *Genome Biology* 2005 6:R46.

Spear, A.D. (2006): Ontology for the 21st Century. An introduction with Recommendations (BFO Manual), http://www.ifomis.org/bfo/documents/manual.pdf.

Tabery, J. (2004): Synthesizing Activities and Interactions in the Concept of a Mechanism, *Philosophy of Science* 71.

Vieu, L. and Aurnague, M. (2007): Part-of relations, functionality and dependence, in Aurnague, M/Hickmann, M./Vieu, L. (eds.): *The Categorization of Spatial Entities in Language and Cognition*, 307–336. Amsterdam.

Representing 'casualties' for epidemiological data processing

Filipe Santana[1], Roberta Fernandes[1], Daniel Schober[2], Cristine Bonfim[3], Zulma Medeiros[4,5], Fred Freitas[1]*

[1]*Informatics Center, Federal University of Pernambuco (CIn/UFPE), Recife, Brazil,*
[2]*Institute of Medical Biometry and Medical Informatics (IMBI), University Medical Center Freiburg, Freiburg, Germany,*
[3]*Social Research Board, Joaquim Nabuco Foundation, Recife, Brazil,*
[4]*Parasitology Department, Aggeu Magalhães Research Center, Oswaldo Cruz Foundation, (CPqAM/Fiocruz), Recife, Brazil,*
[5]*Pathology Department, Institute of Biological Sciences, University of Pernambuco, Recife, Brazil,*

ABSTRACT
This paper presents an ontological approach to formally describe the events related to the passing from life to death, supporting the retrieval of mortality cases (e.g. available in mortality databases). Such representations are needed to support public decision making related to highly disabling diseases like infectious or neglected tropical diseases.

1 INTRODUCTION

Epidemiologic policies are defined as plans for actions to support preventive and control measures for diseases or others health related problems, based on knowledge about individual and collective health. One of the most important data sources to generate *information to provide policies for decision making* is mortality data (Selig *et al*., 2010).

Casualties and diseases, like tuberculosis (TB), can be considered sentinel events for health monitoring systems. (Selig *et al*., 2004). Together with HIV/AIDS and Malaria TB is one of the three most devastating worldwide diseases) (Hotez et al., 2006), resulting in nine million newly reported TB disease cases and 1.7 million casualties each year. Such diseases require new preventive strategies, like the WHO-Stop TB program (WHO, 2009), which should ideally be based on reliable information stored in data bases, providing access to reliable epidemiologic as well as individual health care data. This should ideally be embedded in their local contexts, ultimately contributing to an enhanced understanding of the dynamics of disease spread (Cohen, 2000).

Therefore, policy efforts to help fight diseases in specific countries should embrace local information and patient data in an epidemiological database management system (Selig *et al*., 2010). These should be based on ontologies to provide the rich embedded contextual data in an integrated way and as demanded above. Exploiting consensual knowledge, as formalized in ontologies, can produce new epidemiological insight and thus help in policy management and decision-making processes (Topalis et al., 2011). In order to investigate such capabilities, we created the Neglected Tropical Disease Ontology (NTDO) (Santana *et al.,* 2011).

Ontologies, from a formal point of view, intend to describe the consensus on the nature of entities in a given scientific domain, independently of linguistic variation. Accordingly, formal ontologies are expressed by means of a formal semantics, like Description Logics (DL) (Baader et al., 2007). Nowadays, most DL ontologies are , are shared in the World Wide Web Consortium (W3C) recommended exchange syntax Web Ontology Language (OWL)[1].

The aim of this study is to ontologically formalize foundational events in the life cycle of patients, i.e. events related to the passing from life to death, and representations thereof as needed to track, store and retrieve mortality cases. Ontologized mortality databases can support decision making against relevant casualty events. Here, we used TB cases to exemplify and create this subset representation. We particularly focused on information required to support health policy management, e.g. in the case of NTD and TB using ontologic representations. We hope to show that such ontologized epidemiological data can be exploited by logics reasoners and hence render implicit data explicit and more useable by retrieval and disease monitoring tools.. Specifically we provide patterns for robust and DL-compliant ontological representations of epidemiologically important entities like Birth, Disease, and Death.

2 MATERIAL AND METHODS

NTDO (http://www.cin.ufpe.br/~ntdo) imports and re-uses the upper level ontology BioTop (http://purl.org/biotop) (Beisswanger et al., 2008). NTDO was expanded in a middle-out approach, leveraging on established ontology construction guidelines (Rector, 2003; Schober et al., 2009). The pathogen transmission pattern was based on Santana *et al.* (2011), providing the basis to describe the tuberculosis airborne transmission, its respective pathogens and affected persons.

To perform the data retrieval, the Brazilian Mortality Information System (*Sistema de Informação sobre Mortalidade* - SIM[2]) database was converted (from dBase to SQL), and views were created (using PostgreSQL v9.0) in order to extract demographic (age, sex, among others) and epidemiological data (deceased person, place of casualty event and casualty basic cause). After, we used JENA API to generate RDF triples with individual assertions; and mappings to the NTDO (OWL2) respective classes to enable RDF-querying over the mortality data using SPARQL[3] Query Language.

*To whom correspondence should be addressed: fss3@cin.ufpe.br

[1] http://www.w3.org/TR/owl2-overview
[2]http://portal.saude.gov.br/portal/saude/visualizar_texto.cfm?idtxt=21377
[3] http://www.w3.org/TR/2008/REC-rdf-sparql-query-20080115/

The mortality data gathering was approved by the Ethics Comitee of Health Sciences Center (CCS) (Federal University of Pernambuco - UFPE), as a subpart of the project "Ontologias e as Doenças Tropicais Negligenciáveis" (CAAE - 0112.0.172.000-11), in English, "Ontologies and Neglected Tropical Diseases".

3 RESULTS

3.1 Casualty representation

In this section, all ontology modeling processes, e.g.pathological processes, casualty and transmission process, and data analysis, e.g. mortality data retrieval, are scrutinized.

Our ontological representation of casualty cases, as foundationally relevant for any epistemological analysis, follows the model described in Figure 1.

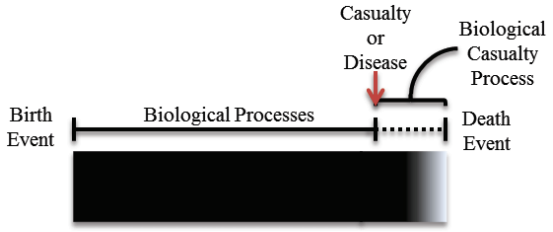


Figure 1:Organism life cycle model, including a Birth event and a Casualty Process Model.

As *birth*, *life, disease, and death* seem to be diffusely delineated concepts people tend to argue about, we introduce a semiformal notion according to Koshland (2002), who defines. "Life" as the capability of a program which has a description of the ingredients and their interaction kinetics (like the genome and the metabolome), capable of mutation and hence prone to selection . The bearer of such Life capability is an organism compartmentalized into cells and organs, which metabolize substances to generate energy for adaption, regeneration and segregation.

In addition, the notion of "birth" is based on the notion of "Living Birth", as defined by the Brazilian Geography and Statistics Institute (IBGE) as 'the expulsion or complete extraction of a product from the maternal body, after conception which after the detachment of the maternal body, breaths or gives any other life-sign, e.g. heartbeat, umbilical cord beat, or movements from voluntary muscle contraction, the umbilical cord being cut or not, and the placenta being detached or not[4]. On the other hand, 'death' means the complete extinction of any life-sign in any moment after a 'Birth"- event, i. e. cessation of the vital functions without resuscitation.

Therefore, a typical lifecycle of a living organism begins with a conception event followed by a pregnancy process which ends with a birth event (a point in time locating when it happened). Here, we are considering only the processes which happen after the conception. It is important to note that *Events* here are known to exhibit a certain behavior relative to a process (Herre *et al.*. 2007). Each organism has a lifespan and, at a later point in time, its body starts a biological death process, which can be caused by natural means, disease or casualties (as stated by the World Health Organization in "cause of death" definition[5]). It culminates, independently of how long such processes may take, with death.

3.2 Representational challenges

Many challenges were faced in the attempt to create a sound representation of such a structure, such as preserving identity, asserting correct cardinalities, getting support of sound background theories, and, last but not least, representing the resulting ontology in a decidable DL that could encompass the expressivity employed in the definitions. We will discuss each of these items in the following. An initial definition of a *CasualtyEvent* could be

CasualtyEvent equivalentTo *Event*
and (**hasLocus** some *GeographicLocation*)
and (**hasPatient** some *DeadOrganism*)
and (**hasProcessualPart** some *BiologicalDeathProcess*)
and (**hasCasualtyInstant** some *PointInTime*)

Such a definition brings many imprecisions with regard to preserving identity and correct cardinalities. First of all, the representation purpose of this class is conveying information on the casualty of a single living organism. However the axiom does not express any cardinality constraint. Moreover, there is no guarantee that the living organism that is dying and the resulting dead organism coincide, retaining identity between these two. It makes a living organism coincides with the definition of a phased sortal, i.e., one which starts by enjoying a certain phase (in our case, a living phase) and eventually turns into a new phase (for us the phase of a dead organism).

This indeed constitutes an interesting representation problem, given that it brings about the philosophical issue of representing most (if not all) rigid classes (Guarino & Welty, 2000) as phased sortal, not to mention that it additionally provokes a discussion whether a *DeadOrganism* is still an *Organism* or not and until exactly when.

Instead, a sound solution resides on representing *LivingOrganisms* subsumed by a *gfo:Presential* (GFO, Herre *et al.*. 2007), which is a *biotop:MaterialEntity*. A *Presential* exists only at exactly one time interval (in the ontology called a *Chronoid).* As described in GFO, *Chronoids* possess two inherent and external time boundaries, *RightTimeBoundary* and *LeftTimeBoundary* (subclasses of *TimeBoundary*), (Fig. 2), which can coincide in a single *Chronoid*. Accordingly, *LeftTimeBoundary* and *RightTimeBoundary* represents the beginning and end, respectively, of an inner *ProcessualEntity*, realized by an object. The mereological sum of *Chronoids* represents the notion of a time region (Herre *et al.,* 2007).

[4]http://www.ibge.gov.br/

[5] http://www.who.int/topics/mortality/en/

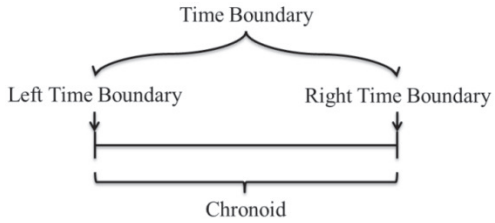


Figure 2: A *Chronoid* time interval and its boundaries along the time axis.

The axiom below should be included then:

MaterialEntity subClassOf *Presential*

departing from the premise that *LivingOrganisms* are *MaterialEntities*. Of course, we are also assuming that *MaterialEntities* are formed (*LeftTimeBoundary*) and destroyed (*RightLeftTimeBoundary*). Below are the GFO describing a *MaterialEntity* according to its *TimeBoundaries*:

MaterialEntity equivalentTo *ParticularEntity*
and (**existsAt exactly** 1 *TimeBoundary*)

We also need events that take place in an instant, like death. For this purpose, we define an Instant Event as an event in which its time boundaries coincide (using the agreement operator ≐).

InstantEvent equivalentTo *Event*

and *(***projectsTo o hasLeftTimeBoundary ≐ projectsTo o hasRightTimeBoundary***)*
and *(***hasInstant** *some TimeBoundary)*

The property **hasInstant** is defined as one of them:

hasInstant = projectsTo o hasRightTimeBoundary

Another subtle aspect regards the range of the **hasPatient** property. For our purposes, this object property should be functional (i.e., each relation instance admits only one element in the domain and range), since health notifications are about a single individual. Nevertheless, it is not originally like that in the relation definition, which allows for more than one element from the range. In the ontology, this could mistakenly lead to the interpretation that a single casualty event can indeed represent the death of many individuals at the same time. Accordingly, we created the subproperties **hasCasualtyPatient**, **hasConvalescentPatient** and **hasDeadPatient**, all functional properties.
Even when we substitute the unknown cardinality ("some", in the axiom) by a defined cardinality (e.g. **hasPatient exactly** 1 *DeadOrganism*), the problem with preserving identity persists, and still in the case when a *LivingOrganism* is transformed into a *DeadOrganism*.
This indeed constitutes an interesting representation problem, given that it brings about the philosophical issue of representing most (if not all) rigid classes (Guarino &Welty, 2000) as phased sortal, not to mention that it additionally provokes a discussion whether a *DeadOrganism* is still an *Organism* or not and until exactly when.

A good way to circumvent such representational problems - and probably also the common choice of the ontologists who designed all of the other biological ontologies that we found in the literature to the extent of our knowledge – is not representing a *DeadOrganism* at all.
Bearing this in mind, we represented each event (*BirthEvent*, *CasualtyEvent* and *DeathEvent*), stating its participant, the place the event takes place and the cause.

BirthEvent equivalentTo *InstantEvent*
and *(***hasLocus** *some GeographicLocation)*
and *(***hasPatient** *some LivingOrganism)*

Naturally, this birth modeling is very simple and cannot comprehend all the situations of parenthood that can occur. However, since we are interested in modeling casualties and death, we will not dive into these issues and go on with the definitions of Casualty and Death:

CasualtyEvent equivalentTo *InstantEvent*
and *(***hasLocus** some *GeographicLocation)*
and *(***causedBy** some
(ProcessualEntity and (not BiologicalProcessualEntity)))
and *(***hasCasualtyPatient** some *LivingOrganism)*

DeathEvent equivalentTo *InstantEvent*
and *(***hasLocus** some *GeographicLocation)*
and *(***hasDeathPatient** some *LivingOrganism)*
and *(***hasDeathPrimaryCause** some *ProcessualEntity)*
and *(***hasDeathPatient =**
hasDeathPrimaryCause o **hasConvalescentPatient)**

Note that the last condition including an equality role-value map (=) assures that the dead individual is the same one who was participated in a prior casualty, thus preserving identity. For our modeling purposes, our goal was finally attained; nonetheless, for DL reasoning with the ontology, additional measures are still to be taken. If equalities are not built over property chains of functional properties in role-value-maps, then inference becomes undecidable (Schmidt-Schauss 1989). Nonetheless, that constraint could only be fulfilled by our ontology when its properties were functional. The property **hasDeathPrimaryCause** is functional but the initial **hasPatient** is not. The solution for this was creating sub-properties **hasDeathPatient** and **hasConvalescentPatient,** which are both functional and have as range the concept *LivingOrganism*. We now go on with the last necessary definitions of the modeling.

BiologicalDeathProcess equivalentTo BiologicalProcessualEntity and (**precededBy** some (*PathologicalCondition* or *CasualtyEvent* or (*BiologicalProcessualEntity* and (not (*BiologicalDeathProcess*)))))
and (**hasConvalescentPatient** some *LivingOrganism*)
and (**hasInstant** some *TimeBoundary*)

It is important to note that the used notion of causation depends on the process observer. Summarizing, the figure 3 has all assumptions taken here

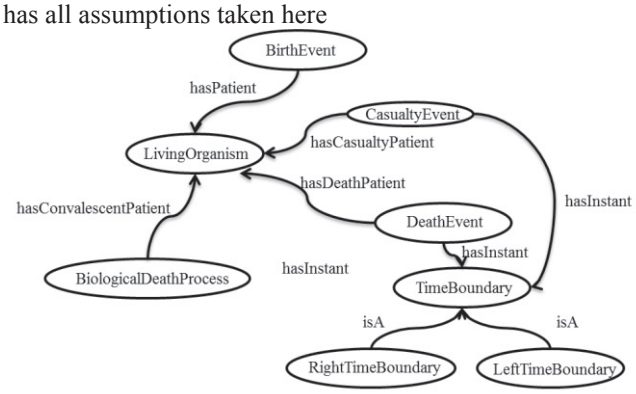


Figure 3: Graph-model of a DeathEvent, CasualtyEvent and BirthEvent.

which describes the transitional profile of a convalescent *LivingOrganism,* from life to death, followed by its cause. It is important to note that the causation here is defined by the process observer, which means that the main cause of a *BiologicalDeathProcess* could not be the one defined for it.

Following, few agreements are required to express the time sequence between one event and a process, for example to express that a person in a given time a *BiologicalDeathProcess* started and when the *DeathEvent* begins, the end of a *BiologicalDeathProcess* and the *DeathEvent* can coincide.

3.3 Disease and Transmission Representation

Our TB representation follows the distinction between disease and disorder (Schulz, 2010). The differences between sign and symptoms were not taken into account.

The transmission path was created based on a vector borne transmission model (Santana *et al.,* 2011), which was adapted to fit TB and its airborne transmission cycle.

The model states that in an airborne transmission process, the 'vector' (for TB, air or dust) is an agent and the 'pathogen' (*Mycobacterium tuberculosis*) is an additional participant and can cause a *PathologicalProcess* (*Tuberculosis*), due to host favorable conditions.

3.4 Data Analysis

A data pre-analysis was performed in order evaluate suitability of potentially epidemiologically relevant data from different mortality databases. The SIM dataset contains information about most death events occurring in Brazil and was deemed suitable for populating our knowledge base.

SIM data are divided by years and has full demographical information covering the causal circumstances which led to the death of a victim. Using TB casualty data generated between 1996 and 2003, we selected cases by main cause of death, age, sex, among others. The data was added to the knowledge base as instances of classes, like *Person*, *GeographicLocation* and *DeathEvent*. An example SPARQL query to show which death happens by place, by causation, due to which disease.

```
SELECT ?deathEvent ?placeocurr ?condition
WHERE { ?deathEvent rdf:type ntdo:DeathEvent;
        biotop:hasLocus ?locusocurr;
        biotop:causedBy ?condition.
        ?locusocurr ntdo:name ?placeocurr.
        ?condition rdf:type ntdo:Tuberculosis.}
```

Using this example query, it collects 205 cases from Pernambuco State, in which 161 cases occurred in Recife. It denotes this location as a main point for health policy actions for the population.

For our work, the choice of SPARQL, instead of SQL, lies in the graph based structure of the RDF data: SIM has two different data schemas (before 2000 and after 2001), whose differences does not affect SPARQL queries. Furthermore, this choice provides reasoning over data which cannot be found in simple SQL queries.

Acknowledgements: This work was sponsored by the German DFG grant JA 1904/2-1, SCHU 2515/1-1 GoodOD (Good Ontology Design) and German Ministry of Education and Research (BMBF)-IB mobility project BRA 09/006.

4 REFERENCES

Baader,F. *et al*. (2007) The Description Logic Handbook. Theory, Implementation, and Applications, 2nd edn. Cambridge University Press, Cambridge.

Beisswanger, E. *et al*. (2008). BIOTOP : An Upper Domain Ontology for the Life Sciences. *Appl. Ontology*, 3(4), pp.205-212.

Gruninger,M.and Fox,M. (1994). The role of competency questions in enterprise engineering. In IFIP WG 5.7, Workshop Benchmarking. Theory and Practice, Trondheim/Norway.

Guarino, N. & Welty, C. (2000). Ontological Analysis of Taxonomic Relationships. In A. Lander & V. Storey, eds. *Proceedings of ER-2000: The International Conference on Conpectual Modelling*. Springer-Verlag LNCS, pp. 1-15.

Herre, H *et al*.. (2007). General Formal Ontology (GFO): A Foundational Ontology Integrating Objects and Processes. Part I: Basic Principles. Research Group Ontologies in Medicine (Onto-Med), University of Leipzig.

Hotez, P.J. *et al*. (2006). Incorporating a rapid-impact package for neglected tropical diseases with programs for HIV/AIDS, tuberculosis, and malaria. *PLoS medicine*, 3(5), p.e102.

Koshland, D.E. (2002). The seven pillars of life. *Science (New York, N.Y.)*, 295(5563), pp.2215-6.

Rector,A.L. (2003) Modularisation of domain ontologies implemented in description logics and related formalisms including OWL. In Proceedings of the international conference on Knowledge capture - K-CAP'03. ACM Press, New York, USA,p. 121.

Santana, F. *et al*. (2011). Ontology patterns for tabular representations of biomedical knowledge on neglected tropical diseases. *Bioinformatics*, 27(13), p.i349-i356.

Selig, L. *et al*., (2004). Óbitos atribuídos à tuberculose no Estado do Rio de Janeiro. *Jornal Brasileiro de Pneumologia*, 30(4), pp.335-342.

Topalis, P. *et al*. (2011). A set of ontologies to drive tools for the control of vector-borne diseases. *Journal of biomedical informatics*, 44(1), pp.42-7.

World Health Organization (2009). *Global tuberculosis control: A short update to the 2009 report*, Geneva.

Continuation-like Semantics for Modeling Structural Event Anomalies

Niels Grewe
University of Rostock, Rostock, Germany

ABSTRACT

Biomedical ontologies usually encode knowledge that is easily standardised and applies always or at least most of the time. But for certain applications (e.g. phenotype ontologies), encoding information about aberrations from a norm is becoming increasingly important. Many of such aberrations are related not only to physiological structures but also to the processes that these structures participate in. We suggest a method for dealing with certain anomalous features of such processes, chiefly delays and interruptions. This modeling scheme draws inspiration from the use of continuations in the analysis of programming languages and applies a similar idea to ontology modeling.

1 INTRODUCTION

The portion of reality under scrutiny by life sciences is much more exposed to the phenomenon of variability than, for example, chemistry or physics. Consequentially, many biological truths only hold 'normally' or 'for the most part.' Biomedical ontologies, if they are considered to be information artifacts modeling or representing some portion of the underlying reality at all, usually strive to capture only the aspects that are subject to some regularity, because it seems that no knowledge whatsoever can be gleaned from random aberrations.

In some areas, however, systematic considerations of the deviations from the normal case are of indisputable importance. One example is medical diagnostics, where pathological (and hence aberrant) phenotypes are a primary means for making inferences about the cause of a patient's condition. Ontologies that provide structured access to phenotypical information are thus becoming valuable tools for researchers and clinical practitioners alike.

Examples of such ontologies include the Mammalian Phenotype Ontology [13] or the Human Phenotype Ontology, [10] which both make use of the *Phenotype, Attribute and Trait Ontology* (PATO), which seems to have emerged as an accepted standard for specifying information about phenotypes. [8]

The problem of the relationship between clinically normal and pathological is by itself troubling enough for the formally minded ontology engineer, and has, for example, driven research into the use of nonmonotonic logics (e.g. default logic) for this kind of application [5]. But it should also be noted that the problems arising from the distinction are further aggravated by the fact that the term 'phenotype' is everything but a mono-categorial term. Phenotypes can describe traits not only pertaining to the concrete bodily structures, but also those which describe locations of such structures, dispositions or processes (cf. table 1).

Abnormal phenomena in each of these categories seem to deserve separate treatment; something that is neatly reflected in the fact that PATO defines the classes *process quality* and *physical object quality* as disjoint from one another. But the disjointness does not hint that both categories are completely unrelated: We will, for example, always assume that a quality of a process has something to do with the continuants participating in that process (e.g. the process quality *rate of osmosis* of an osmosis process will, among other things, depend on the concentration of solvent molecules and the permeability of the membrane for the molecules in question).

HPO ID	Phenotype	Related Category
HP:0010442	*Polydactyly*	Material Object
HP:0001100	*Heterochromia iridis*	Quality
HP:0008522	*High-frequency deafness*	Disposition
HP:0001696	*Situs inversus*	Location
HP:0000823	*Delayed puberty*	Process

Table 1. Phenotypes in different ontological categories

This suggests that it might be desirable to spell out process related phenotypes in terms of qualities of continuants.[1] Such definitions are conspicuously absent from the *process quality* subtree of PATO, but its members are extensively used, for example, in definitions of the HPO.

One example is the process quality *delayed*, which features in the definition of 47 classes in HPO, whether informally or explicitly referencing the PATO class PATO:0000502 (e.g. *delayed eruption of primary teeth*, HP:0000680). While this only accounts for less than half a percent of all HPO classes, it is an example of a certain type of process anomaly that could be termed a structural anomaly (as opposed to 'material' anomalies, such as increased or decreased frequency, etc.). This kind of anomaly seems to be relatively uninvolved with concrete biological problems that usually arise from this kind of reasoning. Structural anomalies thus seems to be a useful subject for an initial case study of how anomalies of processes could be treated, which is what I will attempt in what follows.

2 THE SEMANTICS OF PROCESS ANOMALIES

2.1 An Analogy from Holes

In order to get a better picture of what features accurately characterise the anomalies of processes, it is useful to consider analogues in continuants as a starting point. The reason for this is that our grasp of occurrent entities is usually weaker than our grasp of continuant

[1] This issue should be separated from the issue of causal explanations of processes: For example, a patient's tachycardia could be explained by an elevated level of norepinephrine in that patient's blood. But this is a causal explanation that could be part of a physician's diagnosis, not an explanation of what it means of a process to be a tachycardia, e.g. a certain state of the heart and the nervous system.

objects, since the latter stay around for close examination as long as we like.

One continuant analogue readily presents itself if one considers some of the more serious siblings of delays, namely interruptions or disruptions (PATO:0001507) of processes.[2] If we consider instances of this kind of entity, we might speculate that there is some likeness between them and holes in continuants. The reason for this might be that we observe that every hole in a continuant corresponds to a discontinuity in the surrounding material, [6] just as interruptions always coincide with discontinuities of processes.

But the analogy is imperfect at best for several reasons: (1) The surroundings of holes are mostly continuous, so that we can without any hesitation distinguish a hole in a piece of cheese from a gap between two distinct pieces of cheese. But since we specify processes as extending along a single temporal dimension, distinguishing between them is no longer easy. Unless we want the difference between a gap and a 'hole' to be blurred, this suggests that we need an identity criterion for processes and events[3] that does not depend on temporal continuity.

(2) There seems to be no room for gradations of hole intensity, but clearly a delay and an interruption in a process are interfering with the process in a similar way, but with a different severity.

(3) Whether there are holes in a continuant is not at all affected by whether we think that it is normal or essential for the thing to have holes. This is not the case with interruptions and delays. For an episode within an event to be an interruption or a delay requires that we look at the normal or canonical course of the event. For example, if Mary gets on a train in Berlin and off the train in Brussels, one cannot say that her travels have been interrupted simpliciter. We rather need to know whether she was traveling from Berlin to Brussels (no interruption) or from Berlin to London. In this case, this could be a interruption, but only if the normal course of events would not have involved a stop in Brussels.[4]

2.2 Semantical Considerations For Modeling Process Anomalies

While the continuant analogy is not really fruitful in an explanatory way, it does draw attention to the peculiarities of delays and interruptions that have to be taken into consideration:

1. We do need identity criteria to re-identify events that contain interruptions.
2. We need to account for the differences between different kinds of structural anomalies (at least for delays and interruptions).
3. We need to establish the relationship between the anomalous and the normal form of the event.

Our proposal to tackle 1 and 3 will follow quite straightforwardly from our formal treatment of the matter (as presented in section 3), but 2 deserves some additional clarifications. For one, we need to state an ambiguity about the meaning of 'delayed'. The corresponding process quality *delayed* is defined in PATO as follows:

> A duration quality of a process inhering in a bearer by virtue of the bearer's duration which starts later than the natural start time. (PATO:0000502)

It seems that this definition does not encompass everything that would be called a delay. For example, Mary might be entitled to the claim that her travels from Berlin to London were delayed even if the delay did not result from the first train leaving later than it should have (with respect to the timetable) but rather from some unforeseen stop in Brussels. This concern is further amplified by the realisation that the phrase 'natural start time' (of the process, that is), needs to involve some reference to an overarching process regarding to which the process in question is said to be delayed. For example, *delayed eruption of the primary teeth* might mean something different if one regards the normal developmental process of a mouse or of a human being as the frame of reference.

Secondly, our intuitive answers to the question of the duration of a process are highly dependent on the severity of the process anomaly. While one usually would affirm that a process is still in effect during a episode that might be labelled a delay (and hence the delay contributes to the overall duration of the process), one would be hesitant to state the same thing about a disruption of a process: When there is a disruption of a process, we usually claim that it is not taking place and hence the disruption episode should neither count as a part of the process in question nor should it contribute to the overall duration of that process.[5] These issues should be kept in mind when considering a modeling approach to continuations.

3 THE CONTINUATION MODEL

3.1 Continuations in Computer Science

Our approach to modeling structural anomalies of processes relies on the concept of continuations, which has been successfully employed by computer scientists to tackle a variety of seemingly divergent problems in the realm of programming language design and programme analysis.[6]

Roughly speaking, a continuation is an abstract data structure that represents a certain point in the control flow of a programme by specifying the state of the computation at that point and how the computation will continue. A continuation thus specifies the '(meaning of the) "rest of the program"'. [15, p. 132]

```
hypotenuse(a, b)
{
  return sqrt(sum(square(a),square(b)));
}
```

Listing 1. Imperative Style Computation

[2] We will use the term 'interruption' and 'disruption' interchangeably.

[3] In this paper, I will not make a principled distinction between events and processes, eventhough it is definitely necessary to give one. [2] It seems to me that the occurrent entities relevant for the current discussion are mostly events, but often it is just more natural to speak of them as 'processes'.

[4] It might also be that Mary's train ride is interrupted in Brussels, but not her journey, e.g. if she decides to rent a car in Brussels.

[5] Though interruptions of subprocesses might be closely correlated with delays of their superprocess. For example, if Mary's train ride is interrupted a few times, she is effectively not riding the train during those interruptions. Still, the total duration of her journey increases through these interruptions because they can count as delays of the journey.

[6] For a historical outline of the research in continuations, which also highlights their diverse areas of application, see [9].

It is convenient to approach the topic of continuations by giving an example of their use. One such use is the transformation of a computer programme written in an imperative language into a notation that can be interpreted in a functional way – something that is very useful when specifying the denotational semantics of a programme.

Let us consider a common control flow operation in imperative programming languages: Returning values from a subroutine to the caller of that subroutine. For example, a function called $sum(a, b)$ in a computer programme might compute the sum of a and b and then return the computed value to the caller, which in turn might do additional computations with the obtained value and return the result thereof to its caller. With continuations, the control flow statement '*return*', required for returning values, can be disposed of. Instead, each function or subroutine can be written as taking an additional argument, namely the function which should be called with the result of the computation as an argument. That function is then the continuation of the subroutine in question because it specifies how the computation will continue.

This kind of programme formulation is aptly called 'continuation passing style' [14, p. 421] and its peculiarities will become clear from the differences between the pseudo-code snipplets 1 and 2: Wherever there is a return from a function call in listing 1, there can be identified a lambda term in listing 2 that effectively encodes what remains to be done with the result of the present computation.

```
hypotenuse(a, b, k)
{
  square(a,
    (λs_a . square(b,
      (λs_b . sum(s_a ,s_b ,
        (λs_ab . sqrt(s_ab , k)))))));
}
```

Listing 2. Continuation Passing Style Computation

For our present purpose continuations will show their usefulness if we do not consider their ephemeral variants that are merely applicable at a given point in the execution of the programme, but rather continuations that allow the present execution state of the programme to be stored alongside the information about how the execution is going to proceed. Such continuations are powerful enough to serve as models for various design patterns such as cooperative multitasking (coroutines), or exception and interrupt handling.

In the latter case some external intervention requires that the normal execution is suspended in order to take some immediate actions. With continuations, this can be conceptualized as saving the present continuation of the normal execution process and passing it to the subroutine that handles the interrupt, which will call it as its continuation after performing the necessary tasks.

3.2 Process Continuations

3.2.1 Preliminaries These characteristics of continuations seem to come in quite handy when it comes to the structural anomalies of processes that we are considering here. Our strategy will thus be to describe processes by associating them with their corresponding continuations such that for every point of time (except for the last) at which the process is in effect there exists a continuation of the process. That continuation describes the present state of the process and how it will continue.

Since continuations in the realm of functional programming are purely mathematical concepts, they are devoid of any relation to time and just implicitly specify the required order of computation. This is an important difference to the intended use in the realm of process modelling.

Consequentially, the way this proposal needs to be spelled out is highly dependent on the underlying ontology of time. But while all major top level ontologies (e.g. BFO, DOLCE or GFO) provide at least some account of time, it seems that a commonly accepted, standard ontological account of temporal phenomena has yet to emerge. Hence we restrict ourselves to making clear some of the prerequisites of our approach, all of which should be achievable no matter what top level ontology one chooses:

- Since process continuations need to capture the present state of the process, the underlying ontology needs to contain complex ontological entities to model such states, e.g. through states of affairs [1] which represent the fact of something's being such-and-such (for example, a tomato's being red is a state of affairs composed of the tomato and the quality *red* inhering in that tomato).
- Processes can be made up from subprocesses, hence processes can, but need not, have temporal parts.
- Since processes usually involve things changing, each process needs to be associated with (at least) an initial or input state and a final or output state. [2, p. 86] (In a weaker sense, a process might also be an episode of absence of change. In this case, the initial and final state will be identical.)
- I will assume that the underlying formalisation of time is such that two processes in direct succession coincide at a common boundary (something that is made explicit in the BFO top level ontology by the class *ProcessBoundary*. [3] This way, it is possible to claim that the final state of the first process might serve as the initial state of the second process. With regard to the first process, the boundary will called a right boundary, with regard to the second process the boundary will be called a left boundary.

  This requirement is sufficient to express 'conventional' change, where the separation of an event into subevents is such that the result of the preceding event is 'picked up' by the succeeding event, as is the case in metabolic pathways. Hence, the requirement might not be sufficient to express continuous change or so called 'Cambridge change' [7] where the change occurs between to contrary or contradicting states. To handle this kind of change, more complex formal machinery, such as the theory of boundaries sketched in the GFO [4, pp. 45–46] might be needed. Adapting the modeling strategy presented here should be easily possible.

3.2.2 Process Continuations and Anomaly-Invariant Descriptions With these provisos, a continuation of an event or process can be defined as follows:

DEFINITION 1. *κ is a continuation of the event e iff*

1. *κ is a continuant.*

2. *for every timepoint t and every independent continuant c, if e is in effect at t and κ exists at t and e is ontologically dependent on c at t, then κ is ontologically dependent on c.*
3. *there exists some proper subevent e_c of e and a timepoint t, such that the right boundary of e_c is at t and the left boundary of κ's life-time is also at t.*
4. *there exists some proper subevent e_s of e and some state of affairs s_c such that*
   a. *s_c is the final state of e_c and κ is ontologically dependent on s_c.*
   b. *the left boundary of e_s coincides with the right boundary of κ's life-time and s_c is the initial state of e_s.*

In this definition, clause 1 is more than just a play on words. Continuations also have to be (dependent) continuants because they fulfill the canonical definition of a continuant as a thing that is wholly present at every point of its existence. The reason for this is that we want to assume that the continuation comes to be once all the conditions relevant for advancing the course of events obtain.

A crucial part of these conditions is specified in clause 2: If the process is ontologically dependent on some entity at a given stage (meaning that the entity participates in the process), then the continuation cannot exist without that entity's continued existence.[7]

With clause 3, the definition stipulates that a continuation has to be the current continuation of at least one subevent of the overarching event e, namely of the subevent up to which the event has successfully progressed. This requirement goes hand in hand with clause 4a, which specifies that the continuation depends on the state of affairs that is the final state of the subevent of which the continuation is the current continuation. I will call this state of affairs the *context state* of κ. Conversely, by clause 4b, that state must also be the initial state of the succeeding subevent, so that the continuation really specifies how the event will continue.

This definition allows for a great deal of variability. It does not, for example, stipulate that the subevents related by the continuation are contiguous, something that is crucial for the purpose of modeling interruptions. Still all crucial information about the event is represented in its continuations. It is hence useful to define the *continuation set* of all continuations of e as well:

DEFINITION 2. *Let e be an event, then K_e is the continuation set of e iff*

1. *for every continuation κ, if κ is a continuation of e, $\kappa \in K_e$.*
2. *for every proper subevent e_s of e, if K_s is the continuation set of e_s, then for every $\kappa_s \in K_s$, $\kappa_s \in K_e$.*

The second clause is expendable if transitivity of the subevent relation is assumed. From the vantage of classical mereology, this assumption is quite plausible, but there may be some rationale for dropping it in the case of processes [11].[8] But even if one adopts such a view, it should be possible to claim that there can be interruptions or delays during episodes that are not subevents in a restricted sense. With this kind of arcane subevent relation, the continuation set of e will contain more than just continuations of e. The definition of continuation sets is thus neutral with regard to this kind of ontological decision.

But the continuation set alone is not enough to capture a process in its entirety, because it is easy to observe that for the very end of the event there cannot be a continuation (clause 4b of definition 1 would be violated). One has thus to take into account the final state of the entire event:

DEFINITION 3. *Let e be an event, K_e the continuation set of e and s the final state of e. Then $\langle K_e, s\rangle$ is the event description of e.*

The notion of an event description for individual events can then be used to formulate class-level definitions of event types, by specifying *continuation signatures* that characterise types of events:

DEFINITION 4. *$\langle \Sigma, S\rangle$ is a* continuation signature *iff*

1. *Σ is a set of continuation types*
2. *S is a state type*
3. *there exists some s, κ_1,... κ_n such that*
   a. *s is an instance of S*
   b. *The instantiation relation maps κ_1,... κ_n to exactly one element of Σ*
   c. *$\langle\{\kappa_1, ...\kappa_n\}, s\rangle$ the event description of some event.*

DEFINITION 5. *Let e be an event, E an event type and $\langle \Sigma, S\rangle$ the continuation signature of E. e is an instance of E iff*

1. *there exists an event description $\langle K_e, s\rangle$ of e such that*
   a. *s is an instance of S and*
   b. *for every $\kappa \in K_e$, κ is an instance of some element of Σ*
   c. *for every type $T \in \Sigma$ there is an instance of T in K_e.*

In this view, event types are distinguished not only by their instances bring about, but also by how they bring it about. They are thus stricly linear; variance in events, as is caused by conditional or alternative subevents, would thus require additional aggregation of event types.

3.3 Formalising Delays and Interruptions with Process Continuations

I will further claim that event types, defined by continuation signatures, are invariant to structural anomalies. This claim will be substantiated by showing that (a) both normal and anomalous tokens of an event class have the same final state s and (b) the continuation set of the class is not modified by anomalies.

(a) is quite trivial to show but it also leads to the conclusion that abortions of events cannot be captured with the present modeling scheme because the final state s never obtains if the event is aborted prior to its normal termination.

[7] The dependence relation might be a generic one, though. For example, a game of chess depends on a certain set of chess pieces at every stage of the game. But for the game to continue, it is not necessary that the pieces involved remain numerically identical. I can very well continue playing the game if I replace one white pawn with a different one, provided that I place it in the correct position.

[8] For example, one might wish to claim that depressing the accelerator pedal is a subevent of driving a car, and that moving the foot down is a subevent of pushing the accelerator pedal, but that moving the foot down is not a subevent of driving a car – obeying the intuition that depressing an accelerator pedal is in a strong sense 'part' of driving a car, while foot movement is not. I do not, however, hold any strong opinions on the matter.

That the continuation signature of E is neither affected by delays nor by interruptions will become clear once we have given proper definitions of these anomalies.

DEFINITION 6. *Let e be an event of type E and K_e the continuation set of e. The proper subevent e_d of e is a delay of e iff*

1. *e_d is a proper subevent of a delay of e.*
2. *or*
   a. *e_d is temporally contiguous.*
   b. *there exists $\kappa \in K_e$ such that κ exists at least as long as e_d lasts*
   c. *the right boundaries of e_d and of κ's life-time coincide.*

This definition does justice to the intuition that delays contribute to the overall duration of a process. The episode e_d is part of the overarching event, but it does not contribute anything to advancing the normal course of events, because the continued existence of the continuation for the next 'true' subevent of the process requires that all participants and the final state of the previous 'true' subevent also continue to exist, and hence no changes relevant to the process can occur.[9] Also, defining delays this way does not change the continuation set of the event and since no special assumption about the event were made, the continuation signature can not be affected either.

DEFINITION 7. *Let e be an event of class E and K_e the continuation set of e. The event e_i is an interruption of e iff*

1. *e_i is a proper subevent of an interruption of e.*
2. *or*
   a. *e_i is temporally contiguous.*
   b. *the left and right boundaries of e_i lie between the left and right boundaries of e.*
   c. *the temporal extensions of e and e_i do not overlap.*
   d. *there exists a continuation κ and a state of affairs s, such that*
      (1) *$\kappa \in K_e$*
      (2) *s is the final state of e_i.*
      (3) *s is the context state of κ.*
      (4) *κ existed at or before the left boundary of e_i.*
      (5) *a left boundary of an episode of κ's life-time coincides with the right boundary of e_i.*

This definition is a bit more complex due to the fact that it needs to account for the intuition that interruptions do not contribute to the overall duration of the process. It basically assumes that an interruption is something that fills a 'gap' in the process. Interruptions further differ from delays in that the necessary prerequisites for continuing the process are not present during the interruption. Consequentially, neither can a continuation be present during the interruption. That continuation is rather present sometime before the interruption (most likely at its left boundary) and it reappears once the prerequisites for continuing the process have been reestablished. In this respect, continuations turn out to be a bit awkward, but not any more awkward than ordinary objects that exist only intermittently [12, pp. 195–199], for example a table that is disassembled before it is moved to another room where it is reassembled. Likewise, the continuation will be the same continuation when it 'reappears' and no change to the continuation set needs to be made in order to accommodate interruptions. Again, the same argument as with delays reveals that the continuation signature will also stay the same, thus allowing the interrupted event to be subsumed under the same event type as the event modulo interruption.

[9] Readers should note that in PATO the process quality *delayed* is not attached to the superprocess that experiences the delay, but the subprocess immediately succeeding the delay.

4 CONCLUSION

I believe that the process continuation model of delays and interruptions presented in this paper can be used to provide compelling formal definitions for certain process related phenotypes that correspond to structural anomalies of these processes. It also provides a better understanding of the normality–abnormality divide in the category of occurrents.

It should, however, be noted that such definitions are usually not needed for some of the tasks that phenotype ontologies are usually employed for, e.g. for establishing genotype–phenotype mappings. These use cases usually do not require further analysis of the anomalies. But while these applications do not benefit from using explicit definitions for anomalies, they might prove much more fruitful for other use cases, for example in applications that try to detect such anomalies in datasets.

The approach sketched here also has a few blind spots that provide interesting avenues for future research. For example, it is strictly not possible to arrive at a satisfactory understanding of abortions or missing process parts using continuation sets. This is because it by definition treats processes as *complete* processes and the removal of subevents from these processes does not leave continuation sets unscathed. To me, this suggests that this kind of anomaly does not fall into the same category as interruptions and delays.

Another interesting point to consider is that some kinds of delays seem to lack the discrete nature that is required for definitions presented here. Think, for example, of a train that is being delayed not because it is stalled at some station for a specific period of time, but rather because it can only travel at reduced speed. Here the delay is continuously accumulated while the proper process remains in effect.

Apart from these things, there is the latent issue of proper treatment of all temporal phenomena in ontologies. I have tried to avoid this issue here by giving the general requirements of my approach with regard to temporal modeling. But still a sensible and generally agreeable scheme for dealing with time and occurrent entities remains a considerable desideratum for all ontology modeling.

ACKNOWLEDGEMENTS

Thanks to the participants of the Rostock workshop 'Collectives in Space and Time' (June 23rd–25th, 2011) for their insightful and stimulating comments on an earlier version of the paper, to Ludger Jansen and Johannes Röhl, Rostock, for their helpful remarks on the second draft and three anonymous referees for their extraordinarily learned comments. This work is supported by the German Science

Foundation (DFG) as part of the research project 'Good Ontology Design' (GoodOD).

REFERENCES

[1]David M. Armstrong. *A World of States of Affairs*. Cambridge: Cambridge University Press, 1997.

[2]Antony Galton and Riichiro Mizoguchi. 'The water falls but the waterfall does not fall: New perspectives on objects, processes and events'. In: *Applied Ontology* 4 (2 2009) pp. 71–107.

[3]Pierre Grenon. *BFO in a Nutshell: A Bi-categorial Axiomatization for BFO and Comparison with DOLCE.* IFOMIS Report 06/03. Leipzig 2003.

[4]Heinrich Herre et al. *General Formal Ontology (GFO). Part I: Basic Principles*. Onto-Med Report 8. Leipzig 2006.

[5]Robert Hoehndorf et al. 'Representing default knowledge in biomedical ontologies: application to the integration of anatomy and phenotype ontologies'. In: *BMC Bioinformatics* 8 (377 2007) DOI: 10.1186/1471-2105-8-377.

[6]David K. Lewis and Stephanie Lewis. 'Holes'. In: *Australasian Journal of Philosophy* 48 (1970) pp. 206–212.

[7]Chris Mortensen. 'Change and Inconsistency'. In: *The Stanford Encyclopedia of Philosophy*. Ed. by Edward N. Zalta. 2011. URL: http://plato.stanford.edu/archives/fall2011/entries/change/.

[8]Christopher J. Mungall et al. 'Integrating phenotype ontologies across multiple species'. In: *Genome Biology* 11 (2010) R2. DOI: 10.1186/gb-2010-11-1-r2.

[9]John C. Reynolds. 'The Discovery of Continuations'. In: *Lisp and Symbolic Computation* 6 (3/4 1993) pp. 233–248.

[10]Peter N. Robinson et al. 'The Human Phenotype Ontology: a tool for annotating and analyzing human hereditary disease'. In: *American Journal of Human Genetics* 83 (5 2008) pp. 610–615. DOI: 10.1016/j.ajhg.2008.09.017.

[11]Johanna Seibt. 'Free Process Theory: Towards a Typology of Occurrings'. In: *Axiomathes* 14 (2004) pp. 23–55. DOI: 10.1023/B:AXIO.0000006787.28366.d7.

[12]Peter Simons. *Parts. A Study in Ontology*. Oxford: Clarendon Press, 1987.

[13]Cynthia L. Smith, Carroll-Ann W. Goldsmith and Janan T. Eppig. 'The Mammalian Phenotype Ontology as a tool for annotating, analyzing and comparing phenotypic information'. In: *Genome Biology* 6 (2005) R7. DOI: 10.1186/gb-2004-6-1-r7.

[14]Gerald Jay Sussman and Guy L. Steele Jr. 'Scheme: A Interpreter for Extended Lambda Calculus'. In: *Higher-Order and Symbolic Computation* 11 (4 1998) pp. 405–439. DOI: 10.1023/A:1010035624696.

[15]Christopher P. Wadsworth. 'Continuations Revisited'. In: *Higher-Order and Symbolic Computation* 13 (1-2 2000) pp. 131–133. DOI: 10.1023/A:1010074329461.

The Ten Commandments of Ontological Engineering

Ludger Jansen[1,2,*], *Stefan Schulz*[3,4]

[1] *Institute of Philosophy, University of Rostock, Germany*
[2] *Philosophical Institute, RWTH Aachen University, Germany*
[3] *Institute of Medical Biometry and Medical Informatics, Freiburg University Medical Center Freiburg, Germany*
[4] *Institute for Medical Informatics, Statistics and Documentation, Medical University of Graz, Austria*

ABSTRACT

The realist approach to ontology design has recently been criticized for being idle philosophical red herring, without any advice for ontology developers. We present guidelines for ontology design in the realist spirit and demonstrate not only how these guidelines are motivated from an realist understanding of ontologies, but also indicate how they are thought to improve the performance of scientific ontologies, especially in biomedicine and the life sciences.

1 WHAT IS AN ONTOLOGY?

Though the realist account for ontological engineering has by now been described on text book level (Jansen/Smith 2008; Munn/Smith 2008), there is still considerable controversy about the impact and legitimacy of its philosophical background (Merrill 2010a, 2010b; Maojo et al. 2011), and what the methodology consists of in the first place.

In the present paper, we try to summarize the realist paradigm in ten easy rules or commandments. We do not aim to present any new prescriptions; we rather try to compile and condense ontology design rules that have hitherto been presented at scattered places. We will argue that the way how they fit into the realist account follows from the realist conception of what an ontology is. As a result it will follow that realism, as argued for in this paper, is not an ideology, but an engineering approach for designing good ontologies which have a higher probability to be sustainable, interoperable, and adequate to their domain.

Since the term 'ontology' came into use in information science, it has been used for a variety of information artefacts. Both intension and extension of the term is debated. By different scientists, ontologies are said to be: (i) information artefacts, (ii) representations (e.g., of conceptualizations), (iii) formal structures, (iv) theories, (v) hierarchies of types or universals. If we take a closer look at these competing characterisations, they turn out to be compatible, as they describe different aspects of ontologies. Ontologies can, e.g., be representations and information artefacts at the same time, as the latter are nothing but artificial representations. Those who describe ontologies as formal structures, state the way how ontologies represent what they represent. Those who describe ontologies as theories, refer to what is the benchmark of what we represent, i.e. some theory about a certain domain (and many times it will be the best theory for that domain available to us). Those who, finally, describe an ontology as a hierarchy of types, universals, or classes, state what is represented in an ontology and what it is the benchmarking theories are theories of.

As these characterizations of ontologies are complementary, we can put them all together. Hence, an **ontology** is an artificial representation (an information artefact, that is), that represents types or universals of a certain domain and the relations that hold according to a certain theory in a formal structure. More specifically, **scientific ontologies** are information artefacts that represent types of things and their relations of a certain domain according to the best available scientific theory, in order to support knowledge storage, processing and eliciting in the sciences. An immediate consequence of this definition is that an information artefact that is not intended to be useful for science cannot be considered to be a scientific ontology.

We will now present our set of ten commandments. As in the biblical paradigm, we start with three 'stage-setting' commandments that characterize our general approach, while the remaining seven commandments will lay out the details for the design of classes and relations within an ontology.

2 REALISM , MULTI-PERSPECTIVALISM, AND ADEQUATISM

I. You shall be a realist.

Realism, in general, is the thesis that there is a mind-independent world. Scientific discourse is, indeed, only possible if there is a consensus about the observer-independent existence of entities, which can be described in terms of invariant properties (e.g., having a volume, a size, a chemical constitution etc.). This assumption is pragmatically useful even if we cast some doubt on humans' ability to perceive and understand the objects in the world in their totality.

There is quite a lot of disagreement concerning the content of ontologies (cf. Table 1): Some define ontologies as representations of concepts, others as representations of non-conceptual entities. Even lexical representations of

words or terms are found to be called ontologies, though the latter would more appropriately be called terminologies. However, ontologies and terminologies may look much the same, as they both come along as structured lists of terms. The crucial difference is whether these terms are intended to represent themselves or something else. In terminologies, terms are *mentioned*, whereas in ontologies they are *used*. If a term is mentioned, it functions as a representation of the type of word that is instantiated by the word token used; it has, as medieval logicians called it, material supposition. Terminologies, that is, are inventories of linguistic entities, whereas a term when it is being used (and not only mentioned) normally represents some type of non-linguistic entities. In ontologies, as opposed to terminologies, words are used and not only mentioned. Thus, as a rule, ontologies are inventories of extra-linguistic entities (although there could also be specific ontologies of linguistic types). Thus in ontologies, both natural language labels and terms are only instruments for the representation of other things.

Table 1. Competing Definitions of Ontologies

Representation of Words
'An ontology is a machine-readable representation of a domain's terminology and the relationships among the terms in the domain.' (White2011)
Representation of Concepts
'an explicit representation of a conceptualization' (Gruber 1993)
'an Ontology gathers a set of concepts that are considered relevant to a given domain' (Velardi et al. 2001)
Representation of Entities
'representation of some pre-existing domain of reality' (Rodriguez 2007)
'representational artifacts whose representational units are intended to designate classes or types in reality and to relate them to each other' (Schulz/Johansson 2007)

Now how do concepts fit into this? What would a representation of concepts (a 'conceptology'?) be? As often, the word 'concept" is used with much liberty, and not very uniformly. Sometimes it is used interchangeably with 'word' or 'term'. A representation of concepts in this sense would simply be a terminology. Sometimes, however, the word 'concept' is used either for a group of words that share (nearly) the same meaning or for the meaning of these words itself (Klein/Smith 2010). 'Mumps' and 'Parotitis epidemica' are different words, but they have the same meaning in a given domain; they can be said to represent the same concept. But that these words have the same meaning means nothing but that they represent the same type of entities.But then we have to represent types of entities again, and a concept representation would collapse into an ontology. However this will be resolved by conceptualists, the essential point of Realism is that the goal of an ontology is not the representation of words or meaning-entities (be they mental, social or abstract), but that it aims at a representation of the world. Our cognitive and linguistic apparatus is very important for this endeavour. It is, however, only of instrumental importance and not itself the object of representation within an ontology.

II. You shall not make for yourself an idol, whether in the form of a certain perspective or application.

The world that is investigated by science in general (and the biomedical world from which we borrow our examples in particular) is very rich. There are subatomic particles, atoms, molecules, organelles, cells, organs and organisms, there are packs, herds and swarms, habitats and ecosystems. There are physical and mental events, there are natural and social entities. Any of these levels can become important for the representation of biomedical knowledge. In particular, some relations or predicates are granularity-dependent and do not translate into statements of other levels of granularity. E.g., the statement that the right kidney touches the liver presupposes the granularity level of organs; it is not equivalent to, say, the statement that some molecules in the right kidney touch some molecules in the liver.

This implies that the realism advocated for in the first commandment is not dogmatic in that it is not committed to monopolizing one perspective (or even a non-perspective). The point is rather to explain how various perspectives interrelate to each other. Nor do realists have to claim that they are already in the possession of the ultimate truth: Ontologies can and must be revised as science progresses. Realists can also easily admit cultural influences on the context of discovery of scientific theories, because realism is not a theory about the discovery of theories but about their meaning and validity. Realists can also admit the mind-dependence of many entities, although they will reject (like Searle 1995) any claim to the effect that all entities are mind-dependent; they can also allow for 'fiat' boundaries (Smith/Varzi 2000) and for vague boundaries (Bittner/Smith 2001; Schulz/Johansson 2007).

III. Remember your domain and keep it holy.

Ontologies have to be adequate to their domains, and domains come along on different granularity levels. Despite this richness of the world, there is a strong tendency for Smallism and Reductionism. Smallism is the tendency to prefer smaller entities and negatively discriminate against larger and complex entities (Wilson 2004); Reductionism is the attempt to eliminate (talk about) higher levels in favour of (talk about) lower levels. But if we talk only about, say, atoms, a lot of information about the higher levels is lost which is of utmost importance for the life sciences. It is impossible to translate 'An appendectomy is a surgical removal' or 'Democracy is a form of government' into talk about atoms and their movement. If we want to do fully justice to

our domain, we need to account for all levels that are relevant for a domain. Thus all of these things are to be counted in when we are to set up the inventory of the world.

3 TERMS, DEFINITIONS, AND RELATIONS

IV. You shall not make wrongful use of names.

Even if words are not the normal object of representation in an ontology, words are important for the ontologist as instruments to refer to the types and classes to be represented. Names of representational units in an ontology (be they names for classes or for relations) are ideally unambiguous and self-explaining. To this end, naming conventions have been proposed (Schober et al. 2009). Words and terms used in practice tend to be ambiguous. Their makeup is not always fully compositional; figurative use is common, and literal interpretations are often misleading: A complicated pregnancy is a pregnancy, but a prevented pregnancy isn't, as little as a planned biopsy is a biopsy, or a suspected asthma a kind of asthma. Such idioms should be avoided, as they bear the risk of incorrect subclass assertions such as 'SuspectedAsthma subClassOf Asthma'

The ambiguity of many terms often remains unrecognized, especially if there is one 'standard' reading such as in 'Diabetes mellitus' (frequent) and 'Diabetes insipidus' (rare). 'Diabetes' (without modifier) is vastly used in the sense 'mellitus', and few doctors are aware of the ambiguity. Furthermore, acronyms abound in biology and medicine, and most of them are ambiguous. 'CT' generally means 'X-ray computed tomography', but competing readings such as 'connective tissue' coexist.

To avoid ambiguity, class or relation labels in good ontologies are often somewhat artificial and not commonly used in oral and written communication. Yet this is a price worth to be paid for clarity. Ontology labels are not supposed to provide the lexical basis for text mining or information extraction systems. For these and other use cases, terminologies or vocabularies need to be linked to the ontology as external resources. Homonymy and synonymy are linguistic phenomena that should be addressed by language resources and not by ontologies.

V. For most of your terms you shall provide unambiguous definitions.

The first requirement is, plainly, that definitions are given at all. This is important as well for the human user as for automated processing of the ontology. Often, the human user needs additional information to figure out which of several possible meanings the term used is intended to have. And as the computer has no prior knowledge about term meanings at all, definitions are the best way to provide this information. The second requirement of this definition is, then, to define in the proper way, for otherwise the definition will not fulfil these purposes. A well put definition will (i) not confuse the definition with information about the etymology of the term, its usage etc., (ii) take care that the defined term (the definiendum) fits to the definition given (the definiens), and (iii) be internally consistent and, e.g., not try to cover homonyms within one and the same definition. The third requirement is, finally, to take one's own definitions seriously and use the defined term according to the definition given and only according to the definition given. Should the desire arise to use a term in several meanings within an ontology, different classes with different class names have to be asserted, each with its specific definition. .

Note that this commandment cannot be extended to all terms in ontologies. First, the highest classes cannot be defined for lack of appropriate superclasses and specific characterizations. And second, especially in the bio-medical field it will often be difficult to find full definitions, and the ontologist must be contend to name some superclasses as necessary conditions for the application of the respective class term.

VI. You shall use a top-level ontology.

Top-level terms are highly abstract and often denote philosophically derived categories like 'continuant', 'occurrent', 'dependent entity' or 'independent entity' in BFO, 'perdurant' or 'endurant' in DOLCE and others. Nevertheless, they are highly important for ontology engineering.

First, they force ontology developers to think about ambiguous terms. And second, they support ontology mapping. For instance, 'tumour' can denote a physical entity (a mass of tissue), just as a malignant process a patient is suffering from. These are distinctly different: as a material entity, a tumour has a size and a weight, whereas a malignant disease has a duration. Similarly, 'Allergy' can denote both an allergic disposition in a healthy patient (who has no signs of disease as long as the allergen is absent) and an allergic reaction. Both things are distinctly different. A physician treats a patient with an allergic disposition differently from a patient with a manifest allergic reaction. Only clear-cut top-level categories are able to support disambiguation of terms and expressions such as explained in IV. Upper level ontologies also enforce a distinction between real entities and information entities. 'Gender: unknown' refers to an information entity (e.g. a medical record) on the gender of a patient. 'Gender male' refers to the really existing gender quality of a person.

Second, top-level categories support ontology mapping, as these are a relatively small number of terms and they are both domain-independent and domain-transcendent. Domains can vary widely with regard to the types that feature therein, but if ontologies subsume their types under the domain-independent top-level categories, this is one clear indication for their interrelations to start with.

Finally, upper-level ontologies can enforce type constraints, provided that the upper-level categories are disjoint. For instance, if the relation **inheresIn** has its range constrained to material objects, any assertion on the inherence of something in a different category, e.g., a process would render the ontology inconsistent.

VII. Honor interoperability, so that your ontology has a long life and can be re-used by others.

Most of the rules discussed here make good sense, even if your ontology were the only ontology in the whole known universe. But it isn't. There are a growing number of information artefacts that claim to be ontologies and are called by that name. Setting up a good ontology is difficult, time consuming and expensive. Thus it is desirable that ontologies can be re-used or used in connection with other ontologies for different sub-domains. Good naming conventions, unambiguous definitions and the use of a (common) top-level ontology are already good means to enhance interoperability. But keeping an eye on interoperability has also an influence on the choice of your types and on their definitions. For short, good ontologies should avoid eclecticism and parochialism:

Avoid Eclecticism. Eclectics go along and pick the things that suit them out of whatever system they encounter. In ontology design, eclecticism concerns the choice of types in an ontology. These types have, for sure, to be chosen wisely. An ontology usually has the task to represent one domain only, and there are pragmatic restrictions for its size. One variety of eclecticism is national bias. For example, in the National Cancer Institute Thesaurus (NCIT)[1], *American Indian or Alaska Native* is the only sub-type of *Underrepresented Minority,* although there are, of course many more 'minority groups presently underrepresented in biomedical and behavioural research', as the NCIT defines the latter type. Many types in the NCIT are, taken literally, ambiguous, and seem to contain a tacit reference to the US. E.g., there is nothing like an *Underrepresented Minority* as such: A group can be underrepresented in America, while being overrepresented in, say, China. A minority can be underrepresented in American cancer research, but overrepresented in, say, the patients of St Louis Hospital. This is an obvious hindrance for the integration of cancer research done elsewhere and very much annoying for cancer researchers using the NCIT outside the US.

Avoid Parochialism. While eclecticism is about the choice of types, parochialism is about definitions and other term properties. Parochial definitions are built on the assumption that the domain to be covered or the entities described in the ontology itself are all there is in the world. This erroneous assumption is, as a rule, not only false, it is also a hindrance to the understanding of human users and to interoperability. The NCIT, e.g., uses *Clinical Study* synonymous with *Study*, though there are, of course, plenty varieties of non-clinical studies. The NCIT also subsumes *Action* under *Clinical or Research Activity*, although a lot of 'thing[s] done' (thus the definition of *Action* in the NCIT) are done outside of research business. *Underrepresented Minority*, again, is defined as a group 'underrepresented in biomedical and behavioural research' (while there are plenty of other fields where a group can be underrepresented), and *Funding* is subsumed under the semantic type *Governmental or Regulatory Activity*, leaving out companies, charities or endowments as potential money-givers.

VIII. You shall not confuse ontology and epistemology.

Epistemology describes what an agent sees, knows or records about a domain, whereas ontologies, ideally, describe the entities in that domain as they exist. Especially in medicine, the reference to a term or an ontology class does often not mean that a related individual exists. Clinical decisions are often triggered by mere suspicions, due to lack of time or resources. E.g., children who exhibit a clinical picture suspicious for meningitis are treated as if they had meningitis. As Bodenreider et al. (2004) and Ingenerf and Linder (2009) underline, legacy medical classification systems such as ICD are 'infiltrated' by epistemic notions, such as in classes like 'Tuberculosis of lung nodular bacteriological or histological examination not done' in ICD-9-CM.

It is a repeatedly expressed desideratum that epistemic aspects should be treated in information models. Yet there are overlaps between ontologies and information models which give rise to conflicting representations, requiring sophisticated mitigation strategies (Cheetham et al. 2009). The very same complex information (e.g., a clinician's hypothesis of a stenosis of the left carotid artery) can be represented to different proportions in clinical ontologies and clinical information models and creates interoperability problems (Garde et al. 2007). A way out of this dilemma could be to represent epistemic aspects or recorded information in a separate branch of the ontology (Schulz et al. 2011).

IX. You shall not produce ontology artefacts.

Ontologies are systems of types, the essential features of which are given by text or (better) formal definitions, and whose instances are similar in that they share these essential features. However, not all classes, which are formally construable using logic, are the extensions of types. Dog is a type, because all its instances are similar in that they are not dogs. Non-Dog, however, is not a type is this sense, because there is nothing that its instances share and that distinguish them from the instances of Dog. Classification systems, however, often introduce such classes in order to artificially

[1] http://ncit.nci.nih.gov/ncitbrowser/

produce exhaustive partitions, such as in ICD-10, where Angina pectoris (I20) has as subtypes:

> *Unstable angina* (I20.0),
> *Angina pectoris with documented spasm* (I20.1);
> *Other forms of angina pectoris* (I20.8), and
> *Angina pectoris, unspecified* (I20.9).

Of these, I20.8 refers to the logical complement to the union of I20.0 and I20.1; it is thus only defined negatively. Such artefacts like I20.8 and I20.9 are needed for, e.g., statistical purposes, but they should not be contained in ontologies proper.

X. You shall not covet to use relations with ambiguous semantics, or at any place where are do not apply.

The development of ontologies has been triggered by the desire to collect more information about a domain than is normally contained in a terminology. For this purpose, terminologies were more and more enriched with additional information by encoding interrelations between the entries of a terminology. We must, however, bear in mind that such information is to be understandable by the human user and processable by a computer. Ontology developers should, therefore, respect the semantics of formal relations. That is: Use different relations if you want to say different things.

Avoid isA-overload. In some cases, natural language is a good guide to distinguish between formal relations, as differences between formal relations often become obvious when the respective statements are spelt out: *Thumb* isA *Finger*, but it is not the case that *Finger* isA *Hand*; rather *Finger* partOf *Hand.* Sometimes, however, we use the same word in natural language (often the copula "is") to express quite different formal relations. We say, for example, both 'Fish is an animal' and 'Fish is a food'. But only the first sentence expresses a subtype relation: '*Fish* isA *Animal*' is true because any and every fish is necessarily an animal. A fish cannot cease to be an animal without ceasing to exist. To be food, however, is not necessary for being a fish. To serve as a food is a role that is played by some fishes only, and only contingently so. Thus being *Food* is only a *Role* played by some fish, and only contingently so: Many a fish is never been eaten; and in those cases, where a fish is been eaten, it comes along with the end of its existence as a fish. This fact also indicates a difference in subject, as the latter sentence is naturally construed to contain 'fish' as a mass noun. Using the language of OntoClean, we also say that being fish is a rigid property, whereas being food is a non-rigid property (Guarino 2009).

Be aware of implicit quantifiers. Ontologies involve statements about types of entities. The semantics of relational statements between types of entities normally involve (implicit or explicit) quantifiers that range over instances of these types. Often, formal relations have an all-some semantics: '*Human* hasPart *Head*' is true, if and only if for every instance *x* of the type *Human* there is at least one instance *y* of the type *Head*, such that *y* is part of *x*. The all-some structure is asymmetric, which leads to the effect that relations that are converse on the instance level are not converse on the type level. While 'Human #10 hasPart Head #12' is true if and only if 'Head #12 partOf Human #10', this does not work for types of entities: '*Head* partOf *Human*' is false, for there are plenty of heads that have horses or cats as their possessors. Similar, '*Uterus* partOf *Mammal*' is true, for every instance of the type *Uterus* is part of some mammal, but '*Mammal* hasPart *Uterus*' is false, for many instances of the type *Mammal* do not have an uterus, for example all males.

Use singular numerus. It is with respect to the relations used that ontology labels should be used in their singular form. To use, e.g., the relational phrase 'is a' with a plural term, is in most cases wrong for grammatical reasons, let alone for the asserted formal semantics of the relation.

All these rules are important for the realist, because they avoid nonsense or logical error. As a realist ontology is intended to represent features of the world, both logical error and nonsense statements are unwanted guest to be expelled from the ontology. Moreover, logical contradictions proof to be fatal for automated processing with, e.g., a reasoning programme, because a contradiction entails any statement, including all false statements – which, again, contravenes realist ambitions.

4 DISCUSSION

After having stated our 'Ten Commandments for Ontological Engineering', we can now summarize them by pointing out their contribution to a better performance of the ontologies created with their help:

> Commandments enhancing sustainability:
> II, IV, V, VIII, IX, X
> Commandments enhancing interoperability:
> I, II, IV, V, VI, VII, IX, X
> Commandments enhancing adequacy:
> I, II, III, V, VIII, IX, X.

The commandments presented here are not all of the same standing. As in the biblical prototype, the first three commandments set out the background assumptions and the habit of mind that are to guide ontology development, without itself prescribing any concrete actions. This task is fulfilled by the remaining seven commandments which fill in the details. It might be objected that these latter advices are independent from the realistic stance as they are 'just good ontology design guidelines' that serve their purpose also when the designer does not embrace realism. But this is not

in contradiction to our case, as can be explained by carrying the analogy to the biblical prototype further: The concrete prescriptions are motivated by the realism, multi-perspectivalism and adequatism outlined in the first three commandments (as the biblical commandments are motivated by theism). But the prescriptions may perchance be also observed by ontology designers of other than realist observation (as also an atheist can honour his parents and refrain from murder). The point is not that you have to be a realist to follow these guidelines (you do not have to be a theist to honour your parents), but that realism implies these advices (if you are a theist, you are to honour your parents). In setting out these guidelines we thus show that realism has a bearing on ontology design by motivating a coherent set of guidelines for ontology development.

Realism has recently been dubbed 'the 'philosophical' approach to the development of ontologies' (Haux 2011). It is true that realism is motivated by the thoughts of eminent philosophers, notably Aristotle (hence the honorific name 'Aristotelian approach'). It is false, however, that the realist approach is doomed to fail because Aristotle did not count his wife's teeth correctly and natural sciences outdated Aristotelian thinking (thus Maojo et al. 2011). It did not, simply because realism is not a scientific theory but a meta-scientific theory – the thesis that scientific knowledge is about the world. Although it may be considered to be an advantage that realism conforms to the ideas of so eminent a philosopher as Aristotle, it is false that this is the only advantage of realism. The presentation of our guidelines shows that realism and its implications are not only of aesthetic value because they conform to a certain philosophical 'ideology', but that they are also of instrumental value from the engineering point of view: They enhance readability by the human user, automated processing, inter-coder reliability and interoperability between different sub-domains. In short: Realism helps to make ontologies a much smarter tool for the use of science.

ACKNOWLEDGEMENTS

Work for this paper has been supported by the DFG grant JA 1904/2-1, SCHU 2515/1-1 GoodOD (Good Ontology Design). Many thanks to the anonymous referees for OBML for helpful comments and suggestions.

REFERENCES

Bittner T., Smith B. (2001) Granular Partitions and Vagueness. In: *Second International Conference on Formal Ontology in Information Systems, FOIS 2001. Proceedings*. ACM, 2001.

Bodenreider, O., Smith, B., Burgun, A. (2004) The Ontology-Epistemology Divide: A Case Study in Medical Terminology. *Proceedings of FOIS 2004*. Amsterdam: IOS.

Cheetham E. et al. (2009) Using SNOMED CT in HL7 Version 3; Implementation Guide, Release 1.5, http://www.hl7.org/v3ballot/html/infrastructure/terminfo/terminfo.html (9/9/2011).

Garde S., Knaup P., Hovenga E., Heard S. (2007) Towards semantic interoperability for electronic health records. *Methods of Information in Medicine* 46(3), 332–343.

Gruber, T.R. (1993) A Translational Approach to Portable Ontologie. *Knowledge Acquisition* 5, 199-220.

Guarino N. (2009) An Overview of OntoClean. *Handbook on Ontologies*, second edition, Springer, pp. 201-22.

Haux R. (2011) Editorial. *Methods of Information in Medicine* 50, 201-202.

Ingenerf J, Linder R. (2009) Assessing applicability of ontological principles to different types of biomedical vocabularies. *Methods of Information in Medicine* 48, 459-67.

Jansen L., Smith B., eds. (2008) *Biomedizinische Ontologie. Wissen strukturieren für den Informatik-Einsatz*, Zürich: vdf.

Klein G.O., Smith B. (2010) Concept Systems and Ontologies: Recommendations for Basic Terminology. *Transactions of the Japanese Society for Artificial Intelligence* 25(3), 433-441.

Maojo V. et al. (2011) Biomedical Ontologies: Toward Scientific Debate. *Methods of Information in Medicine* 50, 203-216.

Merrill, G. (2010a) Ontological Realism: Methodology or Misdirection. *Applied Ontology* 5(2), 79-108.

Merrill, G. (2010b) Realism and Reference Ontologies: Considerations, Reflections, and Problems. *Applied Ontology* 5, 189-221.

Munn K., Smith B., eds. (2008) *Applied Ontology. An Introduction*, Frankfurt am Main.: Ontos.

Rodrigues, J.M et al. (2007) A road from health care classifications and coding systems to biomedical ontology: The CEN categorial structure for terminologies of human anatomy: Catanat. In: K.A. Kuhn et al. (eds.), *MEDINFO 2007*, Amsterdam: IOS, vol. 1, 735–740.

Schober D. et al. (2009) Survey-based naming conventions for use in OBO Foundry ontology development. *BMC Bioinformatics* 10:125.

Schulz S., Cornet, R.; Spackman, K. (2011) Consolidating SNOMED CT's ontological commitment. *Applied Ontology* 6, 1-11.

Schulz S., Brochhausen M., Hoehndorf R. (2011) Higgs bosons, mars missions, and unicorn delusions: How to deal with terms of dubious reference in scientific ontologies. *3rd International Conference on Biomedical Ontologies (ICBO 2011), University at Buffalo, July 2011.*

Schulz S., Johansson I. (2007) Continua in Biological Systems. *The Monist* 90(4), 499-522.

Schulz S., Karlsson D. (2011) Records and situations. Integrating contextual aspects in clinical ontologies. *Bioontologies SIG 2011, Vienna, July 15-16, 2011.*

Velardi P., Missikoff M., Basili R. (2001) Identification of relevant terms to support the construction of Domain Ontologies. *ACL-EACL Workshop on Human Language Technologies*, http://acl.ldc.upenn.edu/W/W01/W01-1005.pdf (9/9/2011).

White, S. (2011) What the Heck is an Ontology?, http://www.catalysoft.com/articles/WhatIsAnOntology.html (9/9/2011).

Wilson, R.A. (2004) *Boundaries of the Mind. The Individual in the Fragile Sciences: Cognition*, Cambridge: Cambridge University Press.

Towards Improving Phenotype Representation in OWL

Frank Loebe [1]*, Frank Stumpf [1], Robert Hoehndorf [2] and Heinrich Herre [3]

[1]Department of Computer Science, University of Leipzig, Germany
[2]Department of Genetics, University of Cambridge, UK
[3]Institute of Medical Informatics, Statistics and Epidemiology (IMISE), University of Leipzig, Germany

ABSTRACT

Phenotype ontologies are used in species-specific databases for the annotation of mutagenesis experiments and to characterize human diseases. The Entity-Quality (EQ) formalism is a means to describe complex phenotypes based on one or more affected entities and a quality. EQ-based definitions have been developed for many phenotype ontologies, including the Human and Mammalian Phenotype ontologies. We analyze the OWL-based formalizations of complex phenotype descriptions based on the EQ model, identify several representational challenges and analyze potential solutions to address these challenges. In particular, we suggest a novel, role-based approach to represent *relational qualities* such as *Concentration of calcium in blood*, discuss its ontological foundation in the General Formal Ontology (GFO) and evaluate its representation in OWL and the benefits it can bring to the representation of phenotype annotations. Our analysis of OWL-based representation of phenotypes can contribute to improving consistency and expressiveness of formal phenotype descriptions.

1 INTRODUCTION

In recent years, molecular biology has made significant progress in understanding the mechanisms underlying human disease. Several studies investigate disease mechanisms in animals that serve as models for humans [30]. In particular, the targeted modification of the genetic markup of these organisms provides a powerful means to investigate the molecular mechanisms associated with heritable diseases in humans [8]. Large-scale mutagenesis projects are now underway with the aim to characterize the outcomes of null-mutations for every gene in an organism. The observable characteristics of these modified organisms (their phenotypes) are represented in model organism databases and can be utilized to suggest candidate genes for diseases for which no molecular origin is currently known [20].

To standardize the terminology used in describing phenotypes, multiple species-specific phenotype ontologies were developed. For example, the Mammalian Phenotype Ontology (MP) [33, 7] is used to characterize phenotypes in mice and other mammals, and the Worm Phenotype Ontology (WPO) [31] is used to characterize *C. elegans* phenotypes. The Human Phenotype Ontology (HPO) [29] describes phenotypes in humans and is applied for describing human diseases and individual patients.

To translate phenotypes across species and enable their comparison with human phenotypes and disease, a syntax for phenotype decompositions has been developed [5, 37, 26]. In this syntax, phenotypes are represented by a combination of a quality and one or more entities. The entities represent the entities that are affected by a phenotype and are either physiological processes and functions (from the Gene Ontology [2]) or anatomical structures as represented by species-specific anatomy ontologies. The Phenotypic Attribute and Trait Ontology (PATO) [9] is an ontology of qualities which is used to describe *how* an entity is affected within a phenotype. Entity-Quality (EQ) based specifications of phenotypes have been developed for several species-specific phenotype ontologies [26], including the HPO [29], MP [33, 6, 7], WPO [31], and others, thereby intergrating pre- and postcoordinated biomedical ontologies [32, 26].

Recently, mechanisms became available to enable the automated translation of phenotypes across different species [26, 20]. In these methods, ontologies are integrated through species-independent ontologies, and automated reasoning over the integrated ontologies enables the automated comparison of species-specific phenotype information across multiple species. This approach crucially relies on the formalization of phenotype information in ontologies and model organism databases. With the increasing application of ontologies for data analysis, improving the representation of phenotype ontologies has the potential to directly affect and advance scientific analyses and discoveries.

The EQ model is an important and widely used means for formalizing phenotype information in ontologies [4]. In greater detail, its main idea is to combine an 'entity class' (the E in EQ) from an anatomy or process ontology with a 'quality class' (the Q) from PATO. For example, the class *eye* (MA:000261 in the Mouse adult gross anatomy ontology (MA) [14]) as the E and the color *red* (PATO:0000322) for Q can be combined to form the class *Red eye*. The typical formal interpretation of EQ statements is that the combination refers to a specialization of the quality class Q such that it inheres in instances of the entity class E [26, p. 3],[25]. In the example, this yields the class *red that inheres in an eye* (cf. Fig. 1).

Relational qualities involve at least one additional entity besides E. In the semantics of EQ, a second entity can be attached to a quality via the relation `towards` [26, p. 3–5]. An example of this kind is the *concentration of iron in the spleen*, which can be formalized as a quality *concentration of* (PATO:0000033) inhering in *spleen* (MA:0000141) and connected via `towards` to *iron* (CHEBI:18248

Figure 1. EQ model. (Gray indicates the optional part for relational qualities.)

*to whom correspondence should be addressed

in the ontology of Chemical Entities of Biological Interest [21]), in order to define *abnormal spleen iron level* (MP:0008739).[1]

The term 'relational quality' as nowadays found in the bio-ontology community is typically used without further analysis, e.g., in [26] and, through [25], can be traced back to [27] where it seems to be meant synonymously with the more widely used term 'relation'. Notably, in the context of formal ontology, by 'relational qualities' sometimes constituents of particular relation instances are referred to (in contrast to the overall relation instances themselves), termed 'relational roles' in sect. 3.3.

While EQ descriptions characterize a phenotype, a related question pertains to the formalization of the *annotation* of organisms, genotypes and genes with EQ-based phenotype descriptions. In model organism databases such as the MGI database [6], genotypes like Add2^{tmLlp} (MGI:2149065) are annotated with a class like *abnormal spleen iron level* (MP:0008739). The intended meaning of this annotation is that organisms of a particular mouse strain that exhibit the described genotype (a targeted mutation of the *Add2* gene) within a specific environment will develop the *abnormal spleen iron level* phenotype. This complex relation can be simplified to improve performance of specific information retrieval tasks into a view in which the genotype is equivalent to the intersection of phenotypes and individual mice instances of their phenotypic annotations.

Only few efforts formally explore the compositional nature of phenotypes, i.e., how atomic phenotypes can be combined into more complex phenotypes such as in disease descriptions or in genotypes annotated with multiple phenotypes. In particular, the naive combination of phenotypes such as *red eye* with *short tail* is based on class intersections, and these lead to contradictory class definitions due to the disjointness of *color* (the super-class of *red*) and *size* (the super-class of *short*) [19]. More challenging are combinations of qualities which are hidden in the taxonomy of biomedical ontologies. For example, asserting that *red eye* is a sub-class of an *abnormal eye morphology* will imply that *red eye* is both a subclass of *morphology* and *color*. This will lead to another contradictory class definition due to the disjointness of *color* and *morphology* [18].

2 REPRESENTING PHENOTYPES IN OWL

2.1 Basic Problems

We see three *basic problems* that need to be addressed regarding the representation of phenotypes and the interpretation of EQ descriptions in terms of the Web Ontology Language (OWL) [36], in order to utilize automated and semantically correct reasoning to its full extent.

I. ontological foundation of complex phenotypes
II. representation of phenotypes in formal languages
III. ontological foundation of phenotype annotations

The first problem concerns the ontological foundation of complex phenotypes. To address this problem, we attempt to gain a clear understanding of the ontological nature of complex phenotypes and rely on an ontological framework for the explanation and foundation of complex phenotypes which does not depend on the expressive power of OWL. Once we obtained an understanding of the ontological nature of complex phenotypes, we investigate how to represent them in OWL, as a case of the second problem. The next step is to apply this theory to existing descriptions of complex phenotype, such as those found in phenotypic annotation of diseases and genotypes in model organism annotations.

2.2 Issues of Formal Representation

The first basic problem requires further attention, but is widely discussed in biomedicine and formal ontology, e.g. see [24, 19, 35]. In the present paper, our focus is on the second problem and its application to formalizing phenotype annotations. In this regard we identify five interrelated *particular issues* that affect our analyses.

1. ontological adequacy / coherence of ontological interpretation
2. invalid permutations / ambiguities
3. relational expressiveness
4. consistency of domain modeling
5. formal reflection of annotations

Referring to *ontological adequacy*, we intend to find OWL representations that are close to the ontological understanding of phenotypes as *qualities*, similar to established ontological theories of phenotypes [25, 26].

While several approaches allow for representations of individual EQ statements in OWL, combining multiple EQ statements by means of their intersections may create incorrect [19, sect. 4.2, p. 3117] and sometimes contradictory statements [18]. For instance, consider the following OWL concept:

```
(red that inheresIn some eye) and
(short that inheresIn some tail)
```
(1)

Concept (1) is necessarily empty, because no instance of `red` is equally an instance of `short`. Furthermore, this formalization faces the problem of *permutations* (issue two), arising from the commutativity and associativity of intersections in OWL. In particular, the parentheses in example (1) are merely auxiliary for reading. The concept is formally equivalent to `(red that inheresIn some tail) and (short that inheresIn some eye)`. As a consequence, queries will deliver incorrect results if this mode of combining EQ statements is used.

The next two issues concern primarily phenotypes based on relational properties, like *iron concentration in the spleen*. *Relational expressiveness* is used for referring to limitations of the arity of relations that can be specified with an EQ description. The current model does not allow for relational qualities of an arity greater than two. This may lead to undesirable consequences, since several applications of biomedical knowledge representation require relations of higher arity [34, 10]. This issue has been identified as a particularly important challenge for representing EQ-based phenotypes [25]. Closely connected to the number of arguments is the question of *inter-modeler consistency / harmonization*, cf. also [10]. This fourth issue refers to the question of how to link (a class representing) a relation to (classes of) its arguments such that it is as unambigious as possible which argument connects to the relation in which way. In the current EQ model confusion can arise,

[1] Despite continued use of this example, we will not go into detailed ontological analyses of the relationship between iron and spleen, e.g., as particulars / individuals. In particular, iron as an amount of matter / quantity or collection would deserve special treatment, cf. e.g. [12, 13].

e.g., on whether *calcium concentration of blood* should be formalized as `concentration that inheresIn some blood and towards some calcium` or instead as `concentration that inheresIn some calcium and towards some blood`. The different positions may correlate with the community/background of modelers, e.g. whether a biologist or a chemist makes the assertion. Corresponding decisions are not only relevant for formalization, but likewise influence querying. For the particular case of concentrations, [13] proposes inherence in those entities that are concentrated in another in the context of an ontological analysis, i.e., inherence in calcium in the example. We comment on this in sect. 4, with hindsight regarding our analysis.

The fifth and final issue is the orientation and clarification of how *annotations* are interpreted, for any account of phenotype representations. This immediately links back to the ontological reading of phenotype representations and the third basic problem above.

3 ANALYSIS OF ALTERNATIVE APPROACHES

3.1 Spectrum of Solutions

In general, different approaches may be pursued in order to tackle the issues presented for the second basic problem. Like in [25], quality models that are fairly distinct from the EQ model may be (re-)considered. Another general change would be to concentrate on entities, i.e., primarily on the parts of an organism occurring in EQ descriptions, and to construct phenotype descriptions centering on them. E.g., the scheme E `that hasQuality some` Q follows this line of thought.[2]

We, however, focus first on solutions that limit the number of changes to the established interpretation of EQ descriptions. The latter are meanwhile widely in use, cf. e.g. [4], as are phenotype ontologies with their basic presupposition of providing (sub)concepts of *quality*. Therefore, the migration to new proposals should be facilitated by an approach with less changes compared to more radical revisions.

3.2 EQ Interpretations with regard to Annotations

What appears unavoidable is a more complex provision for annotations, at least if complex phenotypes formalized in OWL/description logic (DL) [3] shall be composable in terms of the usual intersection. Implicitly, this has already been observed in [19], to some extent also in connection with the EQ formalism. The following adheres to the understanding of annotations as outlined in sect. 1 and is inspired by the notion of *phenes* in [19]. Nevertheless, the subsequent variant differs in order to minimize changes to PATO and phenotype ontologies.

In order to solve especially the permutation problem of combined EQ descriptions, formally it suffices to have an "encapsulating" relation available. For instance, while (1) suffers from unwanted permutations, this is avoided in (2), where the encapsulating relation is termed `hasPheno`.

```
hasPheno some (red that inheresIn some eye) and
hasPheno some (short that inheresIn some tail)        (2)
```

Naturally, the question arises which ontological reading applies to `hasPheno`. We interpret (2) as a concept for classifying organisms (by two phenotype descriptions). The `hasPheno` relation belongs to an interpretive view/pattern that overlays common interconnections of entities, centering on the organism. In terms of the example, one may consider an organism O that has an eye E as its part, while there is a red R that inheres in E. Thus O is indirectly related with R in terms of common relations like inherence and part-of. In the phenotype view, this allows us to view O, as *phenotype bearer*, to exhibit R as a *pheno* of O. The latter connection is reflected by the `hasPheno` link between O and R. We require that each `hasPheno` link is "justified" by a chain of basic relations like *inheres-in*, *part-of*, *has-function*, *participates-in*, etc., that connects the entity in the pheno role with the one in the phenotype bearer role (PB in Fig. 2–4 below).

This approach leaves existing ontologies intact, resolves the first two particular issues identified, and accounts for the fifth, as well.

3.3 Enhancements for Relational Qualities

3.3.1 Purely Formal Extension On the remaining issues of relational expressiveness and consistency of domain modeling, we first observe that the current relational EQ model forms a special case of reifying (only binary) relations with *fixed* auxiliary relations, cf. the structural part of [1]. The main uncommon feature is the naming of those auxiliary relations as `inheresIn` and `towards`,[3] rather than using names counting arguments like $\texttt{argument}_1$ and $\texttt{argument}_2$. With the latter, an extension to n-ary relations is straightforward, which would solve the expressiveness issue. However, with fixed auxiliary relations there is no support for consistent domain modeling because the assignment of "values" to arguments is arbitrary. This may be the reason why all published variants of this pattern that we are aware of eventually suggest the *variable*, relation-specific naming of auxiliary relations [34, sect. 5.1], [28, 1].

Therefore, we do not see that changing the interpretation of relational EQ statements could be sidestepped, if inter-modeler consistent domain modeling is to be supported any further. Striving at the same time for ontological adequacy somewhat systematically, we adopt the model of relations and (relational)[4] roles from the General Formal Ontology (GFO) [15, 16], cf. also [23, 22].

3.3.2 Ontological Alternatives Using Relations In brief, relations in GFO are considered as categories of relators. *Relators* are

[2] Notably, this scheme is seen as equally eligible as phenotype description as the basic EQ scheme Q `that inheresIn some` E in [25, sect. 2.3, p. 5]. Giving preference to the basic EQ scheme appears to have been an arbitrary choice. In terms of their relationship to annotated entities the two schemes differ evidently. Nevertheless, the entity-focused scheme shares analogous problems to those expounded for the basic EQ scheme, in particular the permutation problem.

[3] Admittedly, `inheresIn` is meant to link to the ontological notion of inherence, whereas `towards` is introduced for rather technical reasons in [25] (circumventing an inherence relation of higher arity). It remains to be explored in greater detail whether `towards` can be adequately reinterpreted in terms of the notion of *external dependence*, see [11, esp. sect. 6.2.7].

[4] There are more types of roles in GFO, but for brevity we use roles and relational roles as synonyms herein. Note further that from here on 'role' is reserved for the ontological interpretation, whereas the meaning as set of pairs / as binary relation in the context of description logics and OWL is referred to as 'OWL property' or 'DL role'.

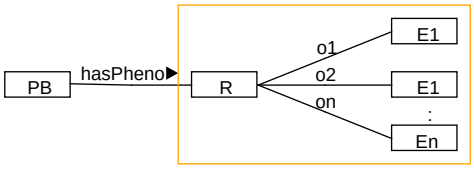


Figure 2. Roles-as-properties: Ontological roles encoded as OWL properties.

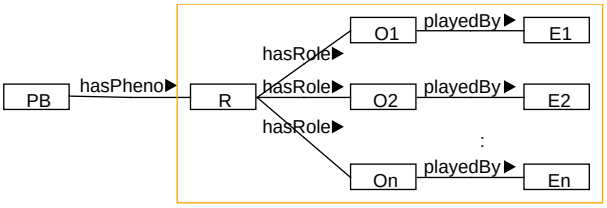


Figure 3. Roles-as-classes: Ontological roles modeled as classes in OWL.

ontological individuals akin to qualities, but with the power to mediate / connect entities. A relator consists of *role* individuals (via `hasRole` / `roleOf`) and each role individual, besides depending on the relator, depends on a *player* (via `playedBy` / `plays`). The term 'player' is relative to this approach; in general, arbitrary entities can play a role within a relation. At the categorial / class level, each relation R is associated with a set of role categories that forms the *role base* for this relation. Basically, that means for each relator of type R that its roles must instantiate one of the role categories in that set, cf. [22, sect. 3.3.3].

The GFO model of relations and roles can be encoded into an OWL representation in two obvious ways, termed *roles-as-properties* (Fig. 2) and *roles-as-classes* (Fig. 3). Common to both cases is to represent phenotype descriptions involving a relation R and (kinds of) entities $E_1, \ldots, E_n$ as argument restrictions. Either, corresponding to Fig. 2, roles are left implicit in the OWL properties $o_1, \ldots, o_n$, or, regarding Fig. 3, role categories are explicated as OWL classes $O_1, \ldots O_n$ (in between R and the E_i). Consider the example of *iron concentration in the spleen*, with the relation *concentration* and assuming that its two role categories are labeled *concentrated* (played by those entities concentrated in another) and *concentrator* (played by those entities within which another entity is concentrated). Then roles-as-properties yields in OWL

```
hasPheno some ( concentration and
                (concentrated some iron) and
                (concentrator some spleen) ),
```

whereas roles-as-classes leads to

```
hasPheno some ( concentration and
 (hasRole some (concentrated that playedBy some iron)) and
 (hasRole some (concentrator that playedBy some spleen)) ).
```

The first of these cases equals the above approach of using variable, relation-specific names for the auxiliary relations [34, sect. 5.1], [28, 1]. The second uses only two OWL properties `hasRole` and `playedBy` (and their inverses, possibly), but here this is unproblematic because the roles of the reified relation explicitly account for what is missing with fixed auxiliary relations without roles. Of course, both of these proposals will require a syntactic extension of the EQ model in order to capture the corresponding roles within EQ statements. Moreover, the roles-as-properties way may be simpler to reinterpret in other top-level ontological theories, because the roles presupposed by GFO are less explicit compared to roles-as-classes.

3.3.3 Ontological Alternative Using Relations and Qualities
The previous subsection suggests two ontologically inspired ways of understanding relational qualities like *concentration of* (PATO:0000033, hereafter CO) in EQ statements that cure the immediate deficiencies previously described. Both are based on a purely relational reading of CO (and relational qualities, in general), i.e., CO is merely considered as a noun form of the phrase *is concentrated in* (CI). For example, '(a particular amount of) *iron* I is concentrated in a (particular) *spleen* S' is a "relational proposition", stating that I is concentrated in S. This proposition can be true of false, depending on whether the relation CI applies to I and S or not, but there is nothing to be measured (neither quantitatively nor qualitatively).[5] In noun form, yet somewhat artificially, one may equivalently refer to 'there is concentration of I in S' (note that I and S are particulars).

However, we hold that CO comes in a second flavor, which is more amenable to specialization with notions like *increased concentration of* or to expressing specific values, e.g., $0.5g/l$. In phrases like 'the concentration of X in Y is $0.5g/l$, it appears more adequate to us to view CO as a proper quality which can be numerically quantified. Of course, immediately the question arises what that quality inheres in, which must be something that "includes" X and Y, not only one of the two. Here, computing the value of CO is instructive, which is based on values of qualities inhering solely in either X or Y, say, the weight of X and the volume of Y. The relationship between X and Y (of type CI, say) is characterized by the value within the CO phrase (in the second reading). Therefore, our current attempt of capturing relational qualities according to this analysis is to view them as inhering in particular relators, say a CI relator between X and Y. Admittedly, this is a deliberate, but no imperative choice among the possibilities within GFO. Other candidates for bearers of these qualities would be the overall relational fact, or one might consider the mereological sum of X and Y, in analogy to the inherence of relators in [11, sect. 6.2.7].[6] Regarding implementation in OWL, though, note that neither facts nor mereological sums are readily available on the basis of relators/relations and their arguments.

Eventually we arrive at a third approach, depicted in Fig. 4, where the relation is characterized by a quality. In the example, that means that CI is distinguished from CO, the latter being understood as a quality that inheres in CI relators / instances. Accordingly, we refer to this approach as *relator-based-quality*. Note that the intuitive term 'relational quality' experiences a formal-ontological reinterpretation from relations in the previous cases roles-as-properties and

[5] Pursuing this line of thought further in the example, one may wonder what remains as the actual difference between CI and relations like 'is contained in' and 'is part of'.

[6] If the latter option is to be followed, a more detailed analysis is required, though. Thinking of an amount of iron I concentrated in a spleen S, the question arises whether the mereological sum of I and S would differ from S. More generally, there may be interaction between the relation under consideration and forming a mereological sum of the relata.

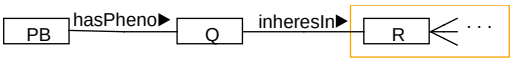


Figure 4. Relator-based-qualities: Relators characterized by qualities.

roles-as-classes to qualities proper (which are not relations) in the relator-based-qualities approach. Looking again at *iron concentration in the spleen*, assuming the roles-as-properties approach for modeling a relation `isConcentratedIn` (with roles like above) and a relational quality `concentration` yields in OWL

```
hasPheno some ( concentration that
        (inheresIn some (isConcentratedIn and
                         (concentrated some iron) and
                         (concentrator some spleen))) ).
```

This approach appears ontologically plausible to us currently, following the explanations above. Moreover, from the point of view of representation, it exhibits the beneficial property that CO is a "unary quality" like *color*, in the sense that it inheres in a single entity (a CI relator, which in turn accounts for the relational character of the quality). Any general account of qualities and quality values should thus be applicable to CO as it is to qualities like *color*. Furthermore, linking qualities to relators does not prescribe an overly specific relation model, but allows for adopting either of the approaches roles-as-properties and roles-as-classes in formalizing relations and roles, or even other theories (for which the quality bearer may require re-inspection).

4 DISCUSSION

Due to spatial limitations we focus the subsequent discussion mainly on aspects of the enhancements for relational qualities, where Table 1 compactly summarizes the approaches herein. Only minimal coverage of the introduction of the `hasPheno` relation can be given here. The latter is inspired by, but deviates from the notion of *phenes* and the *hasPhene* relation in [19]. Phenes may be understood as quality-like entities that reflect / abstract complex aspects that an organism is involved in. Accordingly, one immediate difference is that phenes are additional entities regarding those reflected aspects, whereas `hasPheno` bridges directly to one of the entities within those aspects. In any case, further comparison of the strengths and weaknesses of both views is a future task.

In connection with the general annotation-oriented interpretation, all three approaches for an improved account of relational qualities are designed to satisfy the issues identified in sect. 2.2, possibly varying in their degree of ontological adequacy. Concerning major disadvantages, clearly, all cases lead to significantly greater complexity of the representation through a considerable extension of vocabulary elements (see Table 1 for details). Concerning the "style" of reification embodied in roles-as-properties and roles-as-classes, there are also further unintended *technical issues*, surveyed in [10, sect. 2.2] (only with respect to roles-as-properties). At least in terms of reasoning, more precisely consistency checking and verifying entailments, those technical issues present no negative effects. Ibidem a number of potential *modeling shortcomings* are presented, in brief: (1) impeded manageability of the ontology, (2) purely technical nature of the additional vocabulary elements or at least an unclear ontological status, and (3) modeling diversity due to arbitrary splittings of reified relations, e.g. of reifying a 6-ary relation in terms of two ternary ones.

Feature descriptions, followed by feature matrix:

A	role information	E	max. nr. of relevant vocabulary elements (fixed / per n-ary relation)
B	unlimited arity of relations	F	add. characterization of relations
C	variable arity of relations		
D	straight-forward database support		

Feature	EQ	RP	RC	RQ
A	no	yes	yes	yes
B	no (yes)	yes	yes	yes
C	no	yes	yes	yes
D	yes	no (?)	no (?)	no (?)
E	2 / 0	$0 / n+1$	$2 / n+1$	$X+1 / Y+1$
F	no	no	no	yes

Table 1. Summary of the main features of the discussed approaches (EQ: entity-quality, RP: roles-as-properties, RC: roles-as-classes, RQ: relational-quality). Entry (B,EQ) reflects the discussed extensibility of EQ. X, Y stand for the respective numbers of the RP or RC columns, depending on the relation model combined with RQ.

We disagree with all of these, yet to different degrees. Concerning (1), we agree that more vocabulary is involved which requires additional attention in *ontology maintenance*. But this can be countered by the mutual disjointness of relation, role, and non-relational classes and the use of distinct subsumption hierarchies / graphs for each category, within which relations, roles, and other classes can be organized manageably. Extra effort that remains is to determine role names for each relation when introducing the latter, which is a source of inter-modeler differences.[7] The *use of the ontology* may be less affected, if there are effective intermediate representations and user interfaces, cf. [25, p. 1]. (2) is wrong in the light of the GFO approach to relations and roles, where these are ontological entities and thus not of purely technical nature.[8] Criticism (3) appears not applicable in our case, because the reification directly uses roles instead of arbitrary k-ary "parts" of an n-ary relation (where $k < n$).

Moreover, we see significant advantages in modeling and expressiveness that arise from the use of roles. For instance, relations are not only unconstrained in the number of arguments per relation, but one may even use anadic relations (i.e., with a variable number of arguments) and such with optional arguments. Similarly, symmetry properties of relations derive naturally from allowing for multiply instantiable role categories in the context of a role base. That means, a relation may be instantiated by relators that have several individual roles instantiating the same role category.

Notably, it is also symmetry of this kind that produces doubts on the treatment of concentration in [13, sect. 3.2]. Hastings et al. present a fairly detailed analysis of substance mixtures (among other topics) which we can follow to a large extent. This analysis is aimed at formalizing the notion of concentration in description logics. In this connection and transferred to the original EQ model (sect. 1

[7] However, one may adopt linguistic principles in some cases. E.g., for binary relations that can be appropriately named by verbs, participles can be used as rolenames in many cases. E.g., if *concentration of* (PATO:0000033) is traced back to *to concentrate*, the role(name)s of the *concentrated* and the *concentrating* may be formed.

[8] Admittedly, the roles-as-classes approach is closer to the ontological view of GFO, whereas roles-as-properties is a mainly technical simplification of the former. But this is not the technical nature critized in [10].

and 2), the consistency of domain modeling is achieved – for concentration only – by simply declaring that concentrations inhere in the entity, say `calcium`, that is concentrated in another, say `blood`. This likely means for EQ that the concentration is linked to that other entity by means of `towards`, and thus `concentration that inheresIn some calcium and towards some blood` is the preferred formalization, cf. sect. 2.2.[9] In their analysis, however, this choice itself is not explained. Considering other relational properties than concentration, an analogous decision would have to be made for each relational property (and established among modelers), which appears less attractive than finding more general rules. Closing the circle to symmetric relations, for these it is not possible to distinguish one of the arguments (at least, not based on their roles only). For instance, for a phenotype like *increased distance of the eyes*, it appears completely implausible to select one eye in which a distance `inheresIn`, whereas it is `towards` the other eye. Especially the relator-based-quality approach, despite its own unresolved choices, see sect. 3.3.3, avoids such arbitrary fixing.

A practical deficiency of all three approaches that might be of potential importance is that the increased complexity prevents a straight-forward integration of corresponding annotations into the relational schemas of annotating databases. However, we have not yet explored alternatives in this connection, and this problem may re-occur due to an in-principle incompatibility of various aims, including the provision for n-ary relations vs. simple database implementation.

5 CONCLUSION

In this paper we report on the (work-in-progress) state of our analyses and improvement proposals concerning the Entity-Quality (EQ) model. A simple general modification in the understanding of qualities in PATO is argued to be necessary. Moreover, three variants of extended support for relations / relational qualities are presented.

Much work remains to be done or completed. The approaches detailed herein rely on theoretical analyses thus far. For further assessment, an experimental evaluation should be conducted, e.g. exploring the efficiency of reasoning over ontologies which rely on one or another approach. Despite our (preliminary) decision to minimize changes to the EQ interpretation to the greatest possible extent, we still see many interesting open theoretical issues in the EQ model, respective ontologies, and phenotype understanding and representation in general. For instance, we are convinced that not all concepts of PATO should be regarded ontologically properly as qualities. The not yet elaborated connections between `hasPheno` and `hasPhene` in [19] are named above. Accordingly, further alternatives, which possibly involve larger re-interpretation of existing resources, should be studied and compared. On that basis EQ syntax extensions and possibly changes to phenotype ontologies can be devised.

ACKNOWLEDGEMENT

We are grateful to the reviewers for constructive criticism and for additional pointers to the literature.

[9] At least, this is what we read from sect. 3.2 in [13]. It is not actually stated whether and how the concentration relates directly with the mixture, e.g. blood. One can also find more informal statements which suggest different interpretations, e.g. in sect. 2.2 of [13].

REFERENCES

[1]Nary relationship. `http://www.gong.manchester.ac.uk/odp/html/Nary_Relationship.html`, 2009.

[2]Michael Ashburner, Catherine A. Ball, Judith A. Blake, David Botstein, Heather Butler, J. Michael Cherry, Allan P. Davis, Kara Dolinski, Selina S. Dwight, Janan T. Eppig, Midori A. Harris, David P. Hill, Laurie Issel-Tarver, Andrew Kasarskis, Suzanna Lewis, John C. Matese, Joel E. Richardson, Martin Ringwald, Gerald M. Rubin, and Gavin Sherlock. Gene ontology: tool for the unification of biology. *Nature Genetics*, 25(1):25–29, May 2000.

[3]Franz Baader, Diego Calvanese, Deborah McGuinness, Daniele Nardi, and Peter Patel-Schneider, editors. *The Description Logic Handbook: Theory, Implementation and Applications*. Cambridge University Press, Cambridge (UK), January 2003.

[4]James P. Balhoff, Wasila M. Dahdul, Cartik R. Kothari, Hilmar Lapp, John G. Lundberg, Paula Mabee, Peter E. Midford, Monte Westerfield, and Todd J. Vision. Phenex: Ontological annotation of phenotypic diversity. *PLoS ONE*, 5(5):e10500.1–10, 2010.

[5]Tim Beck, Hugh Morgan, Andrew Blake, Sara Wells, John Hancock, and Ann-Marie Mallon. Practical application of ontologies to annotate and analyse large scale raw mouse phenotype data. *BMC Bioinformatics*, 10(Suppl 5):S2, 2009.

[6]Judith A. Blake, Carol J. Bult, James A. Kadin, Joel E. Richardson, Janan T. Eppig, and the Mouse Genome Database Group. The Mouse Genome Database (MGD): premier model organism resource for mammalian genomics and genetics. *Nucleic Acids Research*, 39(suppl 1):D842–D848, 2011.

[7]C. J. Bult, J. T. Eppig, J. A. Kadin, J. E. Richardson, and J. A. and Blake. The Mouse Genome Database (MGD): mouse biology and model systems. *Nucleic acids research*, 36(Database issue), January 2008.

[8]Francis S. Collins, Richard H. Finnell, Janet Rossant, and Wolfgang Wurst. A new partner for the international knockout mouse consortium. *Cell*, 129(2):235, 2007.

[9]Georgios V. Gkoutos, Eain CJ Green, Ann-Marie Mallon, John M. Hancock, and Duncan Davidson. Using ontologies to describe mouse phenotypes. *Genome Biology*, 6(1):R8, 2004.

[10]Niels Grewe. A generic reification strategy for n-ary relations in DL. In Herre et al. [17], pages N.1–5.

[11]Giancarlo Guizzardi. *Ontological Foundations for Structural Conceptual Models*, volume 015 of *Telematica Instituut Fundamental Research Series*. Telematica Instituut, Enschede (Netherlands), 2005. also: CTIT PhD Series No. 05-74.

[12]Giancarlo Guizzardi. On the representation of quantities and their parts in conceptual modeling. In Anthony Galton and Riichiro Mizoguchi, editors, *Formal Ontology in Information Systems: Proceedings of the Sixth International Conference, FOIS 2010, Toronto, Canada, May 11-14*, volume 209 of *Frontiers in Artificial Intelligence and Applications*, pages 103–116, Amsterdam, 2010. IOS Press.

[13]Janna Hastings, Christoph Steinbeck, Ludger Jansen, and Stefan Schulz. Substance concentrations as conditions for the realization of dispositions. In Ronald Cornet and Stefan Schulz, editors, *Semantic Applications in Life Sciences: Proceedings of the 4th International Workshop on Formal Biomedical Knowledge Representation, KR-MED 2010, hosted by Bio-Ontologies*

2010, Boston, Massachussetts, USA, Jul 9-10, volume 754 of *CEUR Workshop Proceedings*, Aachen, Germany, 2010. CEUR-WS.org.

[14]Terry F. Hayamizu, Mary Mangan, John P. Corradi, James A. Kadin, and Martin Ringwald. The Adult Mouse Anatomical Dictionary: a tool for annotating and integrating data. *Genome Biology*, 6(3):R29, 2005.

[15]Heinrich Herre. General Formal Ontology (GFO): A foundational ontology for conceptual modelling. In Roberto Poli, Michael Healy, and Achilles Kameas, editors, *Theory and Applications of Ontology: Computer Applications*, chapter 14, pages 297–345. Springer, Heidelberg, 2010.

[16]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 1.0.1]. Draft, Research Group Ontologies in Medicine, Institute of Medical Informatics, Statistics and Epidemiology, University of Leipzig, Leipzig, 2007.

[17]Heinrich Herre, Robert Hoehndorf, Janet Kelso, and Stefan Schulz. Proceedings of the 2nd workshop of the GI-Fachgruppe "Ontologien in Biomedizin und Lebenswissenschaften" (OBML): Mannheim, Germany, Sep 9-10. IMISE-Report 2/2010, IMISE, University of Leipzig, Germany, Sep 2010.

[18]Robert Hoehndorf, Michel Dumontier, Anika Oellrich, Dietrich Rebholz-Schuhmann, Paul N. Schofield, and Georgios V. Gkoutos. Interoperability between biomedical ontologies through relation expansion, upper-level ontologies and automatic reasoning. *PLoS ONE*, 6(7):e22006.1–9, 2011.

[19]Robert Hoehndorf, Anika Oellrich, and Dietrich Rebholz-Schuhmann. Interoperability between phenotype and anatomy ontologies. *Bioinformatics*, 26(24):3112–3118, 2010.

[20]Robert Hoehndorf, Paul N. Schofield, and Georgios V. Gkoutos. PhenomeNET: a whole-phenome approach to disease gene discovery. *Nucleic Acids Research*, 2011. in press (advance access: `http://dx.doi.org/10.1093/nar/gkr538`).

[21]Degtyarenko K., P. de Matos, M. Ennis, J. Hastings, M. Zbinden, A. McNaught, R. Alcantara, M. Darsow, M. Guedj, and Michael Ashburner. ChEBI: a database and ontology for chemical entities of biological interest. *Nucleic Acids Research*, 36(Suppl 1):D344–D350, 2008.

[22]Frank Loebe. An analysis of roles: Towards ontology-based modelling. Onto-Med Report 6, Research Group Ontologies in Medicine, University of Leipzig, 2003. Master's Thesis.

[23]Frank Loebe. Abstract vs. social roles – Towards a general theoretical account of roles. *Applied Ontology*, 2(2):127–158, 2007.

[24]Martin Mahner and Michael Kary. What exactly are genomes, genotypes and phenotypes? And what about phenomes? *Journal of Theoretical Biology*, 186(1):55–63, 1997.

[25]Chris Mungall, Georgios Gkoutos, Nicole Washington, and Suzanna Lewis. Representing phenotypes in OWL. In Christine Golbreich, Aditya Kalyanpur, and Bijan Parsia, editors, *Proceedings of the OWLED 2007 Workshop on OWL: Experiences and Directions, Innsbruck, Austria, Jun 6-7*, volume 258 of *CEUR Workshop Proceedings*, Aachen, Germany, 2007. CEUR-WS.org.

[26]Christopher Mungall, Georgios Gkoutos, Cynthia Smith, Melissa Haendel, Suzanna Lewis, and Michael Ashburner. Integrating phenotype ontologies across multiple species. *Genome Biology*, 11(1):R2.1–16, 2010.

[27]Fabian Neuhaus, Pierre Grenon, and Barry Smith. A formal theory of substances, qualities and universals. In Achille C. Varzi and Laure Vieu, editors, *Formal Ontology in Information Systems: Proceedings of the Third International Conference, FOIS-2004, Turin, Italy, Nov 4-6*, volume 114 of *Frontiers in Artificial Intelligence and Applications*, pages 49–59, Amsterdam, 2004. IOS Press.

[28]Natasha Noy and Alan Rector. Defining N-ary relations on the Semantic Web. W3C working group note, World Wide Web Consortium (W3C), April, 12 2006. `http://www.w3.org/TR/2006/NOTE-swbp-n-aryRelations-20060412/`.

[29]Peter N. Robinson, Sebastian Köhler, Sebastian Bauer, Dominik Seelow, Denise Horn, and Stefan Mundlos. The Human Phenotype Ontology: A tool for annotating and analyzing human hereditary disease. *American Journal of Human Genetics*, 83(5):610–615, 2008.

[30]Nadia Rosenthal and Steve Brown. The mouse ascending: perspectives for human-disease models. *Nature Cell Biology*, 9(9):993–999, 2007.

[31]Gary Schindelman, Jolene Fernandes, Carol Bastiani, Karen Yook, and Paul Sternberg. Worm Phenotype Ontology: Integrating phenotype data within and beyond the C. elegans community. *BMC Bioinformatics*, 12(1):32, 2011.

[32]Stefan Schulz, Daniel Schober, Djamila Raufie, and Martin Boeker. Pre- and postcoordination in biomedical ontologies. In Herre et al. [17], pages L.1–4.

[33]Cynthia L. Smith, Carroll-Ann W. Goldsmith, and Janan T. Eppig. The Mammalian Phenotype Ontology as a tool for annotating, analyzing and comparing phenotypic information. *Genome Biology*, 6(1):R7.1–9, 2004.

[34]Robert Stevens, Mikel Egana Aranguren, Katy Wolstencroft, Ulrike Sattler, Nick Drummond, Matthew Horridge, and Alan Rector. Using OWL to model biological knowledge. *International Journal of Human-Computer Studies*, 65(7):583–594, July 2007.

[35]Alexandr Uciteli, Silvia Groß, Sergej Kireyev, and Heinrich Herre. An ontologically founded architecture for information systems in clinical and epidemiological research. *Journal of Biomedical Semantics*, 2(Suppl 4):S1.1–22, 2011.

[36]W3C. OWL 2 Web Ontology Language Document Overview. W3C Recommendation, World Wide Web Consortium (W3C), Cambridge (Massachusetts), 2009. `http://www.w3.org/TR/2009/REC-owl2-overview-20091027/`.

[37]Nicole L. Washington, Melissa A. Haendel, Christopher J. Mungall, Michael Ashburner, Monte Westerfield, and Suzanna E. Lewis. Linking human diseases to animal models using ontology-based phenotype annotation. *PLoS Biology*, 7(11):e1000247.1–20, 2009.

Redesigning an Ontology Design Pattern for Realist Ontologies

Djamila Raufie[1], Stefan Schulz[1,2], Daniel Schober[1], Ludger Jansen[3,4], Martin Boeker[1]

[1] Institute of Medical Biometry and Medical Informatics, Freiburg University Medical Center Freiburg, Germany
[2]Institute for Medical Informatics, Statistics and Documentation, Medical University of Graz, Austria
[3]Institute of Philosophy, University of Rostock, Germany
[4]Philosophical Institute, RWTH Aachen University, Germany

ABSTRACT

Ontology Design Patterns (ODPs) offer solutions for recurring ontology design problems. They promise to enhance the ontology building process in terms of flexibility, reusability and expansion. We analyze ODP repositories and investigate their relation with top-level or upper-level ontologies. In particular, we compare the Action ODP from the NeOn repository to the BioTop upper ontology. In view of the differences in the respective approaches, we ask whether the Action ODP can be embedded into BioTop. We demonstrate that, this requires reinterpreting the meaning of classes of the NeOn Action ODP in the light of the precepts of realist ontologies. As a result, the redesign required clarifying the ontological commitment of the classes in the Action ODP by assigning them to top-level categories. Thus, ambiguous definitions are avoided. Real entities are clearly distinguished from information artifacts. Our approach avoids the commitment to the existence of dubious future entities which underlies the NeOn Action ODP. The redesign is parsimonious in the sense that existing BioTop content proved to be largely sufficient to define the different types of actions and plans[1].

1 INTRODUCTION

Design patterns are popular in software engineering. They have recently also been proposed for ontology building. Such ontology design patterns (ODPs) claim to be reusable solutions to commonly occurring design problems, thus supporting ontology engineers in the efficient development of ontologies. Another advantage is that the resulting artifacts are easier to be handled, as their design principles are explicitly known. We have investigated three main sources of ODPs:

1. The ***Semantic Web Best Practices and Deployment Working Group***[2]: The aim of this group is to guide Semantic Web developers to build reusable OWL ontologies;
2. The ***Ontology Design Patterns (ODPs) Public Catalog***[3,4], distinguishing between extensional ODPs, good practice ODPs and modeling ODPs (Egaña Aranguren, 2008; Egaña Aranguren, 2010); and
3. The ***Ontology Design Patterns.org*** (ODP)[5] under the European NeOn project, distinguishing between *Structural ODPs*, *Correspondence ODPs*, *Reasoning ODPs*, *Presentation ODPs*, *Lexico-Syntactic ODPs*, and *Content ODPs* (Gangemi, 2009; Blomqvist, 2005).

Despite the wealth of available ODPs, at least the latter repositories' content tends to be rather idiosyncratic, mainly due to the fact that its ODPs refrain from a clear ontological commitment and leave the final interpretation to the user.

We defend an ontology engineering approach rooted in a philosophically founded ontological top-level, using the BioTop ontology[6], a publically available upper-domain level for the life sciences (Schulz, 2009; BioTop, 2011). BioTop provides foundational classes and relations embedded in rich axiomatized definitions. Its set of relations is considered to be exhaustive (with defined domains and ranges) so that ontology developers only need to subclass existing classes and define them by adding restrictions using BioTop relations. BioTop is roughly compatible with the major top-level ontologies like BFO (Grenon, 2004), DOLCE[7], and the OBO Relation Ontology[8].

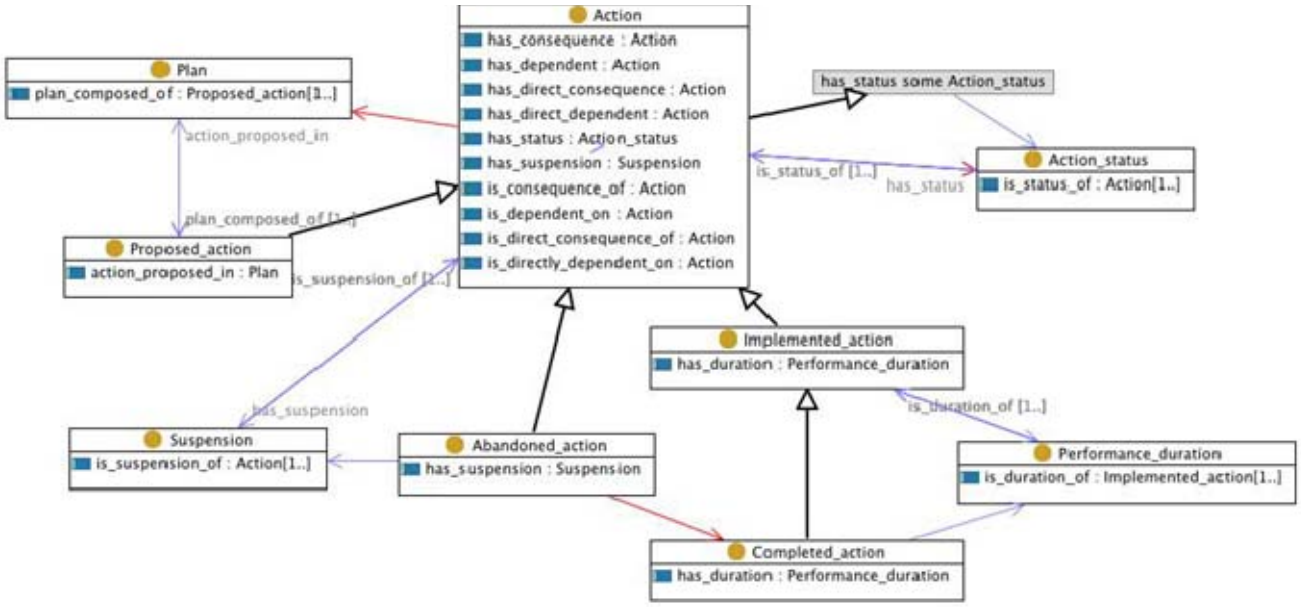
We want to investigate the following:

1. Which elements of ODPs are already expressed by BioTop axioms?
2. To what extent and how can existing ODPs be re-interpreted or adapted for inclusion into top-level or upper-level ontologies, e.g. BioTop?
3. Which ODPs or parts thereof can be redesigned as extensions to BioTop?

* To whom correspondence should be addressed.

[1] An ontology containing the redesign proposed by us can be downloaded from: http://purl.org/steschu/OBML2011
[2] http://www.w3.org/2001/sw/BestPractices/OEP/
[3] http://www.gong.manchester.ac.uk/odp/html/index.html
[4] http://sourceforge.net/projects/odps/
[5] http://ontologydesignpatterns.org
[6] http://www.purl.org/biotop/biotop.owl
[7] http://www.loa-cnr.it/DOLCE.html
[8] http://www.obofoundry.org/ro/

Fig. 1. The Action ODP from the NeOn Project (http://ontologydesignpatterns.org/wiki/Submissions:Action)



We will identify corresponding representations in BioTop by comparing the intended meaning of ODPs and BioTop representations by structural and logical analysis. The example under scrutiny will be the Action ODP from the NeOn repository[9].

2 MATERIALS AND METHODS

Fig. 1 depicts the Action ODP from the NeOn ODP repository. Its aim is to represent actions that are proposed, planned, performed or abandoned, together with their status and duration (Blomqvist, 2010). It also includes action properties such as 'status' and 'duration'. *Action* is described as "The process of doing something. An action is performed by an agent" (Blomqvist, 2010). A link to an *Action_status* class is used to differentiate actions in terms of being proposed, implemented (and possibly completed), or abandoned. As a result, *Proposed_action*, *Abandoned_action*, *Completed_action*, and *Implemented_action* are defined. Together with *Plan*, *Action_status*, *Suspension*, and *Performance_duration*, they can furthermore be related by a set of ten relations like **has_consequence**, **has_dependent** etc. The class *Performance_duration* shows the time interval in which an action is performed. Finally, *Plan* is introduced as a "set of proposed actions and the sequence in which to perform them" (Blomqvist, 2010).
To show that ODPs can be embedded in top level ontologies, our immediate goal is to make the Action ODP compatible with BioTop. Here, our basic assumptions are:

- The number of new representational units should be minimal; especially no new relations (object properties in Protégé) should be added.

3 RESULTS

In comparing the Action ODP with BioTop, we want to investigate exemplarily how an ODP can be adapted to an upper-level ontology like BioTop. In particular we want to know which parts of the Action ODP can be redesigned as an extension to BioTop, which elements of the Action ODP are expressed by axioms already present in BioTop, and to which extent BioTop's ontological assumptions are compatible with the Action ODP.
In BioTop the class *Action* already exists as a subclass of *Process*, promoted by an agent, having a clear role distinction between agent and patient. Processes can have temporal parts, i.e. there might be no time in which all parts of a process are simultaneously present. Processes have physical or abstract entities as participants:

Process equivalentTo
Particular and **hasParticipant** some *Particular*

The object property **hasAgent** is a subrelation of **hasParticipant.** Hence,

Action equivalentTo
Process and **hasAgent** some *Particular*

Additionally it holds that,

Action subClassOf **hasDuration** some *TimeInterval*

In BioTop, *Action* does not have specific modifiers as in the ODP. Yet this does not preclude subclassifying actions in

[9] http://ontologydesignpatterns.org/wiki/Submissions:Action

terms of suspension or completion. The definition tells us that actions, in order to exist, must have an agent and duration. As a consequence of BioTop's realist view of the world, we cannot assert the existence of entities and relations which are supposed to exist in the future. Therefore what is called *Proposed_action* in the ODP is not a subclass of *Action* in BioTop. The expression

hasAgent some *Particular*

posits the dependency of actions on existing agents: if there is no agent, there cannot be an action.
Like the Action ODP, BioTop already includes the class *Plan* with the axiom:

Plan subClassOf
InformationObject and **hasRealization** only *Process*

While in BFO *InformationObject* is not a *Realizable* (which embraces only *Disposition*, *Function* and *Role*),[10] BioTop extends the domain of the relation **hasRealization** to include also *InformationObject* and *ObjectQuality*. Its range is always the class *Process* (BioTop, 2011).

Accordingly, we can define a plan for a specific action *X* as:

X_Plan subClassOf *Plan* and **hasRealization** only *X*

An *X_Plan* is therefore only realized when an action of type *X* is accomplished. This is not the case if *X* is merely proposed. *X_proposed* is not a kind of *X*, because it has no real duration and agent. Proposed actions are no more actions than fake money is money, than a prevented victory is a victory or than pretending doctors are doctors. Proposed actions should rather be seen as the content of proposals, where a proposal could be an information object or some action like a speech act that formulates the plan. A plan can be anything outlining the course of an action, from a gene outlining the plan for proteins, a documentation of medical procedures and the schedule of operations in a hospital to mere mental entities like the intention to have lunch at noon. Likewise, a suspended action *X* is not an action of type *X*, because defining characteristics of *X* may be missing.
A simple example of a surgical action can help illustrate our discussion: *Endoscopic_Removal_of_Foreign_Body_from_Stomach* (e.g. in a child who swallowed a marble). For sake of brevity, we will refer to this procedure type as *X*. In a simplified form we can describe it as follows: every instance of *X* begins with an endoscopy preparation (*a*), followed by the introduction of the endoscope (*b*), endoscopic exploration (*c*), grasping of the foreign body (*d*), and extraction of the endoscope (*e*). A description of this is outlined in the information object *X_Plan*. This plan is realized only by actions that correspond to the sequence *abcde*. If any of these sub-actions is missing, the action is no longer of the type *X*. *X_Plan* is therefore only realized when *X* is fully accomplished. As *X* has necessarily all temporal parts *a-e*, *X_implemented* is not a subclass of *X*, because it may still be in the phase *a* or *b*, lacking the remaining sequential processes c-e. The same applies to *X_abandoned* (e.g. the action is incomplete because no foreign body was found, i.e. no *d* is performed). Rooted in realist philosophy, and in accordance with common sense, BioTop makes a clear distinction between plans as information objects vs. real processes. It is therefore not compatible with the NeOn Action ODP as depicted in Fig.1, which obfuscates the ontological distinction between real and hypothetic entities. We propose the following DL-compliant and realism-rooted model to account for the different "flavors" of actions:

Proposed Action. A proposed action is not an action. It is a refined plan and hence resides in a totally different top-level category, i.e. the category *InformationObject*. Thus a proposed action is rather an action proposal than an action that has been proposed. From this follows that the difference between *Proposed_action* and *Action* is not merely epistemic. Many instances of the class *Proposed_action* (i.e. many action proposals) are never realized, thus having no counterpart among the instances of the class *Action*.
Plans for actions of type X can be refined by adding further restrictions to the realization class of the plan. E.g., the general plan for *Endoscopic_Removal_of_Foreign_Body_from_Stomach* is refined in terms of a patient, a doctor, an operation room, a time slot etc. according to the following pattern:

Specified_X_ByDoctorInOperationRoom _Plan subClassOf
Plan and **hasRealization** only
(*X* and **hasAgent** some Doctor
and **hasLocus** some OperationRoom)

It should be noted that a *Specified_X_Plan* as such does not have an agent. Rather it is the realization of the plan that has an agent, a location, and so on.

The class *Specified_X_Plan* can be fully defined within BioTop:

Specified_X_Plan equivalentTo **outcomeOf** some
PlanSpecificationAction and
hasParticipant some *X_Plan*

The advantage over the NeOn approach is that *PlanSpecificationAction*, as a separate action, may have a different agent: The person who schedules the operation is not necessarily identical with the physician who performs it, and both may be different from the person that has formulated the generic operation procedure (Jansen, 2003).

[10] http://www.ifomis.org/bfo

Implemented Action. Here, the action may be ongoing, and it may still lack some of the features that make it an instance of the type *X_completed*. For instance, the stomach is being explored, but the foreign body not yet found. In such a case, only the initial sequential parts of the original plan have been executed. As a plan is only fully realized at the end of the action, an ongoing action realizes a proper part of the plan. E.g., if the whole plan projects the action parts *a*, *b*, *c*, *d*, *e*, an action which is ongoing in stage *c* has only realized the subplans *a* and *b*. Therefore:

X_implemented equivalentTo
Action and **realizationOf** some
((**abstractPartOf** some *X_Plan*) or *X_Plan*)

Completed Action. Here, the plan has been fully executed, all steps of the plan have been realized, and the action is over:

X_completed equivalentTo
Action and **realizationOf** some *X_Plan*

It can be seen easily from the definitions that all completed actions are implemented action. *Completed_action* is thus a subclass of *Implemented_action*. It should be noted that some actions may be completed as soon as they are implemented. This is the case if their *X_Plan* has only one abstract part, like, for example, looking at the Mona Lisa or sitting on the floor.

Abandoned Action. Here, the action is no longer being performed, but the plan has been executed only partly. In contrast to an implemented action, it is by definition not completed:

X_NonCompleted equivalentTo
Action and (not **realizationOf** some *X_Plan*) and
realizationOf some
(**abstractPartOf** some *X_Plan*)

Furthermore, the NeOn Action ODP introduces the status variable *Suspension* for permanently or temporarily suspended actions. In BioTop we suggest a similar solution, for the lack of a detailed enough time model. However, in order to be consistent with the ontological principles of BioTop, this property needs to be exactly typed. We name it *Inactive*, a subclass of *Quality*, linked to actions by the relation **hasProcessQuality**:

X_Abandoned equivalentTo
X_NonCompleted and **hasProcessQuality** some *Inactive*

An instance of *X_Abandoned* can permanently bear this quality; then the action is aborted. I can also lose this quality when the action is resumed; then it becomes an instance of implemented action, again. Note that all action classes distinguished here are non-rigid in the sense of OntoClean (Guarino, 2009): an *Action* token may first be implemented and eventually completed. Or it may be implemented and then abandoned. It may later be re-implemented and completed.
The restricted expressivity of DL does not allow tracking the identity of individuals across classes. Nevertheless the non-rigidity of these classes is an important guide for human ontology developers.

4 CONCLUSION

The proposed model demonstrates that BioTop provides enough resources and expressivity to represent even complex ODPs, here shown with the different "flavors" of *Action* as proposed in the NeOn ODP. We identified the following advantages of this approach:

1. **It is explicit** in terms of ontological commitment, i.e. it does not leave the interpretation of the meaning of its classes and relations to the user.
2. **It is parsimonious** in the sense that existing classes and relations in BioTop have proved as largely sufficient to define the different types of actions and plans. The only auxiliary classes that had to be created were *Inactive* and *PlanSpecificationAction*. No new relations were necessary, whereas the NeOn approach introduces ten new object properties rendering the pattern more complex than necessary.
3. **It is ontologically clearer** in the sense that ambiguous definitions are avoided. It does not conflate real entities and information artifacts.
4. **It has a simpler and more intuitive notion of existence**. The Action ODP claims existence for dubious future entities, e.g. in the class *Proposed_action.* According to the ground axioms, every instance of *Action* must have some agent. As it conceives of *Proposed_action* as a subtype of *Action*, any instance of *Proposed_action* would need to have an agent, too. But as proposed actions may never be implemented, the Action ODP seems to be committed to postulate the existence of potential or possible entities.
5. **It avoids counterintuitive consequences.** Treating *Proposed_action* as a subtype of *Action* yields, e.g., the consequence that there are actions that are never implemented. In fact, there could be proposals for actions that actually exclude each other: Someone could propose to eat the cake, another could propose to keep it. But as everyone knows, you cannot eat your cake and keep it. The Action ODP, however, would be committed to postulate the existence of both actions.
6. **It is user-friendly.** BioTop's strict division in disjoint partitions and the specification of domain and range restrictions in the definition of object properties guides the user on the right path when extending the ontology. In order to link actions and plans there are no other options than using the relation **realizationOf**; and for mereologically relating information entities there only exists **abstractPartOf**. Hence, compared to self-standing ODPs, patterns that are embedded in a top-

level ontology are more users compliant, as the user profits from inherited constraints.

ACKNOWLEDGEMENTS

This work was supported by the DFG grant JA 1904/2-1, SCHU 2515/1-1 within the project "Good Ontology Design" (GoodOD).

REFERENCES

Blomqvist E. (2005) *Modelling and using ontology design patterns*. http://www.neon-project.org/ (last accessed on July 7, 2011).

Blomqvist, E. (2010) *Ontology Design Patterns. org (ODP)*. Retrieved from http://ontologydesignpatterns.org/wiki/Submissions:Action (last accessed on September 9, 2011).

BioTop (2011) *A Top-Domain Ontology for the Life Sciences*. Retrieved from http://www.imbi.uni-freiburg.de/ontology/biotop/ (last accessed on September 9, 2011).

Egaña Aranguren M. (2008). Ontology Design Patterns for bio-ontologies: a case study on the Cell Cycle Ontology. *BMC Bioinformatics*. **29,** doi:10.1186/1471-2105-9-S5-S1.

Egaña Aranguren M. (2010). *Role and application of Ontology Design Patterns in Bio-Ontologies*. Lambert Academic.

Grenon P. (2004) Biodynamic Ontology: Applying BFO in the Biomedical Domain. *Ontologies in Medicine*, Amsterdam: IOS Press, 20–38.

Gangemi A. (2005) Ontology Design Patterns for Semantic Web Content. Musen et al. (eds.): *Proceedings of the Fourth International Semantic Web Conference, Galway, Irland*, Springer.

Gangemi A. (2007) Towards a Catalog of OWL-based Ontology Design Patterns. *XII Conferencia de la Asociación Española para la Inteligencia Artificial*.

Gangemi A. (2009). Ontology Design Patterns. *Handbook on Ontologies*, second edition, Springer, 221-243.

Guarino N. (2009) An Overview of OntoClean. *Handbook on Ontologies*, second edition, Springer, 201–222.

Jansen L. (2003) Planners, Deciders, Performers. Aristotelian Reflections on the Ontology of Agents and Actions. In: C. Kanzian, J. Quitterer, E. Runggaldier (eds.), *Persons. An Interdisciplinary Approach*, Wien: öbv & hpt, 208–215.

Schulz,S. (2009) Alignment of the UMLS semantic network with BioTop: methodology and assessment. *Bioinformatics*, **25**, i69-79.

OntoCheck: Verifying ontology Naming Conventions in Protégé 4

Daniel Schober[1], Ilinca Tudose[1], Vojtech Svatek[2], Martin Boeker[1]*

[1]*Institute of Medical Biometry and Medical Informatics (IMBI), University Medical Center, 79104 Freiburg, Germany*
[2]*University of Economics, Prague, Nám. W. Churchilla 4, 130 67 Praha 3, Czech Republic*

ABSTRACT

Motivation: The Protégé 4 editor is amended with ontology curation abilities that foster ontology pre-release checks and overall quality assurance. In particular the OntoCheck plugin helps to clean up an ontology with regard to lexical heterogeneity, i.e. enforcing naming conventions and metadata completeness. Once specified naming pattern checks can be stored and exchanged for later re-use. Here we provide a first version of the software tool, illustrate and discuss its applications and highlight potential future expansions.

1 INTRODUCTION

With the advent of the semantic web and RDF-based knowledge representation techniques of-the-shelf ontology editors like Protégé 4 [1] gain widespread use. Although its functionality is sufficient for daily ontology editing tasks, some clean-up checks on the ontology - to be carried out before ontology release - could complement P4 in a useful way.

Here, we introduce a Protégé plugin that checks certain properties of an active OWL ontology (OntoCheck) and allows for amendments (OntoCure) in the areas of Metadata Analysis, e.g. completeness and cardinality checks on mandatory and obligatory annotation properties, and Naming Conventions e.g. lexical analysis and labeling enforcement for representational units (RU) [2].

1.1 Checks on Metadata

Before a new ontology version is released for public use, it should be checked if all mandatory metadata, i.e. annotation properties like natural language definitions, or class labels are present and the ontology is sufficiently described.

1.2 Checks on Naming Conventions

Inconsistent class labeling impairs readability and navigation in ontology class hierarchies. Explicit naming conventions will assist consumers of ontologies to more readily understand what meanings were intended by the authors when looking at annotated data sets. Clear naming conventions on RUs like OWLClassName, labels and property names provide guidance to ontology creators and help developers to avoid flaws and lexical inaccuracies [3] when editing, but especially when interlinking, ontologies. Clear and explicit naming also fosters communication when ontology engineers need to collaborate with external groups to align their ontologies and to ensure effective maintenance of modularity.

Our plugin contributes to such lexical harmonization by validating class names according to specified queries. The presented plugin ensures consistency by testing for defined RU label patterns, e.g. as outlined in the OBO Foundry naming conventions [4], an effort that proposes a set of typographic, syntactic and semantic conventions for labeling classes.

2 RESULTS

The OntoCheck Java plug-in was implemented for the Protégé 4.1 ontology editor using the Protégé OWL API under Eclipse. Example naming conventions were taken from [4]. The OntoCheck plugin (http://www.imbi.uni-freiburg.de/ontology/OntoCheck/) provides a new tab that is organized in the three panels Check, Compare and Statistics. For each we list the testing and curation capabilities together with example applications taken from the practice.

The Check panel

The Check panel (Fig. 1) allows users to first select a metadata element or annotation property (e.g. self-defined, RDF, Dublin Core, RU-meta etc.) of which the value is to be checked for presence (cardinality). E.g. the user can specify that all classes should have at least one natural language definition and one label. The result of a test is displayed to the right in a 'result classes' pane, listing classes failing the test, which can subsequently be enriched with the lacking metadata.

The Check panel further allows verifying whether a particular orthographic or morpho-syntactic naming convention is fulfilled in a selected subtree. In our example a user can select e.g. OWLClassName, rdfs:label, or any other annotation property available, and check all values in the active ontology for all subclasses, leaf-wards from a selected entry node, i.e. to check for

Word Case: mixed case, lower case start, lower case only, upper case start or all upper case.

Word Separator: none, space, hyphen, underscore, dot.

- **Digits**: to check for numeric expression in labels, e.g. to look for cardinality and order indicators.

*To whom correspondence should be addressed:
schober@imbi.uni-freiburg.de

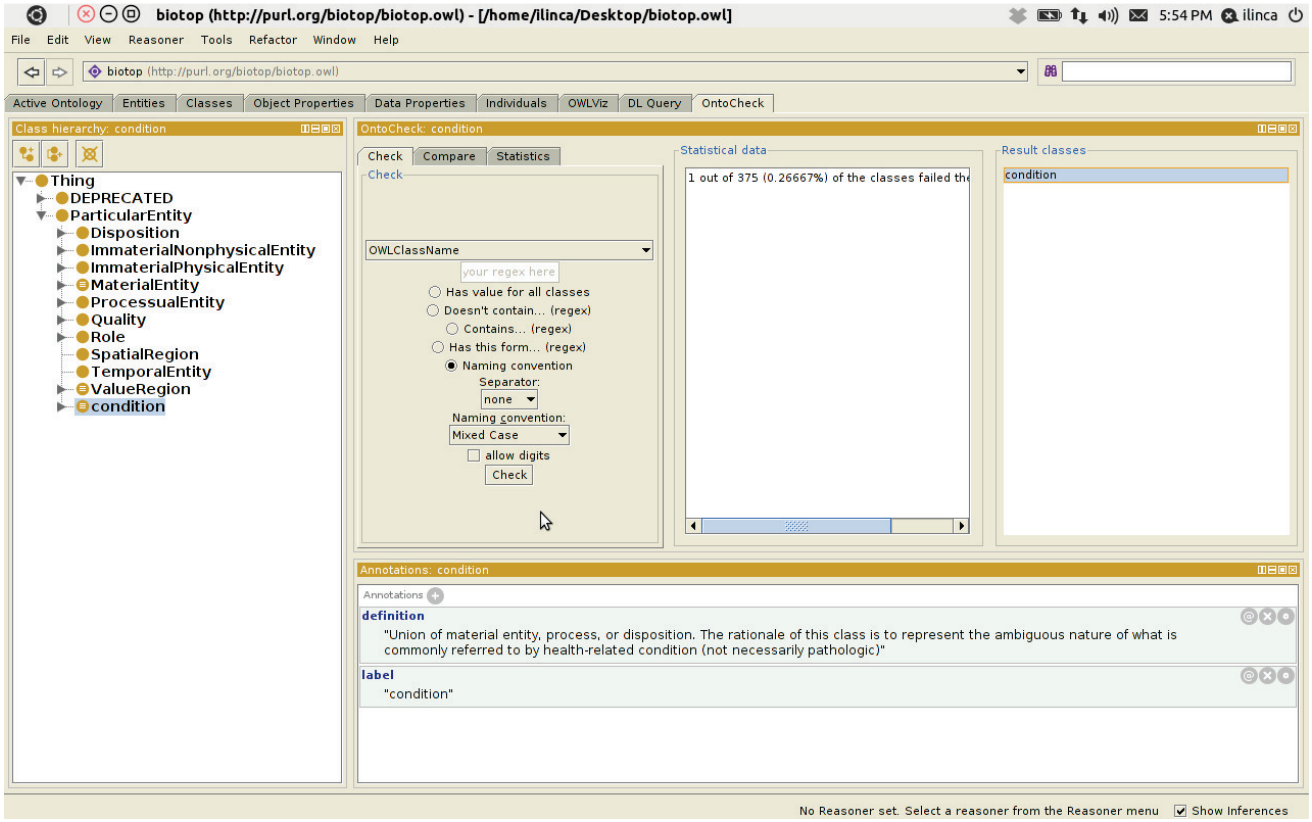


Figure 1: The Check pane within the OntoCheck Tab displays the specification (left) of a test for a MixedCase-no separator naming convention on OWLClassName for the active ontology (biotop). The 'statistical data' view (middle) provides the absolute amount and percentage of classes failing this test. In this case only one class 'condition' violates the convention (by starting lower case). Clicking on it (right) marks it in the class hierarchy pane (left) and opens an edit pane to allow for correction (below). Additional screenshots can be found on the OntoCheck website.

Regular Expressions: to check for presence or absence of a lexical prefix, infix or postfix pattern, selecting e.g. starts with, contains, doesn't contain.

- E.g. all role subclasses should end with the explicit " role"-postfix.
- Class names should not contain Boolean operators, or lexical indications for negations like 'non', 'anti ' or 'dis'.
- Warn on 'metalevel' postfixes like class, '_type', '_concept', and '_relation'.
- Check that qualifier terms (differentia) appear before the part being qualified (genus). E.g. 'NMR_instrument' in place of 'instrument_for_NMR'.
- For object properties check that inverse relations comply to a coherent similar from e.g. for 'has_X' the inverse should be 'is_X_of'. Using (redness) 'has_bearer' (red eye) and (red eye 'is_bearer_of' (redness), which would avoid the hard-to-find relation name 'inheres_in' as inverse of 'is_bearer_of'.
- Checking for punctuation, e.g. if dots are present, allows for the detection of abbreviations, while all upper case can detect acronyms. Else, there should probably be no punctuations in names although sometimes slash and comma are found to narrow down homonyms.

Minimum and Maximum character and word count: to check for potentially unclear names, e.g. being shorter than 4 characters or unreadable names longer than e.g. 50 characters or 10 words.

All specified naming pattern checks can be stored in an external file, as an editor is likely to do the same set of checks on an ontology repeatedly, i.e. before each release The stored Check-list can be exchanged and shared among a group of developers.

The Compare panel

The Compare panel (see webpage for screenshot) allows to compare the values for specified labeling- and metadata-entities. The user first selects a class, chooses two label or metadata annotation properties to compare, e.g. OWLClassName and rdfs:label, and then compares their values for the selected subtree using the 'equals', 'contains'

Ontology	Target Node	RU	Check	Violations [abs, %]
BioTop	root	<rdfs:label>	Upper case start	12 (4)
BioTop	root	<owl:Class rdf:about>	CamelCase	34 (8)
BioTop	root	<ru-meta:definition>	Min card.=1	2 (.5)
DCO	root	<ru-meta:definition>	Min card.=1	37 (8)
DCO	'Disease'	<SNOMED_ID>	Min card.=1	2 (2)
DCO	root	<ru-meta:synonym>	Min card.>2	238 (40)
DCO	root	<ru-meta:label>	Lower case start	4 (3)
DCO	'Drug'	<ATC_ID>	Min card.=1	6 (1)
DCO	root	<ru-meta:shortLabel>	Max Char Count < 20	3 (.5)

Table 1. OntoCheck test cases and detected quantified violations. Target Node refers to the selected Class in the hierarchy for which all subclasses are tested. The RU selected to be checked is described via its owl syntax element. [abs] refers to the amount of RUs of the specified type failing the test. [%] refers to the ratio of abs to the amount of all target node subclasses.

or 'starts with' operator. Case and separator awareness can be checked. As a result, classes with different values in the specified RUs are listed, and can now be rectified.

The Statistics panel

The Statistics panel (see webpage for screenshot) detects and quantifies ontology measures useful for complexity analysis, ontology evaluation and progress monitoring. The developmental version with limited functionality presented here allows displaying the percentage or absolute number of RUs having 'exactly', 'at least' or 'at most' a certain number of subclasses, 'usages' or annotations. This will allow checking classes for having more than one of the same label types, which should not be allowed for any label type other than synonyms.

Further, we list and count all classes with no 'usage' in restrictions other than subclass relations. This allows detecting 'ontological isolates' with no dependencies, which are ignored by any other logical definitions. Such isolates could potentially be removed or hidden in a simplified view of an ontology, focusing on the ontologies' network properties and defined classes linked via object properties. Here, one could order classes according to their 'embeddedness', listing hub-nodes that have many in- and outgoing edges first, as these are likely to represent the more important classes in a formalized domain. As an application is likely to focus on these 'key classes', particular care must be taken to ensure domain coverage in sufficient granularity here.

Given a reasoner is activated, as part of an expressivity analysis, besides counting the defined-to-isolate class ratio, entailment densities and owl flavor element usage could also be counted. Such metrics would enable the judgment whether a certain semantics or OWL flavor was chosen because it is *en vogue* or because it is needed.

To quantify OntoChecks outcome, we carried out an initial statistical analysis in two ontologies, investigating the percentage of found RUs violating each convention-check (Tab 1).

3 DISCUSSION

Rendering labels in ontologies more consistent will pave the way for tools that use lexical information in class names for inconsistency detection, ontology integration, and formalization, e.g. like OBOL [5], which recommends logical definitions for classes by exploiting lexical information from labels. Lexical ontology **alignment tools** such as the PROMPT tool suite [6] will be served with more robust information making automatic alignment and integration easier and more reliable. Recently ontology alignment and transformation techniques have been designed that explicitly rely on naming structures over the ontology graph [7], and thus will particularly benefit from a prior clean-up.

As long as accepted recommendations for standard combinations of single naming conventions are not available, we enable checks on a per-convention basis, rather than allowing checks for defined overall naming convention sets, e.g. the Foundry vs. Manchester vs. Stanford style convention sets. Given these were accessible in a standardized repository, one could envision checks and enforcements of whole naming schemes to be drawn from such libraries. We have recently joined forces with the ontology design pattern community [8] to transform naming conventions into such formal reusable **Naming ODPs**, then accessible to tools like OntoCheck to foster lexical harmonization.

We also investigate the reimplementation of parts of OntoCheck as a webservice in order to foster integration into Semantic Web portals like Watson [9] which would ease reuse for portal and library providers, as semantic metrices can be updated continuously and used for ontology comparison, evaluation, ranking, e.g. helping to select compatible artefacts with similar design principles to be aligned or merged easily.

Future extensions and applications

In its current state OntoCheck works with the active ontology only, ignoring all owl:imports, but this lack can be compensated for checking all namespaces individually. We plan expanding OntoChecks applicability to the whole dependency structure.

At the moment the user has to amend the labels manually, but RUs violating tests could be corrected automatically (OntoCure) in the future.

For an updated list of desired and upcoming features, please visit the OntoCheck webpages.

Check Tab

A future version should:

- Check for **naming clashes** in equal (synonymous) fields for different classes, e.g. if there is a class with equal labels represented with different IDs.
- Check if an **imported ontology differs in naming conventions** from the active ontology.
- Check for violation of **class-subclass naming pattern**, i.e., situations when the head noun of a subclass is not adequately related via a taxonomic correspondence to the head noun of its superclass. It was observed that, for multi-token labels, such violations are nearly always caused either by a modeling error, such as confusion of taxonomy with partonomy, or by a bad naming practice, e.g. 'parsimonious' omission of the true head noun [10]; this is actually a natural consequence of the set-theoretic nature of OWL ontologies.

Compare Tab

In the result list both variant forms will be displayed and the found differences could be highlighted.

Statistics Tab

A future version should:

- Detect abundant pre-, in-, suff- and postfixes and list them according to frequency of occurrence. If a postfix occurs often in siblings, a recommendation could be issued to use this postfix in those labels throughout, or as superclass label.
- Detect logical operators like AND, OR, NOT in names, e.g. BioTop [11] has CarbohydrateMoleculeOrResidue and OligoOrPolymer. These could be potentially correlated with actual logical definitions and disjoints.
- Semantic analysis could probably guide in expressiveness selection, e.g. words indicating cardinality requirements, such as minimal, maximal, exact hint for certain OWL 2 profiles.

4 CONCLUSION

Although in an early development stage, the OntoCeck plugin already proved useful in carrying out pre-release checks for ontologies in different projects [11, 12, 13]. It has helped alerting developers on labeling violations and contributed to keeping these ontologies clean from naming errors. It also rendered our ontologies more complete by curing the lack of metadata. Carried out as pre-release check, the OntoCheck tests contributed to quality assurance [14] in the mentioned projects as has been shown in previous attempts [15]. Ultimately, we hope this Protégé extension will ease lexical post-processing of annotated data and hence increase overall secondary data usage by humans and computers.

ACKNOWLEDGEMENTS

This work was supported by the Deutsche Forschungsgemeinschaft (DFG) grant JA 1904/2-1, SCHU 2515/1-1 GoodOD (Good Ontology Design). Vojtěch Svátek is supported by the CSF under P202/10/1825 (PatOMat). Thank you, Helena R., for the inspiring exchange in Prague.

REFERENCES

[1] The Protégé Ontology Editor and Knowledge Acquisition System, http://protege.stanford.edu/

[2] Smith B, Kusnierczyk W, Schober D, Ceusters W (2006) *Towards a Reference Terminology for Ontology Research and Development in the Biomedical Domain*. KR-MED 2006

[3] Tuason O, Chen L, Liu H, Blake JA, Friedman C (2004) *Biological nomenclatures: a source of lexical knowledge and ambiguity*. Pac Symp Biocomput 2004:238-249.

[4] Schober D. et al. (2009) *Survey-based naming conventions for use in OBO Foundry ontology development*. BMC Bioinformatics, Vol.10, Issue 1, 2009.

[5] Mungall CM. (2004) *Obol: Integrating Language and Meaning in Bio-Ontologies*. Comparative and Functional Genomics, 5:509-520.

[6] Noy NF, Musen MA (2001) *Anchor-PROMPT: Using Non-Local Context for Semantic Matching*. In: Proceedings of the Workshop on Ontologies and Information Sharing, Seventeenth International Joint Conference on Artificial Intelligence (IJCAI-2001), Seattle, WA, SMI technical report SMI-2001-0889

[7] Šváb-Zamazal, O., Svátek, V., Iannone, L. (2010) *Pattern-Based Ontology Transformation Service Exploiting OPPL and OWL-API*. In: EKAW 2010 - 17th International Conference on Knowledge Engineering and Knowledge Management, Lisbon, Portugal. Springer LNCS 6317, 105-119.

[8] Ontology Design Patterns.org (ODP), http://ontologydesignpatterns.org/wiki/Main_Page

[9] d'Aquin M., Gridinoc L., Angeletou S., Sabou M., Motta E. (2007) *Characterizing Knowledge on the Semantic Web with Watson*. In: EON'07 Workshop at ISWC'07, http://watson.kmi.open.ac.uk/editor_plugins.html

[10] Šváb-Zamazal, Svátek V. (2008) Analysing Ontological Structures through Name Pattern Tracking, In: the 16th International Conference on Knowledge Engineering and Knowledge Management (EKAW 2008). Acitrezza 2008, Springer Verlag 2008.

[11] BioTop *A Top-Domain Ontology for the Life Sciences*, http://www.imbi.uni-freiburg.de/ontology/biotop/

[12] The GoodOD Project, http://www.iph.uni-rostock.de/Good-Ontology-Design.902.0.html

[13] Schober D. et. al. (2010) *The DebugIT core ontology: semantic integration of antibiotics resistance patterns.* Stud Health Technol Inform. 2010;160 (Pt 2):1060-4.

[14] Rogers JE (2006) *Quality assurance of medical ontologies*. Methods Inf Med 2006, 45:267-274.

[15] Kismeta Validator v1.1b, *Enterprise Data Standards Validation and Enforcement,* http://www.kismeta.com/Validtr.html

Beyond the Tumour: Breast Cancer Phenotypes

—Towards a pluralistic integration of heterogeneous representations—

Aleksandra Sojic[1,2,3*] and Oliver Kutz[4]

[1]European School of Molecular Medicine, [2]European Institute of Oncology; Milan, Italy
[3]University of Milan, Milan, Italy
[4]Research Center on Spatial Cognition (SFB/TR 8), University of Bremen, Germany

ABSTRACT

Starting from an acknowledgment of the plurality of epistemic motivations driving phenotype representations, our main contribution is a distinction between six categories of *human agents as individuals and groups* focused around particular epistemic interests. We analyse the corresponding impact of these groups and individuals on representation types, mapping and reasoning scenarios, using the example of breast cancer research. We in particular demonstrate a heterogeneity of representation types for breast cancer phenotypes and stress that the characterisation of a tumour phenotype often includes parameters that go beyond the representation of a corresponding empirically observed tumour, thus reflecting significant functional features of the phenotypes as well as epistemic interests that drive the modes of representation. Accordingly, the represented features of cancer phenotypes function as epistemic vehicles aiding various classifications, explanations, and predictions.

1 INTRODUCTION

The representation of phenotypes plays an important role in clinical and biomedical knowledge. Besides functional characterisations, a disease often gets characterised through a distinction between 'normal' and 'abnormal' phenotypes, where 'abnormal' phenotypes often serve as the marks of disease. The 'abnormal' phenotypes associated with a disease are labelled as *phenotypes of disease* (PD). However, the questions of what is 'abnormal' and *what* should be considered as a phenotype of a disease and *how* such a phenotype should be represented are rather contentious. Clearly, the choice of how a PD should be represented is *normative* and *context dependent*. Consider the case of breast cancer and BRCA gene mutations. In the age of genomic medicine, the very definition of disease has changed introducing an asymptomatic diagnosis. So, carriers of BRCA mutation, without having developed any signs of breast cancer, still have a likelihood of over 80% for developing an aggressive cancer phenotype during their life span. Genomic medicine shifts the focus of PD from a traditional organ level approach to the gene level, treating apparently healthy people as 'patients'. For, the 'normal' breast phenotype in a BRCA mutation carrier will be irrelevant in the light of knowledge about 'abnormal', fine-grained phenotypes related to the gene expression patterns of the mutated gene. Although these new directions in biomedicine aim towards an integration of clinical and biomedical knowledge, in most cases the needs of sub-domain knowledge significantly vary. So, a clinician will have different criteria for a representation than a molecular biologist. Regarding the goals of a discipline and the research context, a representation that is relevant for a clinician does not need to satisfy the needs of a molecular biologist who is aiming towards more fine-grained representations. As a result, heterogeneous representations of breast cancer phenotypes were employed in clinical and biomedical knowledge [8, 4, 25].

Taking a very general position, representations of PDs may include images acquired by technologies such as ultrasound, X-ray, and microscopy of histopathological samples. Moreover, representations of PDs are not limited to visual representations, but may include mathematical equations, statistical graphs, molecular markers, microarrays data, and the phenotype specific protein interactions, thus describing PDs according to the needs of and knowledge about a particular domain aspect. In addition, a specific representation of a phenotype should not, in general, be mistaken for the representation of knowledge. Rather, a representation reflects which aspects of knowledge have been targeted by the representation. Accordingly, a representation reflects a scientist's choice of a representation type in order to represent a certain subset of the domain knowledge—therefore, 'choosing a representation' might be a highly intentional act [6]. However, a representation such as a histopathological image will not, itself,

*Corresponding author: aleksandra.sojic@ifom-ieo-campus.it

represent any knowledge unless it gets interpreted. Knowledge within a domain is explicitly represented only if the representations get systematically connected with related interpretations, knowledge claims, and reasoning over the representations. Therefore, besides heterogeneity of PDs, biomedical ontology has to deal with a heterogeneity of reasoning about PDs, comprising different kinds of formal (or logical) representations as well as various types of reasoning. Conversely, the intended reasoning methods or types over PDs also influence the choice of representation of PDs because such representations are mediated by domain specific methods and interventions, employed in the imaging, measuring of the gene expression and other diagnostic techniques [12]. For example, the clinical representation of breast cancer goes beyond the tumour imaging representation. According to the standards of the TNM classificatory system [8], the clinical classification of tumours might consider tumour size (T), lymph nodes involvement (N), and presence of metastasis (M). Of course tumour size is just one feature and is not sufficient for the characterisation of the tumour type. Cancer is a dynamic and complex disease of an organism and the PDs go beyond the characterisation of a tumour's features captured in a static picture. So, for example, knowledge about lymph nodes' status or proliferation marker KI-67 provides additional information about a tumour's phenotype. Likewise, tumour markers provide a view on the PDs through the specific interventions on the representation such as staining samples in order to mark the presence of hormone receptors. Had the estrogen receptor (ER) been detected, the PD would have been described as an ER positive tumour, which significantly differs from an ER− (negative) tumour, which does not respond to the endocrine therapy [7]. Thus, the therapeutic criteria are also considered in the specification of the tumour phenotypes.

2 A PLURALITY OF DOMAIN INTERESTS

Information technologies and formal tools such as ontologies for knowledge representation (KR) are aiming at the integration of heterogeneous knowledge domains and different types of representations. Concurrently, clinicians and molecular oncologists are trying to organise and apply the overwhelming and diverse knowledge about cancer biology. Can these interests of different disciplines meet in a constructive union, while preserving the domain specific representations and reasoning capabilities?

In this and the next section we outline some of the requirements for achieving such a level of interoperability.

We begin by giving a comparative analysis of the distribution and character of knowledge involved in the integration of heterogeneous types of knowledge represented in knowledge bases (KBs). In particular, we distinguish *where*, *how*, and *by whom* knowledge is represented by characterising six epistemic groups, and by discussing how membership to a group impacts the representation as well as knowledge base types. Note that these groups exhibit rich interdependencies and partially overlap.

1. The characterisation of the epistemic groups starts with the societal demands for problem solving, such as, for example, the need for personalised breast cancer therapy. The demands may be represented in the form of standards, platforms and funding policies. In a democratic society, knowledge on this level can be represented as common or shared knowledge available to the members of society; knowledge can be distributed through various channels or common-sense KBs.
2. The second epistemic group to be discussed is at the level of an individual scientist whose 'knowledge base' is a collection of relevant background knowledge, here to be understood as cognitive representations placed in the mind, arguably, in the form of conceptual maps (see [24]).
3. As the third epistemic group, we specify the scientific communities, each of which is composed of the specific disciplinary domain scientists (clinicians, molecular biologists, bioinformaticians etc.). This epistemic group establishes knowledge within a scientific community as a received view, having the form of *explicit* and *inter-subjective* representations expressed in the respective scientific languages, circulated through publications. Like in group (1), knowledge can be distributed in various ways, but related KBs will contain domain specific knowledge.
4. The fourth group comprises scientific communities formed around a particular problem (e.g. breast cancer). As the group contains multidisciplinary teams focused on a particular problem, knowledge will need to be coordinated in such a way that the used scientific terms and reference classes will conform with knowledge within diverse domains. For instance, the biomedical terms might be structured into networks of terms that represent how these terms are interrelated in the domain knowledge. Thus, collaboration here results in merging knowledge from different domains. The representation of the merged knowledge coming from different perspectives on the same problem

might be a 'unified semantic map' (see group (2)) that serves as a semi-formal conceptual model and an intermediate step towards the KB and the formal ontology to be employed in KR.

5. The fifth is the communities of logicians and ontologists who are formalising ontologies according to the needs and specificities of a particular field. Domain knowledge and the merged domain knowledge will be expressed as ontologies written in various formal languages (e.g. refining foundational ontologies such as DOLCE [21], BFO[1], or GFO[2] etc. formalised in OWL[3], first-order logic, etc.)
6. The sixth group involves computer scientists, programmers and engineers, who are designing databases and applying formal ontologies as well as various reasoning tools to large datasets. Technically, a representation built on top of a database involves types and mapping relations *structuring* the data, and can be considered as meta-data. Here the representation integrates the types and mappings with instances (data). Epistemic accuracy of the mappings depends on how well the mappings correspond to the scientific knowledge and the empirical findings of the represented domain (e.g. breast cancer). In contrast to groups (2) and (3), knowledge in a KB is not scattered over various representational spaces or layers, but integrated into one.

Knowledge levels, groups, or layers have of course been discussed previously in the AI literature. For instance, Newell introduced an agent-based distinction between the 'knowledge level' and the 'symbol level' in [23], and [1, 10, 11] analysed layers in formal ontology design. In more detail, Brachman, in 1979, introduced a classification of the primitives used in KR systems at the time [1], distinguishing the following four levels: (i) 'Implementational', (ii) 'Logical', (iii) 'Conceptual', and (iv) 'Linguistic'. Guarino [10, 11] added to these four layers yet another layer, namely the 'Epistemological Layer' for the primitives, situated between the 'Logical' and the 'Conceptual' layers. Our approach differs in that it mainly aims at distinguishing *human agents as individuals and groups* focused around particular epistemic interests, whilst analysing the corresponding impact on representation types. A more detailed analysis of the relationship to previous 'layering approaches' is left for future work.

[1] See http://www.ifomis.org/bfo/

[2] See http://www.onto-med.de/ontologies/gfo/

[3] See http://www.w3.org/TR/owl2-overview/

3 ONTOLOGY INTEROPERABILITY

We next discuss how the six epistemic groups impact on representation types, choice of formalisms, kinds of metadata, mappings, as well as reasoning. We begin by inspecting the notion of an ontology itself.

A plurality of ontologies and formalisms

An often cited definition of the term 'ontology' in computer science was given by Tom Gruber in 1992 [9] (here heavily abridged).

> A conceptualisation is an abstract, simplified view of the world that we wish to represent for some purpose. Every knowledge base, knowledge-based system, or knowledge-level agent is committed to some conceptualisation, explicitly or implicitly.
>
> An **ontology** is an explicit specification of a conceptualisation. [...] For AI systems, "what exists" is that which can be represented. [...] In such an ontology, definitions associate the names of entities in the universe of discourse (e.g., classes, relations, functions, or other objects) with human-readable text describing what the names mean, and formal axioms that constrain the interpretation and well-formed use of these terms. Formally, an ontology is the statement of a logical theory. [9, p. 908–909]

This definition, whilst being controversial, still nicely captures the main differences between the usage of the term 'ontology' in philosophy vs. computer science and artificial intelligence. Namely, consider the following snippets from this definition:

- 'simplified view of the world that we wish to represent for some purpose': an ontology as a technical artefact is not intended to cover the world in its entirety, but only chosen aspects of the world, on specific levels of abstraction, and for given purposes—largely independent of particular metaphysical positions such as realism and antirealism; here, group (4) will typically informally specify the relevant domain knowledge, whilst group (5) is in charge of establishing an agreement on how to formally codify this knowledge.
- 'committed to some conceptualisation': ontologies presuppose various decisions concerning ontological commitments. These originate partly in common sense knowledge (group (1)), precisifications given by members of group (2), and agreements as they are established in groups (3) and (4). Finally, the formal implementation of the ontological commitments is again left for groups

(5) and (6), merging collaborative interests of (1)–(6).

- "'what exists" is that which can be represented': ontological commitments are dependent on the expressive capabilities of selected representational formalisms. The choice of an adequate formal language can only be established as an interplay between logician (group (5)), computer scientist (group (6)), and the domain experts of (3) and (4).
- 'representational vocabulary' and 'human-readable text': there is a 'tension' between the logical vocabulary used, and the natural language concepts and terms it is meant to capture, and, in the case of e.g. breast cancer, various forms of scientific representations such as graphs, mathematical equations, images, 3D models etc. Reconciling this tension requires deep interaction between the various groups of domain experts and formal logicians and computer scientists.
- 'an ontology is the statement of a logical theory': on a technical level, an ontology is seen as equivalent to a logical theory, written in a certain formalism. Clearly, this task is for group (5), respecting the requirements of group (6).

Heterogeneity of formal languages is particularly important in the life sciences, where size of ontologies and needed expressivity vary dramatically. For example, whereas weak (i.e. sub-Boolean) DLs suffice for the NCI thesaurus (containing about 45.000 concepts) which is intended to become the reference terminology for cancer research [26], other medical ontologies such as GALEN[4] require the full expressivity of the OWL language (a decidable fragment of first-order logic), while foundational ontologies typically require at least full first-order logic (see [16]).

An example of a heterogeneous combination of formalisms is discussed in [13], where it is shown that in order to adequately represent the spatial structure of molecules as they are described in chemical ontologies such as ChEBI [2], ontology languages need to be combined with formalisms such as monadic second-order logic. We next investigate how such diversity and heterogeneity is reflected in and how it originates from the different group interests involved in the representation of breast cancer phenotypes.

A plurality of mapping and reasoning types

In biomedical ontologies, metadata in the form of tags, annotation, or more generally documentation, is of particular importance. Indeed, many biomedical

[4] See `http://www.opengalen.org/`

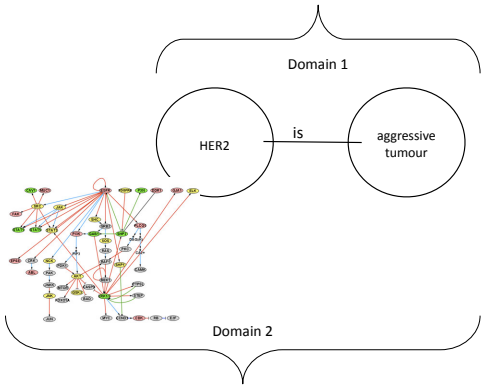


Fig. 1. Knowledge granularity.

ontologies have an extremely shallow logical structure, namely consist only of taxonomies, or even just of sets of concepts, however accompanied with a rich set of metadata. It is clear that the separation of the epistemic groups from Section 2 has a direct impact on the kinds of annotations and metadata that can be expected to be generated. For instance, the particular scientific communities (groups (2) and (3)) need not associate identical sets of concepts as related to a term in use. Had the 'Human Epidermal growth factor Receptor 2' (HER2, also known as ErbB2) been used as a tumour marker in the community of clinical oncologists, it would have been related to the diagnosis of an aggressive tumour with a poor clinical outcome and a low likelihood of a long term survival. On the other hand, among the group of molecular biologists HER2 would be associated with the specific protein-protein interactions that trigger the carcinogenic events.

As interests diverge among and within disciplines concerning ways of describing a problem, distinguishing similarities and difference makers will vary among knowledge domains. So, HER2 will not be the same difference maker for a clinician and for a biologist. The main difference that will be relevant for a clinician will be a difference in the patients survival associated with the expression of HER2 [27]. The biologist who focuses on the cellular signalling pathways might favour a differential expression of the ErbB2 gene while comparing the phenotypes of two types of cell lines [19]. Consequentially, justification of asserted similarities and generalisations will ask for a different kind of evidence in diverse domains. Clinical evidence will be acquired through survival analysis and clinical trials while biologists provide evidence through diverse experimental and explanatory methodologies [18]. Accordingly, the reasoning of

the groups (2)–(4) influence the related mappings and justifications implemented by the groups (5) and (6).

A relation between a term and its reference class gets its justification within domain knowledge as an adequate mapping relationship. The justification is expressed through the claims that support the mapping relations. Regarding the previous example, 'HER2' will be mapped onto a bad prognosis within clinical knowledge, and the mapping will be justified by the statistical data retrieved from the survival analyses (see Fig. 1, Domain 1). Likewise, biological knowledge provides an alternative mapping relation and a related justification to the mapping between 'HER2' and 'tumour aggressiveness', e.g. protein interaction pathways that result in cell proliferation and tumour aggressiveness (see Fig. 1, Domain 2). These diverse patterns of clinical and biomedical reasoning [3] can be perceived as domain specific. A detailed analysis of the mappings within and between knowledge domains asks for a multidisciplinary approach involving a community based process of knowledge production [5]. A group of experts with a common interest is collaborating in establishing standards that help them label and describe the domain of interest [20].

4 DISCUSSION AND FUTURE WORK

Concurrently with the systematisation of epistemic group levels, representation types and knowledge base types, we intend to use the introduced distinctions in order to characterise domain specific knowledge representations for breast cancer phenotypes. Specifically, we are interested in the problem of merging knowledge from different domains and in analysing the 'domain knowledge problems' of [14] further through inspecting a number of examples from molecular oncology and clinical practice. Here, we have demonstrated that such domain problems ask for a plurality of onto-logical formalisms.

We have sketched the intertwined processes involved in the integration of heterogeneous representations as they originate from different epistemic groups that are involved in complex domains such as breast cancer research. Concerning formal representations dealing with the heterogeneities of phenotypes, we propose to endorse a framework that allows to organise the various (domain) representations into an interlinked modular structure, respecting the plurality of formalisms, expressivities and aims, as they are found across diverse scientific communities. A further characterisation of the domain specific epistemic interests, including a deeper understanding of the characterised groups (1)–(6), would provide a more sustainable integration of knowledge about breast cancer, increasing interoperability of represented information and, therefore, applicability of acquired clinical and biological knowledge. A closer understanding of the domain needs would also further support decisions about which formalisms best suit a domain. [15, 22] lay the foundation for a distributed ontology language DOL, which will allow users to use their own preferred ontology formalism whilst becoming interoperable with other formalisms. At the heart of this approach is a graph of ontology languages and translations between them (see [17] for the theoretical development).[5] This graph enables users to:

- relate ontologies that are written in different formalisms with various kinds of mappings,
- re-use ontology modules even if they have been formulated in different formalisms, and
- re-use ontology tools like theorem provers and module extractors along translations.

Indeed, we believe that no attempt at an integration of knowledge can be epistemically sustainable unless it respects the plurality of formal languages and tools, methodologies and perspectives as they result from the heterogeneity of the domain interests.

ACKNOWLEDGEMENTS

Work on this paper was supported by the DFG-funded Transregional Collaborative Research Centre on Spatial Cognition (SFB/TR 8) and by the 'Fondazione Umberto Veronesi' (FUV). We are grateful for the very useful feedback of three anonymous reviewers.

REFERENCES

[1]R. J. Brachman. On the Epistemological Status of Semantic Networks. In N. V. Findler, editor, *Associative Networks: Representation and Use of Knowledge by Computers*. Academic Press, 1979.

[2]P. de Matos, R. Alcántara, A. Dekker, M. Ennis, J. Hastings, K. Haug, I. Spiteri, S. Turner, and C. Steinbeck. Chemical Entities of Biological Interest: an update. *Nucl. Acids Res.*, 38:D249–D254, 2010.

[3]A. D. Evans and V. Patel, editors. *Cognitive Science in Medicine: Biomedical Modeling*. MIT Press, Cambridge, MA, 1989.

[4]D. Faratian, R. G. Clyde, J. W. Crawford, and D. J. Harrison. Systems pathology: taking molecular

[5] DOL is currently under standardisation as Working Draft ISO/WD 17347 in ISO/TC 37/SC 3 'Systems to manage terminology, knowledge and content'.

pathology into a new dimension. *Nature reviews. Clinical oncology*, 6(8):455–464, 2009.

[5]J.-P. Gaudillière and H.-J. Rheinberger, editors. *From Molecular Genetics to Genomics : The Mapping Cultures of Twentieth-Century Genetics*. Routledge, London, 2004.

[6]R. Giere. An agent-based conception of models and scientific representation. *Synthese*, 172(2), 2010.

[7]A. Goldhirsch, W. C. Wood, A. S. Coates, R. D. Gelber, B. Thürlimann, and H. J. Senn. Strategies for subtypes—dealing with the diversity of breast cancer: highlights of the St Gallen International Expert Consensus on the Primary Therapy of Early Breast Cancer. *Annals of Oncology*, 2011.

[8]M. K. Gospodarowicz, B. O'Sullivan, and L. H. Sobin, editors. *Prognostic factors in cancer: International Union against Cancer*. Wiley-Liss, 2006.

[9]T. R. Gruber. Toward Principles for the Design of Ontologies Used for Knowledge Sharing. *International Journal of Human-Computer Studies*, 43(4-5):907–928, 1995.

[10]N. Guarino. The Ontological Level. In R. Casati, B. Smith, and G. White, editors, *Philosophy and the Cognitive Sciences*, pages 443–456. Hölder-Pichler-Tempsky, 1994. Proc. of the 16th Wittgenstein Symposium, Kirchberg, Austria, Vienna, August 1993.

[11]N. Guarino. The Ontological Level: Revisiting 30 Years of Knowledge Representation. In Alex Borgida, Vinay Chaudhri, Paolo Giorgini, and Eric Yu, editors, *Conceptual Modelling: Foundations and Applications. Essays in Honor of John Mylopoulos*, pages 52–67. Springer, 2009.

[12]I. Hacking. *Representing and Intervening: Introductory Topics in the Philosophy of Natural Science*. Cambridge University Press, 1983.

[13]J. Hastings, O. Kutz, and T. Mossakowski. How to model the shapes of molecules? Combining topology and ontology using heterogeneous specifications. In *Proc. of the Deep Knowledge Representation Challenge Workshop (DKR-11), K-CAP-11, Banff, Alberta, Canada*, 2011.

[14]A. Hunter and R. Summerton. A knowledge-based approach to merging information. *Knowledge-Based Systems*, 19(8):647–674, 2006.

[15]O. Kutz, T. Mossakowski, C. Galinski, and C. Lange. Towards a Standard for Heterogeneous Ontology Integration and Interoperability. In *International Conference on Terminology, Languages and Content Resources (LaRC-11)*, Seoul, South Korea, 2011.

[16]O. Kutz, T. Mossakowski, J. Hastings, A. Garcia Castro, and A. Sojic. Hyperontology for the Biomedical Ontologist: A Sketch and Some Examples. In *Workshop on Working with Multiple Biomedical Ontologies (WoMBO at ICBO 2011)*, Buffalo, NY, USA, August 2011.

[17]O. Kutz, T. Mossakowski, and D. Lücke. Carnap, Goguen, and the Hyperontologies: Logical Pluralism and Heterogeneous Structuring in Ontology Design. *Logica Universalis*, 4(2):255–333, 2010. Special Issue on 'Is Logic Universal?'.

[18]A. La Caze. The role of basic science in evidence-based medicine. *Biology and Philosophy*, 26(1):81–98, 2011.

[19]M. Lacroix and G. Leclercq. Relevance of breast cancer cell lines as models for breast tumours: an update. *Breast cancer research and treatment*, 83(3):249–289, 2004.

[20]S. Leonelli. Bio-ontologies as Tools for Integration in Biology. *Biological Theory*, 3(1):7–11, 2008.

[21]C. Masolo, S. Borgo, A. Gangemi, N. Guarino, and A. Oltramari. WonderWeb Deliverable D18: Ontology Library. Technical report, ISTC-CNR, 2003.

[22]T. Mossakowski and O. Kutz. The Onto-Logical Translation Graph. In *Modular Ontologies—Proc. of the Fifth International Workshop (WoMO 2011)*, volume 230 of *Frontiers in Artificial Intelligence and Applications*, pages 94–109. IOS Press, 2011.

[23]A. Newell. The Knowledge Level. *Artifical Intelligence*, 18(1):87–127, 1982.

[24]J. D. Novak and A. J. Cañas. The Theory Underlying Concept Maps and How to Construct Them. Technical report, Florida Institute for Human and Machine Cognition, 2008.

[25]C. M. Perou, T. Sorlie, M. B. Eisen, M. van de Rijn, S. S. Jeffrey, C. A. Rees, J. R. Pollack, D. T. Ross, H. Johnsen, L. A. Akslen, O. Fluge, A. Pergamenschikov, C. Williams, S. X. Zhu, P. E. Lonning, A.-L. Borresen-Dale, P. O. Brown, and D. Botstein. Molecular portraits of human breast tumours. *Nature*, 406(6797):747–752, 2000.

[26]N. Sioutos, S. de Coronado, M. W. Haber, F. W. Hartel, W.-L. Shaiu, and L. W. Wright. NCI Thesaurus: A semantic model integrating cancer-related clinical and molecular information. *Journal of Biomedical Informatics*, 40(1):30–43, 2007.

[27]D. J. Slamon, G. M. Clark, S. G. Wong, W. J. Levin, A. Ullrich, and W. L. McGuire. Human breast cancer: correlation of relapse and survival with amplification of the HER-2/neu oncogene. *Science*, 235(4785):177, 1987.

Information, reality and epistemology: an ontological take

Maurício Almeida and André Andrade*

School of Information Science – Federal University of *Minas Gerais* – Brazil

ABSTRACT

Medical records are crucial resources for the whole healthcare practice. The amount and complexity of information they bear require the use of automation. In this paper we present a framework to represent information recorded in medical records, drawing on Popper three worlds theory. Then, we test such a framework by using a description of a real clinical case. Finally, we offer recommendations of how data can be properly arranged in order to incorporate assorted representations like ontologies, information models and reasoning rules.

1 INTRODUCTION

The medical record is a complex document employed for several purposes in the healthcare realm. Proper documentation of medical encounters is an important task of a physician's activity. Medical records have a myriad of uses in healthcare processes, such as [11]:

- to support patient care: to remind staff of and communicate information, to help in organizing the care process (e.g. care information used in process coordination, clinical decision making, patient demographics);
- to fulfill external obligations: legal requirements, accreditation, reimbursement regulations (e.g. procedure coding), order documentation (e.g. exams, medication), and events (including adverse events, surgeries, sample collections);
- to support administration: in planning, controlling, and refunding the health care institution's services (e.g. medication and medical materials used, equipment use, procedure coding, diagnostic coding, name of professionals);
- to support quality management: by enabling critical assessment and systematic monitoring of processes (e.g. clinical outcomes);
- to support scientific research: by enabling patient selection and statistical analysis (e.g. possibly relevant clinical information, not yet used in clinical reasoning, according to research protocols);
- to support clinical education: by providing information for critical review and case examples (e.g. contextual information about consultation setting).

As a consequence of those multiple uses, medical information is a mix of facts, impressions, measurements, rules, and knowledge recording. A classification of kinds of information is required for automatic processing by computers, as well for system interoperability.

Formal ontology allows robust reasoning, but restricts representation to reality entities. Non-realist information (called here "epistemological information") [5] comprises some representations of symptoms, since there is no way to ascertain the truth value of these assertions: "Neither signs nor symptoms form a natural kind, but are rather composite classes – fiat collections of bodily features delineated by certain socially established cognitive practices on the parts of clinicians and patients" [20].

In order to integrate realist and epistemological oriented information, one must clearly define what such kinds of information mean in medical records, why they are important and which sort of automatic operations they should support. Moreover, a clear separation between, on one side, the entities in reality, and on the other side, the information about them, makes easier the understanding of medical records, allowing different logical operations and the use for different purposes.

The goal of the present paper is to explore better approaches to represent information registered in medical records, taking advantage of the best characteristics of well-known techniques. In seeking such goal, we rely on philosophical grounds in order to create a framework of analysis. Then, we test the framework against a sample of medical records distinguishing within it: i) references concerning real entities; ii) reference concerning epistemological entities; iii) other kinds of information contained in the record that are relevant to the clinical practice. Finally, we offer a proposal of a general arrangement encompassing all those representations into information systems.

2 METHODOLOGY

In order to reach better possibilities of medical record representation, we need to organize the kinds of information they enclose. We here take advantage of well-known techniques for dealing with medical information, like ontologies and information models. The methodology is composed by the following steps.

First, we develop a framework of analysis, which draws on inputs from philosophy, particularly, from Karl Popper´s

* To whom correspondence must be addressed.

three worlds and its usefulness in health information science [3]. We also consider recent researches on the medical ontologies, namely, the Basic Formal Ontology (BFO) [10] the Ontology of General Medical Science (OGMS) [20] and the Information Artifact Ontology (IAO) [12]. Despite empirical evidences suggesting the feasibility of such approaches, different views can be found in the literature [13, 17]. In addition, we take into account other significant advances in binding realist ontologies and information models.
As second step, we test such framework over a complex clinical history developed by The New England Journal of Medicine. We choose that source for didactic reasons, benefiting from the journal´s academic focus, which summarizes clinically useful information. Everything represented in the summary is important to physicians and therefore all entities are considered in our scope of computer processable information. The record was analyzed with the aim of identifying underlying propositions.
In order to identify propositions, a domain expert transcribed the records in sentential fragments that make sense for him. The domain expert was asked to identify the reason for recording those entities and the information that is being conveyed by the representation. The transcription draws upon principles of logic and controlled languages [8, 9], which allowed identification of entities recorded in natural language, outside the particular context in which the event took place [23]. In addition, on the classification side, we use the rationale underpinning OGMS. On the logical side, we took in account that some natural language parts of speech do not have room in logical statements. Even though this is a well-known fact, for example with respect to an adverb, the very same one may be relevant to characterize a clinical situation.
Finally, we took apart the found according to their suitability to each approach. Thus, we organize the information of the medical record in four kinds, which are so employed to recommend both a data arrangement and a scenario of collaboration among different representations.

3 FRAMEWORK OF ANALYSIS

There is no consensus about the best way to represent the myriad of situations that occurs during a medical encounter. A useful approach relating reality, cognition and representations was proposed by Popper in his theory of three worlds [15]. Those worlds are described as follows:

World 1: the physical world;
World 2: the world of mental states;
World 3: the world of contents of thought.

Within those three worlds, objects are real on their own and each one can modify each other. One example is the learning of a new language, which is a modification of World 2 (the process of learning) by a World 3 entity (language itself). Popper's theories receive critics [3], but also favorable claims in which it is considered a useful model to understand epistemic information [1].Accordingly, one can find additions and improvements of the Popper´s views, which propose additional sub-divisions into the original layers [4, 14]. A complete discussion of such a theory is, though, beyond the goals of this paper.
Then, we propose a framework of analysis as depicted in Fig. 1, which was created to organize information according the best possibility of representation.

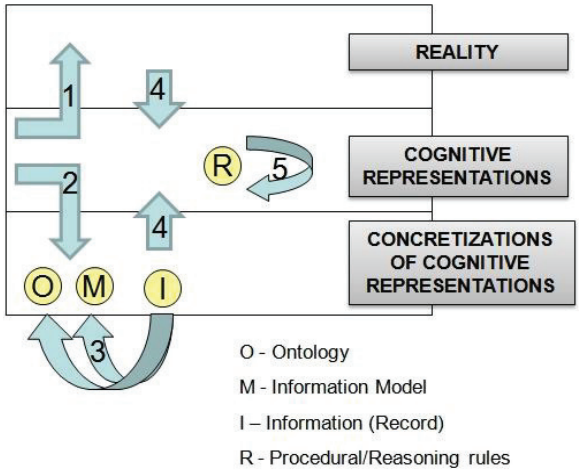


Figure 1 – Framework use for analysis

In this framework, everything begins at level of cognitive representations when a physician observes the reality at the patient side (arrow 1). Each of these entities are filtered by cognition and represented by 2 artifacts (arrow 2). Ontological entities (entity O) are analyzed according to strict philosophical tenets, and are based on reality itself rather than mental representations of a physician. Examples of ontological entities are cells, anatomical features and chemical substances. Information model entities (entity M) stand for cognitive representations of reality, and may include entities without a referent in reality. Examples of these include "severity" of pain and a "feeling well" sensation. Then, the physician creates a record (entity I) to register those representations according to his practical and theoretical knowledge (arrow 3). Constantly, other physicians can interpret records and reality (arrow 4), resulting in new cognitive representations. Finally, the physicians involved in healthcare make judgments, and process past and current information. Some of this processing of information (arrow 5) follows medical training rules, which determine the likelihood of a diagnosis, the correct interpretation of a exam result, to mention but a few. The representation of this process of reasoning is also required for care continuation, a complementary part of the record (entity R). Examples of this include rules to interpret lab data, as hemoglobin level < 12 g/dl means "low hemoglobin level"; and relevant nega-

tive information such as "lack of bowel alteration during episodes".

Our approach to the model is rather more pragmatic – our goal here is to establish a methodology to distinguish real entities from epistemological entities represented as information entities in the World 3. Within the framework created we recognize at least four kinds of information to be separated according to their suitability for information systems:

- Data that represent aspects of the reality;
- Data that represent useful constructs for the medical practice not empirically verifiable;
- Data that represent observations about the reality, not reality itself;
- Data that represent observations about the physician understanding of the clinical situation, not about reality.

Under this model, we still contend to the fact that neither representations of the reality nor representations of thought processes are interpreted in the same way by two people. However, allowing manipulation of the World 3 entities is fundamental for the development of new features in medical systems, such as decision support, inferences and information classification, and discovery.

4 TEST OF THE FRAMEWORK

We here make a preliminary test of the framework by analyzing individual information entities contained in medical records. Figure 2 depicts a small extract of the clinical case available at http://www.nejm.org/multimedia/interactive-medical-case [19]. Once we obtain a sentential fragment from a domain expert evaluation, we thus isolate what could be represented in realism-based ontologies following the rationale of OGMS, BFO and AIO. After that, we arrange other information according to kinds mentioned in section 3. The final results systematize the information contained in a medical record in keeping with the information system that it is suitable for.

"An 88-year-old woman presented to the emergency room with confusion. She began having transient episodes of confusion, dizziness, tremors and anxiety a year earlier. These episodes were unpredictable, lasting for minutes and then abating spontaneously, and had been increasing in frequency since they began. The patient felt well between episodes and reported no abnormal sensation, change in weight, or relation of symptoms to meals, fasting or physical activity."

Figure 2 – Extract of the medical history

In what follows, we present samples of data obtained from the medical record and classified according to kinds proposed in section 3. Fig. 3, 4, 5 and 6 depict such samples.

Data representing aspects of the reality
Physician (BFO Role)
Woman (BFO - Object)
88 years-old (BFO Quality)
Patient report (AOI Information Content Entity)
Confusion, dizziness, tremor (OGMS symptom)
Duration of episodes (BFO temporal region)
Time between episodes (BFO temporal region)
Change in weight (OGMS symptom)
Aspirin (BFO continuant)
Aspirin taken daily (AIO rule)
Physical exam finding of that encounter (OGMS Physical examination finding)
Glucose (BFO Continuant)
Diagnosis of hypoglycemia (OGMS diagnosis)
Insulinoma (BFO continuant)

Figure 3 – Data sample: realist bias

Data that represent useful constructs for the medical practice
… transient episodes of confusion, dizziness, tremors, and anxiety a year earlier (each episode being correlated as caused by a single entity)
No abnormal sensation
… episodes are unpredictable
Confusion
General: well appearing
Chest: clear to auscultation
Abdomen: soft and nontender

Figure 4 – Data sample: data not empirically verifiable

Data that represents observations about the reality
Frequency of episodes
Increase in the frequency of episodes
36º of temperature
76 beats per minute
114/60 mmHg
Glucose concentration
Aspirin dosage

Figure 5 – data sample: observation of the reality

Data that represents observations about the physicians understanding
Insulinoma causing hypoglycemia
Relation symptoms vs. meals

Figure 6 – Data sample: observations of one´s understanding, not reality itself

This data classification was based on both the levels of representation provided in section 2 and the explanation provided in section 3. From the empirical assessment by physicians, the categories suggested from Fig. 3 to Fig. 6

were created. The relation between the proposed framework and the organization of data from medical records can be summarized as follows:

- "Data representing aspects of reality" (Fig. 3) were mapped from processes (1) and (2) to entities (O) (Fig.1);
- "Data that represent useful constructs for the medical practice" (Fig.4) were mapped from the process (1) and (2) to entities (M) (Fig.1);
- "Data that represents observations about the reality" (Fig. 5) were mapped from process (3) to entities (I) (Fig.1);
- "Data that represents observations about the physicians understanding" (Fig. 6) were from processes (4) mapped to entities (R) (Fig.1).

According to the scenario developed so far, we propose a data arrangement to deal properly with all these kinds of data. The data could then be processed by the suitable system and the equivalent representation. The arrangement of data and a scenario of collaboration different systems are depicted in Fig.7:

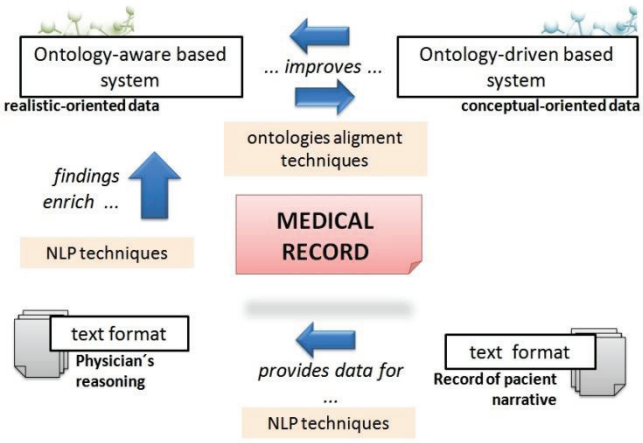


Figure 7 – Final arrangement of medical data

5 DISCUSSION

Medical practice is still heavily grounded in the study of signs and symptoms, which are interpreted by a physician in the search for a diagnosis about a clinical situation. Medical reasoning is a sum of different cognitive practices including induction, abduction and deduction [16]. As a written representation of a medical encounter, a medical record is closely related to medical reasoning practices and how a physician understands pathological process at that time. However, interpretation by computers and semantic interoperability [2] require explicit and shared definition of terms, in order that they can be manipulated without information loss. Our framework attempts make it possible by making clearer the distinctions between reality, medical understanding and the recording of it, while maintaining the medical record as the main data source.

The first important distinction is the separation between information about reality and reality itself. In the context of medical practice, common mistakes have been found when there is no separation of information [6] [22]. For example, a "cancelled surgery" is not a "surgery" which never existed, but a "plan" for a "surgery" which had the content modified". This strategy can be used to allow proper representation of non-existing entities in DL-logic, the most currently embraced family of logic, while maintaining consistency and coherence with realist ontologies tenets [21]. Also, we follow [18] in the description of clinical situations for a useful way of expressing modality, temporal status of symptoms, signs or diseases, and for representing the "subject of care" construct, which is very important for care processes reasons.

Even though one can consider such distinction trivial, we observe that medical systems are currently being developed world-wide without consider what "sort of data" is suitable to what "sort of system". We believe that the commitment to standards addresses only a small piece of the interoperability problem. Our belief is based on the observation that such interoperability problem is yet critical today, despite several standards that have been proposed throughout the years. Our arrangement is an attempt to explore other possibilities of representation for different kinds of systems, considering the real collaboration among them (Fig.7).

With this arrangement proposal, we suggest that it is useful to distinguish what should and what shouldn't be rigorously represented as ontological entities. Our framework suggests that, while the medication (and the analyzed sample) and chemical entities (blood and glucose) are real entities, the results are in fact information about them (data items). For example, the blood glucose measurement refers to the glucose blood concentration at the exact moment of blood sample collection. It is, therefore, empirically verifiable. However, the value of the measurement does not refer to the existence of the enzyme in the real world, and the same entity in reality may be described using different measurement units, laboratory methods and confidence intervals. Besides, the information is analyzed using a sequence of pre-established thinking rules, according to clinical training. For example, in the clinical case present in section 4, the value 40 mg/dl is below normal values (80 mg/dl) and, therefore, suggests the diagnosis of hypoglycemia. It will be interpreted according to a reasoning rule, not according to the structure of reality itself, and is suitable to procedural operations instead of pure logical reasoning. The same reasoning principle holds for other kinds of rates (beats/minute, mg/kg). It is important to emphasize that these rules of interpretation (normality levels) are also based on historical

events which had an almost arbitrary definition of normality [23], and may change at any time.

A pragmatic look at the medical record and the categories of our analysis has shown some aspects that current medical systems solve reasonably well using current relational databases, such as medication dosage and laboratory analysis results. Computerized Patient Order Entry (CPOE) systems are relatively widespread and have successfully replaced free-text orders, though actual improvements in healthcare processes haven't yet come to full extent [7].

Finally, one can argue about examples presented (section 4). For example, the entity "confusion" is exhibited both in data with realist bias (Fig. 3) and data not empirically verifiable (Fig. 4). However, the former represents a condition as reported by a patient; the latter represents a physician´s perception of a patient condition. Also, in Fig. 3, one can claim that diagnosis is not exactly entity pertaining to reality. However, in our framework based in OGMS, a diagnosis is taken as a data record.

6 FINAL REMARKS

In this paper, we present a distinction that includes the representation of entities referring to the physician's or the patient's understanding of a situation. Otherwise said, they represent reality as seen and interpreted by a human being, therefore not objective statements. In the focused record, we found many instances where this distinction is beneficial. We argue that, by separating entities as proposed, one is able to safely talk about the clinical case without harming the interoperability of the medical record. We also suggest an arrangement able to encompass these proposed kinds of data and related systems.

The presented approach is an attempt towards the clarification of critical aspects of data categories. Certainly, it needs more progress in order to have direct impact on the interoperability issue. As future work, we intend to create clear rules to divide kinds of information in a semi-automatic fashion. Then, it will be possible to test our approach against a greater sample. In seeking this, we aim to explore the best characteristics of different systems and data representations.

REFERENCES

1. Abbott, R. Subjectivity as a concern for information science: a Popperian perspective. *Journal of Information Science*, *30* (2). 95-106.
2. Almeida, M.B., Souza, R.R. and Fonseca, F. Semantics in the Semantic Web: a critical evaluation. *Knowledge Organization Journal*.
3. Bawden, D. The three worlds of health information. *Journal of Information Science*, *28* (1). 51-62.
4. Bhaskar, R. *A Realist Theory of Science*. Harvester Press, Sussex, 1978.
5. Bodenreider, O., Smith, B. and Burgun, A., The Ontology-Epistemology Divide:A Case Study in Medical Terminology. in *3rd Conference on Formal Ontology in Information Systems*, (Turin, 2004).
6. Brinkman, R.R., Courtot, M., Derom, D., Fostel, J.M., He, Y., Lord, P., Malone, J., Parkinson, H., Peters, B., Rocca-Serra, P., Ruttenberg, A., Sansone, S.A., Soldatova, L.N., Stoeckert, C.J., Jr., Turner, J.A. and Zheng, J. Modeling biomedical experimental processes with OBI. *J Biomed Semantics*, *1 Suppl 1* (22). S7.
7. Eslami, S., de Keizer, N.F. and Abu-Hanna, A. The impact of computerized physician medication order entry in hospitalized patients--a systematic review. *Int J Med Inform*, *77* (6). 365-376.
8. Fuchs, N.E., Hofler, S., Kaljurand, K., Rinaldi, F. and Schneider, G. Attempto controlled english: A knowledge representation language readable by humans and machines. *Reasoning Web*, *3564*. 213-250.
9. Fuchs, N.E., Schwertel, U. and Torge, S. A Natural Language Front-End to Automatic Verification and Validation of Specifications, LMU München, 1999.
10. Grenon, P., Smith, B. and Goldberg, L. Biodynamic ontology: applying BFO in the biomedical domain. *Stud Health Technol Inform*, *102*. 20-38.
11. Haux, R., Knaup, P. and Leiner, F. On educating about medical data management - The other side of the electronic health record. *Methods of Information in Medicine*, *46* (1). 74-79.
12. IAO, 2011. Information Artifact Ontology. http://code.google.com/p/information-artifact-ontology/
13. Merrill, G.H. Ontological realism: Methodology or misdirection? *Applied Ontology*, *5*. 79-108.
14. Niiniluoto, I. *Critical scientific realism*. Oxford University Press, New York, 1999.
15. Popper, K.R. and Eccles, J.C. *The Self and Its Brain: An Argument for Interactionism*. Routledge, 1977.
16. Pottier, P. and Planchon, B. Description of the mental processes occurring during clinical reasoning. *Revue De Medecine Interne*, *32* (6). 383-390.
17. Rector, A. Knowledge Driven Software and "Fractal Tailoring": Ontologies in development environments for clinical systems *Proceeding of the 2010 conference on Formal Ontology in Information Systems: Proceedings of the Sixth International Conference (FOIS 2010)*, IOS Press, Toronto, Canada, 2010.
18. Rector, A.L. and Brandt, S. Why Do It the Hard Way? The Case for an Expressive Description Log-

ic for SNOMED. *Journal of the American Medical Informatics Association*, *15* (6). 744-751.
19. Ross, J.J., Vaidya, A. and Kaiser, U.B. Lying Low, New England Journal of Medicine, 2011.
20. Scheuermann, R.H., Ceusters, W. and Smith, B., Toward an Ontological Treatment of Disease and Diagnosis. in *2009 AMIA Summit on Translational Bioinformatics*, (San Francisco, CA, 2009), 116-120.
21. Schulz, S., Brochhausen, M. and Hoehndorf, R., Higgs bosons, mars missions, and unicorn delusions: How to deal with terms of dubious reference in scientific ontologies (forthcoming). in *ICBO 2011*, (Buffalo, USA, 2011).
22. Schulz, S., Schober, D., Daniel, C. and Jaulent, M.C. Bridging the semantics gap between terminologies, ontologies, and information models. *Stud Health Technol Inform*, *160* (Pt 2). 1000-1004.
23. Vickers, A.J., Basch, E. and Kattan, M.W. Against diagnosis. *Annals of Internal Medicine*, *149* (3). 200-203.

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 3LGM²
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 Man-tovani, 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 3LGM² 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 (3LGM² V3)
2/2010	H. Herre, R. Hoehndorf, J. Kelso, S. Schulz	OBML 2010 Workshop Proceedings, Mannheim, September 9-10, 2010