

Ontological analysis of controlled vocabularies used in PSI/MSI supported XML standards

Daniel Schober¹, Gerhard Mayer³, Annick Moing⁴, Martin Eisenacher⁵, Steffen Neumann²

1, 2 Department of Stress- and Developmental Biology, Leibniz Institute of Plant Biochemistry, Weinberg 3, 06120 Halle, Germany

4 Metabolome Facility of Bordeaux Functional Genomics Centre, Centre INRA de Bordeaux, 33140 Villenave d'Ornon, France

3, 5 Medizinisches Proteom Center (MPC), Ruhr-Universität Bochum, D-44801 Bochum, Germany

dschober@ipb-halle.de 1
sneumann@ipb-halle.de 2
mayerg97@ruhr-uni-bochum.de 3
moing@bordeaux.inra.fr 4
martin.eisenacher@ruhr-uni-bochum.de 5

Abstract: Besides a plethora of formal ontologies, the requirement for simple data annotation has led to an increased use of so called controlled vocabularies (CV) in multiple omics communities. We analyze two of those CVs from an ontological viewpoint, highlight typical modelling errors and propose more adequate solutions. Discovered errors are discussed in the light of the OOPS ontology pitfall framework and the OBO Foundry naming conventions. As a result the CVs could be improved and the OOPS catalogue could be amended and expanded with new, previously missing error categories. In an outlook we discuss potential reasons for the error prevalence and analyse what criticism is justified for CV semantics and what 'errors' are more valid for formal ontologies rather than CVs. We conclude that although many design principles valid for description logics ontologies are not relevant for semantically flat CVs and in turn there is a need for CV-best-practices that are not appropriate for description logics ontologies, there is room for improvement in the analysed CVs. The scope difference between CVs and formal semantics however should affect policy providers, which should narrow down the scope of their policies, i.e. by stating for each policy the expressivity regime for which it is valid.

Introduction

In recent years controlled vocabularies have gained widespread use in the Proteomics community. Here, the CVs support XML data standards with fixed terms to describe variant experimental metadata [MCS+11]. The combined use of an XSD (XML schema definition) branching out into a CV is defining a terminological standard for proteomics data to be stored as XML in repositories, e.g. PRIDE [VCC+13] and Metabolights [HSC+12]. This standardized data can then be uniformly accessed to compare different

experiments (information retrieval). It also fosters information extraction, i.e. the data can be processed further via vendor-independent open source tools that parse the common standard, e.g. ProteoWizard [KCB+08]. A third use case is quality assurance, i.e. the use of validators to enforce consistent CV term usage [MKR+09] in an XML data record and to allow publishers to test for Minimum Information (MI) completeness according to MIBBI (minimum reporting guidelines for biological and biomedical investigations) [TFS+08].

Compared to the proteomics domain, metabolomics data standardization is still in its infancy. Although knowledgebases with extensive search and query capabilities exist for nuclear magnetic resonance spectroscopy (NMR) data [FGD+11], they are not based on agreed-upon XML standards, nor are they using a common vocabulary. As a result they rather represent insular solutions and hence do not allow cross database queries. To alleviate this drawback and serve metabolomics data integration, the recent COSMOS (Coordination of Standards in Metabolomics) EU project¹ will leverage on the existing Proteomics Standardization Initiative (PSI) data annotation and verification setup and develop an open NMR data standard to be approved by the Metabolomics Standards Initiative (MSI) [SFG+07]. COSMOS will also contribute to the PSI mass spectroscopy standards development. This convergence allows to leverage on the existing PSI infrastructure and ultimately will contribute to a more integrated and systems biologic view of the molecular biology domain. The COSMOS standards development work package² will implement an XSD-based exchange format and CVs needed to describe and exchange, but also to query and verify both metabolomics core and contextual data. Future CV development within the COSMOS effort will profit from an analysis of modelling errors within the existing PSI and MSI CVs in order to improve their quality and that of the successor CVs yet to be developed by COSMOS.

Methods

The two main PSI and MSI CVs used within the scope of COSMOS, namely the PSI-Mass Spectroscopy (MS) CV [MMO+13] and the MSI-NMR CV³ were downloaded in their latest versions (as of May 2013). The PSI-MS CV provides terms required for the annotation of mass spectroscopy analysis. The MSI-NMR CV was developed to annotate NMR raw data in metabolomics databases. All terms in both CVs were subjected to a manual analysis of modelling and labelling errors. The CVs were analysed in their native formats by traversing through the term hierarchy manually, looking at each term's label, superclass, relational embedding and definition. OboEdit 2.2 was used to visualize the term hierarchy. Labels were analysed according to the OBO Foundry naming conventions (NC) [SSL+09] using the OboEdit build-in verification manager. Modelling errors were checked manually and listed using an own combined error identifier scheme (a letter to represent the CV and a number to identify the error). These were later

¹ <http://www.cosmos-fp7.eu>

² <http://www.cosmos-fp7.eu/wp2>

³ <http://www.metabolomicscentre.ca/exchangeformats.htm>

mapped (Tab.1) onto error types as defined in the OOPS (Ontology Pitfall Scanner)⁴ error catalogue. Found errors and suggestions for improvement were fed back to the CV developers. For unclassifiable errors new pitfall categories were submitted to the OOPS catalogue maintainers. We particularly applied description logics (DL) best practices to the CVs in order to allow a later crosstalk with the large amount of domain, but also top level and upper level ontologies available from respective ontology libraries in the W3C recommended OWL-DL format. Additional material, i.e. the concise set-up description of the PSI/MSI CV usage and an expanded analysis including further CVs, can be found on our website at <http://msbi.ipb-halle.de/msbi/OntologicalAnalysis>

Results

Criticism of the CVs from an ontologist's standpoint

A. Ontological errors found in the PSI-MS CV

1. The PSI-MS **imports ontologies but then makes little use of them**, i.e. a phenotypic quality ontology⁵ (PATO) is imported, but the imported `bearer_of` or `inheres_in` relations are not used to couple entities to their qualities. The disuse of the quality upper level class further lead the CV authors to source out the definition of a thing's properties from the representational to the terminological level, e.g. using 'instrument attribute' part of 'instrument', rather than formally defining qualities of instruments via axioms such as 'Instrument' `bearer_of` 'Instrument quality'. Most of the logics regarding qualities is put into external mapping files. If the DL expressivity provided by the imported artefacts would be used, many constraints that now sit in the mapping files could be put right into the ontology.

The following assertion displays a misuse of qualities of representational artefacts related to the one mentioned:

```
'electromagnetic radiation chromatogram' is_a 'chromatogram type'  
'chromatogram type' part_of 'chromatogram'
```

It seems the authors see the chromatogram type as an attribute or quality rather than the class for chromatograms. For the first case it should be modelled as outlined above (via qualities), for the second case it should be modelled by direct subclassing, i.e. 'electromagnetic radiation chromatogram' `is_a` 'chromatogram', removing the "type"- or "attribute"-postfixed classes.

2. The **definition does not harmonize with the term classification**, i.e. for 'electromagnetic radiation chromatogram', the definition reads "The measurement of electromagnetic properties as function of the retention time". Here, the 'measurement' head noun suggests a process interpretation, but what the authors are interested in is the `information_content_entity` that is the result of such measurement process.

3. Looking at the partonomic structure 'Sample/sample attribute/sample preparation/MALDI matrix application/matrix application type', it can be argued whether a **process** (sample

⁴ <http://oeg-lia3.dia.fi.upm.es/oops/index-content.jsp>

⁵ http://obofoundry.org/wiki/index.php/PATO:Main_Page

preparation) can be **part_of an attribute** (sample attribute). A logically sound representation would be: 'sample attribute' inheresIn 'Sample', and 'Sample' outcomeOf 'SamplePreparation'.

4. We find 'matrix solution' is_a 'MALDI matrix application'. 'matrix solution' is defined as "Describes the chemical solution used as matrix". We regard the implied '**material object is_a process**' statement as ontologically incorrect.

5. **No Top Level Ontology (TLO) usage.** This can cause problems, e.g. for the term 'source' it is not clear what its hypernym or corresponding ontological superclass is. In a TLO one would make 'source' a 'role'-subclass, and the name would be sufficiently general.

6. In the PSI-MS CV however the authors intend to model an 'ion source'. This **specific meaning should be made clear in explicit naming** ([SSL+09], Naming Conventions, NC 1.1 Use explicit and concise names), as we assume an annotation to be clear and intuitive even when looking at annotated data when the full ontology is not at hand.

7. The lack of a rigid top level structure further leads to quite **specific terms**, such as 'peptide spectrum match scoring algorithm', **residing in root-near top level positions**, whereas in ontologies one would expect to find the very general more domain-independent terms.

8. There is a 'role type' class made a subclass of contact attribute. This should be renamed to reflect that this role is specifically referring to contact roles. But also there is a 'contact role' in the CV already. The **difference in scope** between these two roles **should be made explicit**.

9. Some **definitions are tautological**, e.g. for 'object attribute' or for 'chromatogram type' where the definition reads "Broad category or type of a chromatogram". Good ontological practice demands Aristotelian definitions, i.e. of the form An A is a B, which Cs, where B is the superclass and C the differentia criterion [SSL+09].

10. There is **editorial metadata encoded by means of terms**, i.e. to add a term 'purgatory' to indicate subclasses that can be revived (a predecessor for obsoleting). Naming best practices demand that such helper classes should be marked as such, e.g. adding an underscore prefix: '_purgatory' [SSL+09]. This will prevent people from using it out of scope in annotating any data with it.

11. We find **non-orthogonality (redundancy) between the CV and its imports**: There are **terms with the same name** included, i.e. there is one MS:1000460 'unit' and a UO:0000000 'unit'. The MS term should be renamed 'mass spectrometry specific unit' and its definition be updated accordingly. Another example of such overlap between MS and PATO are 'linear', 'polarity', and 'wave length', each found in both ontologies. Such redundancy is bound to create confusion in users.

12. **No orthogonal scoping**: The CV contains terms from the MS domain but also from totally unspecific domains such as 'contact attribute'. Here FOAF⁶ or an equivalent ontology on administrative metadata could have been used in order to keep this CV orthogonal in coverage from already established CVs. 'chemical compounds' should probably come from ChEBI⁷ and

⁶ <http://xmlns.com/foaf/spec/>

⁷ <http://www.ebi.ac.uk/chebi/>

'file format', 'software' and 'external reference identifier' should come from IAO⁸ (Information Artefact Ontology).

13. There are **terms in the imports that have different definitions and names but share the same synonym**, e.g. we find unit:magnetic flux density unit and ms:magnetic field strength that both have the synonym 'B' **and are likely to refer to the same universal**. To ensure orthogonal (non-overlapping) CVs we propose that ms:magnetic field strength would be made obsolete or set equivalent (intersections in OBO) to the corresponding unit term. Also, we find a pato:pH and a unit:acidity with the synonym pH.

14. The **use of related-synonyms that are not exact synonyms** is ontologically doubtful practice, as the degree to which semantic equivalence is demanded is not specified.

15. **Masked class redundancy**. MS:gas (is_a sample state) is redundant to pato:gaseous conformation. This redundant modelling is repeated for liquid and liquid conformation, and the other matter states like solid.

16. **Missing basic classifications/is_a relations**: For some terms more parents could be specified, i.e. that a 'fragment neutral loss' is_a 'neutral loss' and that 'precursor neutral loss' is_a 'neutral loss'. It could be argued that 'raw data file' could be a subclass of 'file format'. The term 'neutral loss' means an input parameter for the MS instrument, whereas the term 'fragment neutral loss' says that a peptide was modified by loss of a small uncharged molecule like e.g. H₂O or NH₃.

17. Top level nodes like 'spectrum interpretation' are problematic, as an ontology should capture the realm of reality, and where epistemology is tackled, it should be made clear by using an information artefact class. **Epistemic intrusion**⁹ is also manifested in statements like 'predicted isoelectric point' is_a 'isoelectric point' and 'ambiguous residue'. The top level branch 'spectrum generation information' contains epistemic metadata ('information' is this case), which violates good ontological naming conventions [SSL+09], i.e. NC 1.3 (Avoid taboo words such as postfixes like name, type, class, information that refer to the representation level).

18. To add postfixes like 'name' to terms hints for the **use-mention problem**¹⁰: Term definitions like

```
name: cleavage agent name, def: "The name of the cleavage agent."  
[PSI:PI], is_a: MS:1001044 ! cleavage agent details
```

get problematic as soon as someone starts adding a 'has_site' cleavage site restriction, as it is not the name of anything that can cleave anything, but the actual individual of the class 'cleavage agent'. This should be reflected by the terms label, i.e. removing the "name" postfix that only creates confusions on the reality level (NC 1.3 Avoid taboo words).

19. We find 'chromatogram type' part_of 'chromatogram'. In a name like 'chromatogram type', the **'type'-postfix refers to representational metadata and is already implied** by the available subclasses and should not be indicated in the label. Both chromatogram subclasses 'electromagnetic radiation chromatogram' and 'mass chromatogram' should be made direct subclasses of 'chromatogram', rather than introducing an artificial "type" postfix.

⁸ <http://code.google.com/p/information-artifact-ontology/>

⁹ http://ontology.buffalo.edu/medo/Onto_Epist.pdf

¹⁰ http://en.wikipedia.org/wiki/Use%E2%80%93mention_distinction

20. The CV contains **negations in terms such as ‘unknown residue’, ‘unknown modification’ and ‘no threshold’**. In ontologies one would avoid such terms (NC, 2.4 Use positive names) as not-being-known is no intrinsic property of the universal. One would simply not annotate these or use the next more general superclass, e.g. ‘residue’.

21. **Use singular word forms in term labels:** Plural term forms should only be used where the value represents a plurality (NC 2.3 Prefer singular nominal form). For plural terms like ‘mass table options’ and ‘peptide modification details’ this means that all subclasses are pluralities themselves, which e.g. for the subsumption ‘nucleic acid base modification’ is_a ‘peptide modification details’ does not hold, as the modification refers to a singular. In reverse the CV contains cases where a term is formulated in a singular form, but the definition refers to a plurality, e.g. in the case of the term ‘alternate mass’ the definition reads: “List of masses a non-standard letter code is replaced with.”

22. The spectrum interpretation branch contains the statement ‘alternate mass’ is_a ‘ambiguous residues’. Not only violates the plural word-form established naming conventions [SSL+09], but the subsumption statement seems doubtful, given the slang-names used here. It seems the label represents a linguistic **ellipse, which should be avoided**, where the real head noun is missing to render the term shorter. The statement with correct explicit labels would be much easier readable, i.e. ‘residue of alternate mass’ is_a ‘ambiguous residue’.

23. The definitions sometimes contain **orthographic errors**, e.g. in the definition of ambiguous residues: “Children of this term describe ambiguous residues.” Another error here is that the definition is a) tautological, b) refers to the representation level (“term describes”) rather than to the reality level, and c) counts for subclasses rather than for the term itself.

24. The labels sometimes adhere to proprietary -but undocumented- complex naming conventions, i.e. where the label consists of two or more parts separated by a colon, e.g. ‘metabolic labelling: heavy N (mainly 15N)’, or ‘MaxQuant:peptide counts (unique)’. Such **compositions** should be used with care and where possible these **should be refactored into ontological patterns making use of relations**, i.e. by sourcing out the redundant prefix into an own super term. E.g. have the repetitive prefixes like ‘metabolic labelling’ and the Vendor ‘MaxQuant’ modelled as part of a pattern. Other such cases are the terms under ‘StudyVariable attribute’, namely the ones starting with ‘experimental condition’, e.g. ‘experimental condition ‘case’’, ‘experimental condition ‘disease’ and ‘experimental condition ‘control’.

25. **Violations of case and separator conventions:** The CV contains case convention violations, e.g. using capital case start in ‘Feature attribute’. An example for separator convention violations, i.e. using CamelCase word separation is ‘StudyVariable attribute’.

B. Ontological errors found in the MSI-NMR CV

1. **Sparse editorial and administrative metadata** exists for the MSI-NMR CV.

2. It currently consists of 125 classes, of which only half refer to NMR per se while many terms are generic. Domain independent terms such as ‘contact email’ should be factored out into more appropriate external ontologies (see PSI-MS section above) for the sake of **orthogonality between artefacts**.

3. Only 30% of the terms have **natural language definitions**

4. Definitions are often **non-Aristotelian, tautological or cyclic**.

5. To add a root node and **make all terms part_of this root term is a bad practice**, as this is metadata information, which is implied by the namespace and NS ID prefix anyway. It also presses the meaning of the part of relation into service. OboEdit's rule-based reasoner detects that e.g. 'contact attribute' part_of 'Metabolomics Standards Initiative NMR Spectrometry Vocabularies' is redundant and displays a warning. In the definition for contact attribute the assertion: is_a: NMR:1000547 ! object attribute, would be exploited by the reasoner to infer and assert the now user-asserted line

```
relationship: part_of NMR:0000000 ! Metabolomics Standards Initiative NMR Spectrometry Vocabularies
```

, as the rule assumes transitivity over the is_a and part_of hierarchies and states that if B (contact attribute) is_a A (object attribute) and A part_of X (NMR CV), then B (contact attribute) must be part_of X (NMR CV).

6. **Redundant terms:** The term pH is redundantly modelled in both NMR and UO namespaces. This should be avoided in order to maintain orthogonality.

7. Assertions such as

```
name: spectrum type, def: "Spectrum type." [MSI:NMR]
relationship: part_of NMR:1000442 ! spectrum
```

suffer from **use-mention problem** as they conflate ontology (real things) with epistemology (information on things). Its definition is tautological.

8. Capturing **epistemic postfixes** as in 'spectrum type' and 'gene name' is a bad practice and 'spectrum', 'gene' should be used instead. Also, here the 'part_of' relation seems pressed into service. Rather use '1D-Spectrum' is_a 'Spectrum'. Further errors of this type are illustrated in Fig. 1.

9. Given **transitivity of the part_of relation**, from 'spectrum type' part_of 'spectrum', 'nmr spectrum' is_a 'spectrum type', it would follow that an 'nmr spectrum' part_of 'spectrum', which sounds weird.

Another example of epistemological confusion can be found in the statement 'heavy labeled peptide' is_a 'peptide labeling state'. This should be: 'heavy labeled peptide' is_a 'labeled peptide'

10. The triple 'chemical compound attribute' part_of 'chemical compound' is misleading, as it would follow that e.g. a 'predicted isoelectric point (is_a 'chemical compound attribute') part_of 'chemical compound'. An **'is_about' relation with the range IAO: 'information artefact' should have been used here.**

11. Regarding the 'predicted'-affix, in general **only intrinsic properties of an entity should be captured in class names** shown in the taxonomic backbone. Extrinsic (externally asserted) properties need special handling, i.e. via roles.

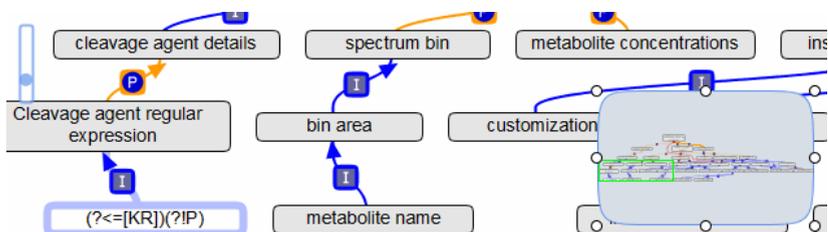


Figure 1: An excerpt of the NMR CV in the OboEdit ontology tree pane, showing some doubtful assertions. Is a certain regular expression really part_of some ‘detail’? Is a ‘metabolite name’ a ‘bin area’? Is a ‘metabolite name’ a ‘spectrum bin’?

Classification of found errors according to the OOPS! Ontology Pitfall catalogue

We here list the pitfalls of the OOPS! Pitfall Catalogue¹¹ which are occurring in the analysed CVs and map them to our own error scheme (Table 1).

Some of the OOPS pitfalls were not found in the analysed CVs, e.g. P1 No polysemy (Homolog to NC 2.1), P3 Creating ‘is’ relationships (This error seems rather ‘made up’ as we are not aware of occurrences of such errors in any ontology). P5 Wrong inverses, P6 Hierarchy cycles (unintended equivalences), P7 Different concepts in one class (Homolog to NC 2.2.), P31 Wrong equivalences, P34 Untyped classes and P35 Untyped properties. Non-applicable pitfalls (due to CV semantics) were P10-P19, P24-28, P30 and P33.

Table 1: The found CV modelling error types mapped to corresponding OOPS pitfalls. Non-occurring, not applicable and missing error types are excluded from the table.

OOPS Pitfall	Errors in CVs	Comment
P2 Creating equivalent classes	A11, A13, A15, B6	If formal semantics is used, reasoning can do this.
P4 Creating unconnected elements	A1, A5, A7	Not using Quality class. This can be tested with OntoCheck [STS+12], or a ‘usage’ test in P4.
P8 Missing annotations	A9, B1, B3	Lack of metadata can be checked by OntoCheck and the OboEdit verification manager.
P9 Missing basic information	A1, A5, A7, A8, A16, B3, B10	Missing statements and restrictions are only a problem, if the application cannot fulfil its use case.
P20 Misused annotation properties	A2, B1	
P21 Using a ‘miscellaneous’ class	A10, A20	Homolog to Naming Convention 2.5
P22 Naming violations	A6, A8, A9, A10, A14, A17, A18, A20,	The OBO Foundry Naming Conventions are much

¹¹ <http://oeg-lia3.dia.fi.upm.es/oops/catalogue.jsp>

	A21, A22, A24, A25, B7, B8, B11	more detailed here.
P23 Misuse of ontology elements	A1, A2, A3, A4, A10, A17, A18, A19, B5, B7, B8, B10	Not exploiting the available formats' semantics.
P29 Wrong transitivity on relations	B9	
P32 Classes with same label. Synonyms as equivalent classes.	A11, B6	Label redundancy check in OntoCheck and OboEdit verification manager.

For the following errors a classification according to OOPS was attempted, but was rather shallow due to lack of OOPS error granularity (see below). For A14 (The use of related-synonyms that are not exact synonyms), one could force P1 into action. For A17, B5 (Adding a root node and make all terms part_of this root term) and B8 (Epistemic intrusion) P23 could be pressed into service. B9 (Transitivity consequences) could be mapped onto P29 (Wrong equivalences on relations).

Discussion

Potential reasons for the errors in the CVs

Although the PSI-MS CV was developed according to quite explicit design principles¹², our analysis revealed some ontological errors. For the PSI-MS CV there can be multiple potential reasons for the prevalence of the identified errors. The PSI/MSI CVs have to serve a wide range of application use cases from long term data storage, database querying to MIBBI standard [TFS+08] verification and journal requirements enforcement [RSU+09]. This results in the CV to be aligned to the least semantic formal use case. The discussion if the analyzed CVs should be called ontologies is old and not fruitful, as the usage scope of these CVs is so different from formal ontologies (as e.g. represented in DL), that modelling principles can hardly be compared¹³. As mentioned in [SSL+09] it is however still justified to apply at least a subset of the DL best practices, i.e. naming conventions to the CVs. The particular improvement for the artefacts under scrutiny here is that the validation rules needed to verify correct CV usage in data annotation can be simpler.

However the way the XML/CV combination is exploited could be achieved with more ease applying e.g. Protégé Frames format, which was particularly developed for data entry guidance and verification and has elaborated data acquisition functionalities. We can only speculate that the decision for XML was taken based on its more widespread

¹² <http://www.psidev.info/node/47>

¹³ From this standpoint a lot of older work had been criticized harshly with doubtful justification. The advent of semantic web and Linked Open Data techniques made even hardliners with a formal DL background give up some of their strong positions.

use. An additional reason was to have a stable open standard, whereas the Frames format seems to be more proprietary, even if the Protégé tool is open source.

In general the CVs are modelled more like the XSDs. Many ontological errors might be the result of the CV maintainers having an XML background rather than an ontology background and hence a) being used to cluster data rather than real world instances and b) being not aware of the best practices in the ontology realm. Further, historically motivated changes in CV scope and modularizations lead to a development with multiple concept changes and merges between previously separated PSI CVs. This is exemplified in the PSI-MS ‘spectrum generation’ branch containing mostly *part_of* relations¹⁴ in its upper area and displaying a partonomic structure rather than a taxonomy. Its sibling, the ‘spectrum interpretation’ branch largely represents a taxonomic structure. At first, we assumed this heterogeneity is due to the different usage scopes, i.e. the partonomy used in resemblance to an XSD to specify spectrum generation parameters in *mzML*, whereas the taxonomy is used for interpretations expressed in *mzQuantML* and *mzIdentML* [JEM+12], but this rather had historical reasons, because the PSI-MS CV came into being by merging two predecessor ontologies¹⁵. In the OBO format there is the possibility to apply so-called ‘subsets’, e.g. to indicate terms that are used for spectrum generation or spectrum interpretation purposes.

Some errors are a result of the PSI-MS CV maintenance being a decentralized effort with difficult consensus of community discussions. Further the personnel of the CV coordinator changes over time which makes stringent policy fulfilment difficult.

In [MMO+13] the authors mention that the CV is used to a) avoid inconsistencies in annotation, b) to have a unique (and preferably short) accession number and to c) give researchers and computer algorithms the possibility for more expressive semantic annotation of data. The authors claim further d) that ‘The CV contains a logical hierarchical structure to ensure ease of maintenance and the development of software that makes use of complex semantics’. Although b) is clearly supported, for a) it can be argued that the validators rather exploit the rules, but less the semantics inherent in the OBO CVs. Regarding claim d) the structure is not completely logic as its not rooted in DL set theoretic semantics and suffers from the above mentioned errors.

No intersections are used in the CVs. That means no equivalence statements and hence detection can be automatized in a reasoner. As the key objective of *mzML*/CV set-up according to [MCS+11] is: (i) creation of a simple format, and (ii) elimination of alternate ways to encode the same information, it could be argued that applying OWL-DL and its capability to automatically detect semantic equivalences would be overshoot for these CVs. On the other hand, as a consequence of this, all needed terms have to be pre-coordinated¹⁶.

¹⁴Most of the *part_of* relations here are in fact *is_about* relations, e.g. in the way an *ms:chemical compound* formula *is_about* a chemical compound rather than *part_of* it.

¹⁵The terms below *MS:1001000* stem from an *mzML* specific ontology, which was merged with the interpretation terms (id above *MS:1001000*) for *mzIdentML* and *mzQuantML*.

¹⁶ http://www.debugit.eu/events/documents/Schober%20PrePostCoord_Lissabon.pdf,
<http://ontogenesis.knowledgeblog.org/1305>

Regarding the prevalence of error A12 No orthogonal scoping, although it would be a good practice to have clearly distinct artefact scope borders and orthogonal domains modelled in separate namespaces, we see how this not easy to accomplish in a pragmatic setup, as this would enforce term inclusion requests to external editors, which might take some time to be implemented, or not achievable at all, as many artefacts currently available in the ontology portals are not updated or maintained any longer. The fact that the CVs often contain terms that are not domain-specific might be due to the difficulties in importing single terms from other ontologies. For OWL a per-term import mechanism exists¹⁷, but for OBO such function is currently missing and importing whole artefacts to leverage on just a few terms makes the overall artefacts unhandy, also leading to the observed unused imported entities and questionable use of qualities (error A1). As mentioned, we feel the use of ‘part of’ seems rather pressed into service here, as the way an instrument attribute is ‘part of’ an instrument is not the same ‘part of’ as in the statement ‘detector part_of instrument’¹⁸. These differences in modelling principles may result from a dissonance caused by heterogeneous expressivities of the MS CV and the imported ontologies. If the import is build according to a more formal DL philosophy, the CV of less formal semantics importing it, is likely to miss out on certain idiom usage. From this we can derive the general demand that an ontology should only import artefacts that have a comparable expressivity. This will impact the NMR CV development directly, as it is planned to re-use and submit terms to OBI¹⁹. Importing such complex DL ontology into the flat NMR CV is bound to cause incompatibility problems through dissonant modelling patterns.

For many errors, tools to detect and alleviate them exist. I.e. for A11 non-orthogonality (redundancy), tools like OntoCheck for Protégé [STS+12] and the OboEdit verification manager/name check can be used elegantly to avoid these. Another error easy to be prevented via these tools is A13 Terms have different definitions and names but share the same synonym. These cases hint for masked redundancy (heteronyms) and should be cleaned up. OntoCheck and the OboEdit verification manager could also detect and prevent errors like A20 Negations in terms. This practice might be justified for CVs, as people are interested in these as simple search attributes. A23 orthographic errors could be detected by the OboEdit text verification check, which is based on a lexicon.

Discussions on the error type mapping to the OOPS classification

Although most of our identified error types could be mapped towards pitfalls in the OOPS, during our analysis we identified further error types which could not be classified according to the current OOPS pitfall list. These five potentially new pitfalls have been submitted to the OOPS catalogue maintainers for inclusion: Specific terms at upper level positions (A7). Tautological definitions (A9, B4), Non-orthogonal scoping (A12, B2), Orthographic errors (A23). Import of artefacts of heterogeneous semantics and scope (A1).

¹⁷ <http://obi-ontology.org/page/MIREOT>

¹⁸ Page 1, case 3 in http://www.columbia.edu/~av72/papers/AO_2006.pdf

¹⁹ http://obi-ontology.org/page/Main_Page

The OOPS pitfalls are not always orthogonal, as some errors could be classified into multiple pitfalls, e.g. B3 Missing definitions, is now classified under P8 Missing annotations and P9 Missing basic information. To guide the evaluator, orthogonality between the pitfalls should be the goal. In other cases a multiplicity of errors could be classified under one and the same pitfall, overloading it with meaning and decreasing error resolution (see Figure 2). Here the pitfalls should be subclassed to allow for a more concise error classification, e.g. the above mentioned P8 should be made a sub-pitfall of P9, hence generating a pitfall hierarchy for easy navigation. At the moment, the catalogue lacks structure as it is presented as a rather flat list of pitfalls that are not sorted i.e. according to expressivity regime or ontology element. Some OOPS pitfalls are named unintuitive, e.g. in P24 Using recursive definition it should be made clear that axiomatic class definitions are meant here, rather than natural language definitions.

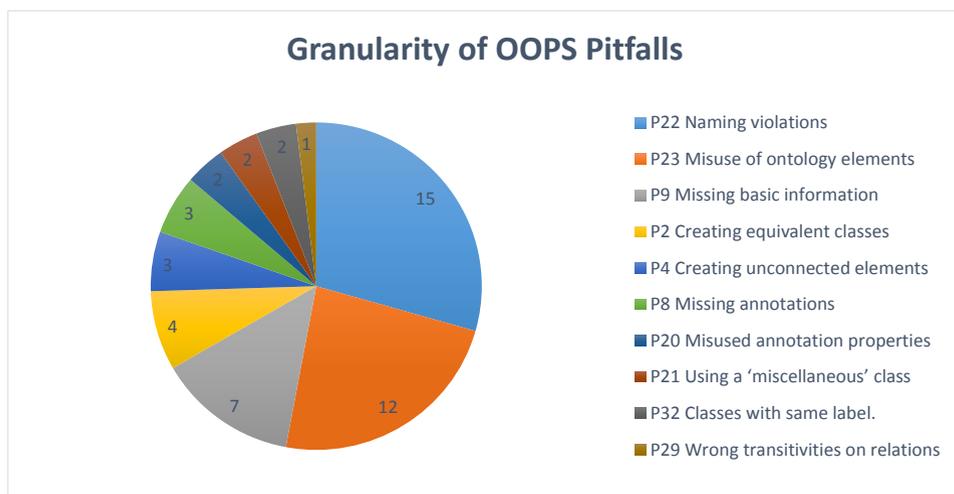


Figure 2: Occurrence of found error types mapped to OOPS pitfalls, indicating the necessity for a more granular pitfall description for the OOPS pitfalls P 22, 23 and P9. These were found to be ‘overloaded’ with each having more than seven error types mapped.

Consequences for the COSMOS NMR CV development

As the COSMOS work package 2 is responsible for the standards development and coordination, it needs to take care not to copy unjustified design decisions from the PSI CV. The NMR CV strives to achieve a compromise in adhering to ontology best practices while still being easy to use and fulfilling its primary use case. The NMR CV will be released with a clear documentation on all its design decisions.

Looking at the distribution of the found errors over error categories and best practice recommendations, it becomes evident that, although most policy providers tackle DL semantics, many patterns are still applicable to the CV domain. However not one single scheme was comprehensive enough to capture all found errors and hence for quality

assurance in the COSMOS NMR CV development still a multitude of patterns from a multitude of providers have to be applied.

Conclusion

Problems arise when ontologies build according to different design principles are to be used in combination, i.e. Unit or PATO imported into the NMR CV or if the OBO NMR CV is to be used with the top level ontology BFO²⁰ in DL. Even different knowledge representation languages (syntaxes like OWL vs. OBO can lead to problems, i.e. due to different quantifier handling²¹ [BTH+11]). As long as these cannot be completely mapped, this dichotomy in bioontologies will prevail, and if they could be mapped then there is no need for the OBO format in the first place.

As some design principles valid for DL ontologies are not that relevant for semantically flat CVs we propose that policy providers such as the OBO Foundry should more clearly specify the scope of their policy set. We propose that best practices are rather issued on a per-expressivity basis, a way that has been taken by the ontology design pattern community to narrow down the scope of their policies to specific expressivity regimes.

We have analyzed the CVs currently used in PSI and MSI efforts from an ontologist's perspective and classified the more prevalent modelling errors according to two ontology evaluation frameworks. This endeavor has contributed to increase the quality of the analyzed CVs and will be valuable in the design phase of the new Cosmos NMR CV. Besides an improvement in the mentioned artefacts, the propagation of new error types to the OOPS together with constructive critique should contribute to improve future OOPS versions and evaluation scenarios.

Acknowledgements

We like to thank the COSMOS and PSI working groups. DS was funded through the EU FP7 project COSMOS grant EC312941 (<http://www.cosmos-fp7.eu>). GM is funded by 'ProteomeXchange' (<http://www.proteomexchange.org>, EU FP7 grant number 260558). ME is funded by the Protein Unit for Research in Europe (<http://www.pure.rub.de>).

References

- [BTH+11] Boeker M, Tudose I, Hastings J, Schober D, Schulz S (2011) Unintended consequences of existential quantifications in biomedical ontologies. *BMC Bioinformatics* 12: 456

²⁰ <http://code.google.com/p/bfo/>

²¹ Slide 15 on <http://de.slideshare.net/dosumis/from-obo-to-owl-and-back-building-scalable-ontologies>

- [FGD+11] Ferry-Dumazet et al. (2011) MeRy-B: a web knowledgebase for the storage, visualization, analysis and annotation of plant NMR metabolomic profiles, *BMC Plant Biology* 2011, 11:104, <http://www.biomedcentral.com/1471-2229/11/104>
- [HSC+12] Haug K, Salek RM, Conesa P, Hastings J, de Matos P (2012) MetaboLights—an open-access general-purpose repository for metabolomics studies and associated meta-data, *Nucleic Acids Research*, 2012, 1–6 doi:10.1093/nar/gks1004
- [JEM+12] Jones AR, Eisenacher M, Mayer G, Kohlbacher O et al. (2012) The mzIdentML data standard for mass spectrometry-based proteomics results. *Mol Cell Proteomics*. 2012 Jul;11(7).
- [KCB+08] Kessner, D., Chambers, M., Burke, R., Agus, D., and Mallick, P. (2008) ProteoWizard: open source software for rapid proteomics tools development. *Bioinformatics* 24, 2534–2536
- [MCS+11] Martens L, Chambers M, Sturm M. et al. (2011) mzML—a community standard for mass spectrometry data. *Mol. Cell Proteomics*, 10, R110000133. <http://www.ncbi.nlm.nih.gov/pubmed/20716697>
- [MKR+09] Montecchi-Palazzi L., Kerrien S., Reisinger F. et al. (2009) The PSI semantic validator: a framework to check MIAPE compliance of proteomics data. *Proteomics*, 9, 5112–5119.
- [MMO+13] Mayer G, Montecchi-Palazzi L, Ovelleiro D, Jones AR, Binz PA, Deutsch EW, et al.; (2013) The HUPO proteomics standards initiative- mass spectrometry controlled vocabulary. *Database (Oxford)*. 2013 Mar 12;2013(0):bat009
- [RSU+09] Rodriguez H., Snyder M., Uhlen M. et al. (2009) Recommendations from the 2008 International Summit on Proteomics Data Release and Sharing Policy: the Amsterdam principles. *J. Proteome Res.*, 8, 3689–3692.
- [SFG+07] Sansone S.A., Fan T., Goodacre R. et al. (2007) The metabolomics standards initiative. *Nat. Biotechnol.*, 25, 846–848.
- [SSL+09] Schober D, Smith B, Lewis L et al. (2009) Survey-based naming conventions for use in OBO Foundry ontology development. *BMC Bioinformatics*, Vol.10, Issue 1, 2009.
- [STS+12] Schober D, Tudose I, Svatek V, Boeker M (2012) OntoCheck: verifying ontology naming conventions and metadata completeness in Protégé 4, *Journal of Biomedical Semantics* 2012, 3(Suppl 2):S4, <http://www.jbiomedsem.com/content/3/S2/S4>
- [TFS+08] Taylor, C. F., Field, D., Sansone, S. A., Aerts, J., Apweiler, R., Ashburner, M et al. (2008) Promoting coherent minimum reporting guidelines for biological and biomedical investigations: the MIBBI project. *Nat Biotechnol* 26, 889–896.
- [VCC+13] Vizcaíno J.A., Côté R.G., Csordas A. et al. (2013) The Proteomics Identifications (PRIDE) database and associated tools: status in 2013. *Nucleic Acids Research* 2013,D1063-D1069, doi:10.1093/nar/gks1262