Towards standardized evidence descriptors for metabolite annotations

Daniel Schober [1,*], Reza M Salek[2] and Steffen Neumann [1]

[1] Leibniz Institute of Plant Biochemistry, Dept. of Stress and Developmental Biology, Weinberg 3, 06120 Halle, Germany
[2] European Bioinformatics Institute (EMBL-EBI), European Molecular Biology Laboratory, Wellcome Trust Genome Campus, Hinxton, Cambridge, CB10 1SD, UK

ABSTRACT

Motivation: Data on measured abundances of small molecules from biomaterial is currently accumulating in the literature and in online repositories. Unless formal machine-readable evidence assertions for such metabolite identifications are provided, quality assessment based re-use will be sparse. Existing annotation schemes are not universally adopted, nor granular enough to be of practical use in evidence-based quality assessment.

Results: We review existing evidence schemes for metabolite identifications of variant semantic expressivity and derive requirements for a 'compliance-optimized' yet traceable annotation model. We present a pattern-based, yet simple taxonomy of intuitive and self-explaining descriptors that allow to annotate metabolomics assay results both in literature and data bases with evidence information on small molecule analytics gained via technologies such as mass spectrometry or NMR. We present example annotations for typical mass spectrometry molecule assignments and outline next steps for integration with existing ontologies and metabolomics data exchange formats.

Availability: An initial draft and documentation of the metabolite identification evidence code ontology is available at https://github.com/DSchober/MIECO. Supplementary material can be found at goo.gl/NCsA7w

*** Contact:** dschober@ipb-halle.de

1 INTRODUCTION

1.1 Background

Metabolomics investigates the distribution and abundance of small molecules in organisms, mainly applying assay methods like Gas Chromatography/ Liquid Chromatography (GC/LC), Mass Spectrometry (MS), Nuclear Magnetic Resonance Spectroscopy (NMR), Ultraviolet (UV) and Infra Red (IR) spectroscopy. To convert this analytical data into usable systemic knowledge, identification and annotation of metabolites in biomaterials is essential (Creek et al. 2014), e.g. to indicate that study X provides evidence by assay Y for occurrence of metabolite Z (in sample Q under condition R).

However, the degrees of confidence in identification statements can vary greatly between researchers and studies and are difficult to communicate among users in a crisp, yet concise and accurate fashion (Schymanski et al. 2014). An author's method for reporting the identification evidence in free text may be dependent on the context and is usually hard to follow by an external recipient, be it another scientist or a computational agent like a search engine. Attempts were made to set up traceable annotation systems to indicate the quality levels of evidence assignments. Different proposals were put forward to allow biologists to indicate their identifications with evidence information in a standardized manner, ranging from domain-specific simple four level schemes (Sumner et al. 2007) to complex domain-independent description logic (DL) based ontologies for automatic evidence reasoning (Bölling et al. 2014). Yet, none of these efforts has gained greater momentum so far.

The PhenoMeNal data standards workpackage (PhenoMeNal website 2016) and the Metabolite Identification Task Group of the Metabolomics Society (Creek et al. 2014) both aim to foster the development and harmonization of metabolite evidence reporting. As part of this endeavor, we briefly review existing schemes, identify their compliance problems and present a domain specific, simple and compliance-maximized ontology to assist metabolite evidence assignments.

1.2 Overview on existing evidence schemes

Among the minimum reporting standards put forth by the Metabolomics Standards Initiative (Fiehn et al. 2007), a four level evidence scheme was proposed to enable researchers to specify the degree of confidence in metabolite annotations (Sumner et al. 2007):

- **Level 1: Confident Identification** based on two orthogonal evidences using defined reference standards measured under identical analytical conditions.
- **Level 2: Putative Identification** based on similar physicochemical properties or library spectra similarities (no authentic reference standard).
- **Level 3**: **Putative Identification of Compound-Class** i.e. classification based on similar physicochemical properties or spectral similarity with a compound class.
- **Level 4: Known Unknowns** that are unidentified, yet can be differentiated and quantified based on spectral data.

This broad numeric classification is proposed for wider usage, e.g. as recommended assay annotation for the metabolomics journal (Metabolomics Journal, 2016). Although easy to use, its drawback is a lack of granularity in the detail of what evidences can be expressed in a formal and search engine-friendly manner. This lack of utility might be the reason why it has been sparsely adopted by the metabolomics community (Everett 2016). Realizing these drawbacks, and based on an earlier suggestion of adding computable numeric indicators (Creek et al. 2014), Sumner et al (2014) proposed a more granular assay-technology centered scheme, allowing to assert numeric weights to its granular evidence components, resulting in an additive quantitative identification score. At the same time, the earlier four level scheme was expanded by Schymanski et al. (2014), providing an enriched scheme with five identification levels that are accompanied by minimal assay data requirements. Level 1 to 3 map to the Sumner levels, but provide a better granularity, with 2a and

2b distinguishing different sets of evidence in absence of a reference standard, whereas level 4 is an addition describing formula based annotations and Level 5 describes the known unknowns via an exact mass.
Besides the aforementioned domain-specific approaches, advanced ontological proposals of more general, domain-independent nature have been introduced. These leverage on automated evidence reasoning, based on description logic (DL) semantics, with axiomatizations by means of fine-grained DL patterns: The Evidence Code Ontology (ECO), (Chibucos et al. 2014) applies a basic pattern around the notions 'Assertion method' and 'Evidence'. At a recent workshop (ECO-OBI Workshop 2016) it became evident that besides domain-specific coverage gaps, restructuring is required to allow for OBI (Bandrowski et al. 2016) usage. The transition from a simple enumerated evidence list towards axiomatised reasoning for auto-classification is reflected in the recent name change: 'Evidence and Conclusion Ontology' (ECO website 2016). Diverging from its initial Gene Ontology inspired simple set-up, this effort now competes with complex DL-heavy approaches like the Semantic Evidence (SEE) reasoning ontology (Bölling et al. 2014) and there is danger of development slow-down due to the added complexity. Coverage is currently sparse on metabolomics technologies, i.e. important top level terms like

```
'mass spectrometry evidence used in automatic as-
sertion' EquivalentTo:
    'mass spectrometry evidence' AND
      (used_in SOME 'automatic assertion')
```

are missing. An analysis of other existing domain ontologies revealed sparse coverage for terms used in metabolite identification and distribution across multiple namespaces (see Supplementary Material), making it necessary to build a coherent artefact[1]. As ECO still is expected to gain a greater user base, we decided to re-use this ontology by importing and expanding it and leveraging on its basic evidence pattern.

2 METHODS

To allow for granular coverage, we apply an ontology driven approach to generate descriptors for small molecule evidence assignments, covering all major assay methods in metabolomics. We imported ECO in Protégé 4.3 and started our additions in an OWL ontology called Metabolite Identification Evidence Code Ontology (MIECO). OBI-core[2] was imported to gain BFO 2, the OBI upper level and core RO object properties. These are utilized to increase the granularity of assertion modes, *i.e.* providing ontology design patterns for composite assertion specifications. As we also aim for backwards compatibility to the Sumner (2007) scheme, we intend to employ DL reasoning to automatically map and classify MIECO evidences onto earlier evidence schemes.

In a first round of ontology design, we derived terms in a data driven bottom-up approach from an in-house LC-MS use case (UPLC/ESI-QTOF MS) based on an untargeted analysis of semipolar root exudate metabolites[3] of *Arabidopsis thaliana* (Strehmel et al. 2014). Based on the natural language identification statements in this paper, we generated an enriched list of around 20 evidence descriptors. (see Supplemental material). Additional input were the alphanumeric terms from Tab.2 in Sumner et al. (2014), which lists the most common descriptors proposed as identification indicators, distributed among the five most prominent assay methods. From that, we started expansion by first identifying an intuitive lexical pattern that is easily understandable by our end users and that covers most of the classes generated from our initial paper text. This will form the basis for later reconfiguration into DL patterns and automatic pattern-based term creation via TermGenie (Dietze et al. 2014), which assists pragmatic pre-coordination of only those terms really required in practice. We exemplarily annotate assignments from Table 2 in the LCMS use case paper (Strehmel et al. 2014) over a range of evidence levels.

3 RESULTS

Our analysis indicates that the simple four level scheme in use today is not granular enough to provide reliable quality indicators and foster quality and trust analysis for metabolite annotation. Formal DL approaches, on the other hand, tend to develop slowly and shielding their complexity from users has been a known problem, preventing end user compliance and community growth. Here we strive for a pragmatic middle way, starting with an intuitive taxonomy of data driven and frequently used descriptors, and later exploiting DLs combinatorial semantics to generate axioms and additional evidence terms from patterns where required via TermGenie, as described above.

We provide a first draft of a taxonomy of pre-coordinated descriptors for metabolite identification. In contrast to emerging DL-heavy approaches, our scheme is less complex, as we focus on simple and easy to use terms for maximum end user compliance. MIECO currently consists of ~ 90 terms, axiomatization being sparse at this early stage.

3.1 Pattern proposals

Metabolite identification refers to assertions that support that a compound under investigation is either of a certain totally defined molecular structure (Identification on the leaf node universal level) or is of a certain backbone structure (annotation on the superclass level). We therefore structure the main evidence axis according to the following assertion taxonomy and compositional pattern:

Assertion e.g. comprising a taxonomy of Annotation/Characterisation/Classification/Identification

of

Molecular structural element e.g. Molecule, class, functional group, element, Isotope

by (

Assay Outcome e.g. Assay outcome with sub-assay details e.g. MS, MS2, LC/RT, Isotope data, adduct data, precursor (quantifier) ion type

used in

Assertion method e.g. Run against reference standard, Comparison to reference database, Author inference)

[1] Term imports and cross-referencing methodologies like MIREOT in OntoFox are still fragile and often provide little help in modularisation and domain border decisions.

[2] http://obi-ontology.org/page/Core

[3] The data is available as MTBLS160 in Metabolights at http://www.ebi.ac.uk/metabolights/reviewerLgTnoHUrFb

We currently try to render this linguistic pattern compatible with ECO, but we hope the patterns will still be intuitive to understand by peer users. The part in round brackets is of 1:n cardinality.

We strive for an end-user compliant naming pattern for future MIECO class names. MIECO labels are currently generated in a bottom up manner as extracted from the use case. The relation of the naming pattern and the MIECO term labels will be, that in a future release we can add the formally consistent pattern-generated TermGenie labels as alternative labels to existing user-preferred (Schober et al. 2009) MIECO labels. The difference to the above lexical pattern is that the Assertion taxonomy has been transformed into an assertion relation (object property) hierarchy:

asserted by (annotated by)
characterised by
classified by
identified by

Overall evidence naming pattern:

MolecularStructureElement [*annotation relation*] AssayOutcome *used_in* AssertionMethod

Example Evidence Class:

Guanosine *identified_by* 'LCMS fragmentation pattern' *used_in* 'similarity to authentic reference standard'

3.2 Example annotations and mappings

Table 1 exemplarily shows evidence annotations via MIECO terms for different feature assignments from the use case paper (Strehmel et al. 2014, modified after supplementary Table 2 therein) and compares these to earlier evidence schemes.

Aside these encompassing overall assignment annotations, MIECO terms can also be used to annotate on a highly granular level, i.e. annotating each evidence contributor for a whole set of experimental LCMS characteristics of a molecule. E.g. for feature #2 (Guanosine), we can assign

- 'MIECO_0000094: Characterisation by sum formula', annotating the Elem. Comp. of $C_{10}H_{13}N_5O_5$.
- 'MIECO_0000005: Characterisation by LC RT similarity', annotating the Retention time 46s
- 'MIECO_0000012: Characterisation by online HD exchange experiment identified substructure revealing exchangeable protons', annotating the evidence of the 6 exchange protons matching the H-functional groups.
- 'MIECO_0000016: Characterisation by collision induced dissociation (CID) MS2 with mass and isotope pattern of quasi-molecular fragment ion in negative ESI mode', annotating the [M-H]$^-$ Quantifier Ion.
- 'MIECO_0000009: Characterisation by m/z value in MS1', annotating the 282.08 m/z of the parent ion.
- 'MIECO_0000010: Characterisation by fragmentation pattern in MS2', annotating the MS^2 fragment ions m/z 150,133.

Table 1. A tabular representation of a subset of heterogeneously annotated compounds from the use case paper (Strehmel et al. 2014) is shown, with additional MIECO term annotation and annotation with earlier evidence schemes (both shaded gray). The last six rows represent single evidence contributors (EC)

LCMS Feature#|	**2**	**20**	**50**	**100**
Assignment	Guanosine	H-Val-Leu-OH	Unknown Indole derivate	Unknown
MIECO Annotation	MIECO_0000001:Complete structural identification by LCMS similarity to authentic reference standard	MIECO_0000001, MIECO_0000002: Characterisation by LCMS similarity to literature reference	MIECO_0000097:Classification based on RT and m/z value in MS2	MIECO_0000098:Unknown assignment based on RT and m/z value in MS2
Verification Level, VL	S	S,L	I	-
Sumner 2007	Level1	Level1	Level3	Level4
Elem.Comp.	$C_{10}H_{13}N_5O_5$	$C_{11}H_{22}N_2O_3$	$C_{10}H_9NO_3$	$C_{20}H_{28}O_{11}$
RT[s]	46	159	391	234
#Exchange Protons	6	4	n/a	6
Precursor. Ion Type	[M-H]$^-$	[M+H]$^+$	[M+H]$^+$	[M-H]$^-$
m/z	282.08	231.17	192.07	443.15
MS^2 fragments	150,133	213,185,132,86,72	177, 174, 161, 159, 148, 133, 132, 117,116, 105, 104	291, 151, 145, 125, 107, 101

Table derived from Supplementary Tab 2 in Strehmel et al. (2014). All acronyms and abbreviations are explained there. RT=Retention time;VL is an inhouse verification level applied in the paper: S=Secure, as RT and m/z match Ref Standard; I=Interpreted, based on CID MS and online H/D exchange; L=Literature reference; -=no assignment possible but distinct m/z pattern (known unknown).

4 DISCUSSION

Reliable metabolite identification is not easy to achieve and communicate in a traceable, reproducible manner. Among the reasons is that identification assertions made by humans often embrace implicitness, e.q. as a consequence of the complexity of the underlying state of affairs. The reason for this is the combinatorial cross dependencies of the steps in a metabolite identification process. This can consist of multiple singular assertions that act together in a non-linear synergistic fashion to ultimately produce an evidence. This is why highly granular models are required and ontologies are currently the best formalisms to capture such detail. If the granular subprocesses of an identification process are not explicitly named, the danger of subjectiveness arises and unreliable identification measures can decrease not only scientific resolution and interpretation, but may also blur further processing, re-use and knowledge generation. However, the reliability is often not additive/linear, but rather an emerging property of inference chains. Multiple *per se* anecdotal evidences can result in a strong evidence, because the parts are reinforcing each other. Emergence is unfortunately not easy to be captured in Boolean symbolic formalisms like ontologies (Haken H., Schober D. 2008), but as a proxy for such knowledge could be introduced by assigning weights to single MIECO evidences and then let a rule-based system judge/compute on the overall evidence of the combined evidence contributors.

Also, as molecular structures are an important part of our annotation pattern, we have to consider at what level in a generalization/specialization taxonomy a distinction between class and individual is made[4].

As next steps we need to increase the amount of pre-coordinated terms beyond use case coverage and we plan to foster interaction and cross play with ECO. The addition of weighted numeric information as proposed in Sumner et al. (2014) will later lead to a quantified scoring scheme. We currently investigate usage within community resources like Metabolights (Haug et al. 2013) and integration into Galaxy Workflows. We currently test curator compliance by annotating experiment entries within Metabolights[5] with MIECO terms, i.e. test if users apply MIECO correctly. Suitability for use in existing metadata representation systems like ISA (Sansone et al. 2012) and format exchange data standards like mzTab or mzIdentML (Jones et al. 2012) are analysed. E.g. mzTab expanded the Sumner et al. (2007) scheme with added identification reliabilities: Small molecule identifications reported in an mzTab file can be assigned a reliability, reported as an integer between 1-3 for proteomics results (1: high reliability, 2: medium reliability, 3: poor reliability). This is easy to use but still untraceable insofar as these indicators cannot be derived from provided granular metadata.

As a next step we need to investigate curator tools for assisted semi-manual annotation, as well as the best storage places for metabolite evidence metadata. We will also look into text mining and embedding into annotation tools for computer assisted high throughput annotation, rendering an ever increasing amount of data accessible to evidence-based threshold filtering for quality data re-use. Here the transition to a quantitative background model for numeric evaluation is necessary.

5 CONCLUSION

Our initial MIECO draft is domain-specific use case (bottom up) driven vocabulary for annotating metabolite assignments with assay specific evidence terms. It was designed in a pragmatic manner, attempting to shield the end-users from DL axiomatizations and subscribing to complexity reduction strategies described earlier (Schober et al. 2014). Although in an early stage, we hope this domain restriction and its simple design, reflecting a lexical design pattern that is intuitive to a domain specialist, will contribute to reducing development times in the future. Our ontology-based metabolite identification and evidence scheme could become a handy asset in judging data provenance and reliability of identification assertions, i.e. allowing to set confidence thresholds for search and retrieval tasks. When more mature, this annotation scheme could gain momentum in the larger metabolomics community and is envisioned to contribute to a more traceable catalog of descriptors for small molecule assignment evidence.

ACKNOWLEDGEMENT

We like to thank the Baltimore ECO workshop participants, Nadine Strehmel, Christoph Ruttkies and the Metabolite Identification task group of the Metabolomics Society for inspiration and feedback. This work has been financed by the EC Horizon 2020 project PhenoMeNal, grant agreement number 654241.

REFERENCES

Bandrowski A., Brinkman R., Brochhausen M., Brush M.H., Bug B., Chibucos M.C., et al. (2016) The Ontology for Biomedical Investigations. *PLoS ONE* 11(**4**): e0154556. doi:10.1371/journal.pone.0154556

Bölling C., Weidlich M., Holzhütter H.G. (2014), SEE: structured representation of scientific evidence in the biomedical domain using Semantic Web techniques. *J Biomed Semantics*; **5**(Suppl 1): S1. doi: 10.1186/2041-1480-5-S1-S1 http://www.jbiomedsem.com/content/5/S1/S1

Chibucos M.C., Mungall C.J., Balakrishnan R., Christie K.R., Huntley R.P., White O., Blake J.A., Lewis S.E., Giglio M., (2014), Standardized description of scientific evidence using the Evidence Ontology (ECO). *Database*. 2014, 2014: bau075-10.1093/database/bau075, http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4105709

Creek, D., Dunn, W., Fiehn, O., Griffin, J., Hall, R., Lei, Z., Mistrik, R., Neumann, S., Schymanski, E. L., Sumner, L., et al. (2014), Metabolite identification: are you sure? And how do your peers gauge your confidence? *Metabolomics*, **10**, pp. 350–353

Dietze H. et al. (2014), Termgenie–a web-application for pattern-based ontology class generation. *J. Biomed. Semant*., **5**, 48, http://bioinformatics.oxfordjournals.org/external-ref?access_num=10.1186/2041-1480-5-48&link_type=DOI

Dunn WB, Erban A, Weber RJM, Creek DJ, Brown M, Breitling R, Hankemeier T, Goodacre R, Neumann S, Kopka J. et al. (2013), Mass appeal: metabolite identification in mass spectrometry-focused untargeted metabolomics. *Metabolomics*. 2013;**9**:p44–66.

ECO-OBI workshop (2016); https://docs.google.com/document/d/1Y-gxKHpHAyS_ngiAS1NkBLT4lttPAO-x8ECHkCf9kSw/edit#heading=h.r8a1r3t5hcrz, accessed 21.6.2016

ECO website (2016) http://www.evidenceontology.org/; accessed 3.9.2016

Everett, J., (2016), New NMR-Based Methods for Assessing Confidence in Known Metabolite Identification, Metabolomics Spotlight Article. In MetaboNews website May 2016, http://www.metabonews.ca/May2016/MetabоNews_May2016.htm#spotlight , last accessed 17.6.2016

Fiehn O., Robertson D., Griffin J., van der Nikolau B, et al. (2007), The metabolomics standards initiative (MSI) *Metabolomics*. 2007;**3**:175–178. doi: 10.1007/s11306-007-0070-6.

[4] i.e. is a metabolite fully identified as Leucine? Or is there a need to further disambiguate the analyte from Isoleucine ? Are trans fatty acids the same as cis fatty acids? ChEBI does not always separate class from instance level.

[5] Metabolights is a dedicated open data repository for metabolomics experiments and metabolite assignment data.

Haken H., Schober D. (2008), personal email communication. 25.8.2008

Jones AR, Eisenacher M, Mayer G, Kohlbacher O. et al. (2012), The mzIdentML data standard for mass spectrometry-based proteomics results. *Molecular and Cellular Proteomics*: MCP. 2012;11(**7**):M111 014381. doi: 10.1074/mcp.M111.014381.

Metabolomics Journal, Instructions for authors (2016), http://www.springer.com/life+sciences/biochemistry+%26+biophysics/journal/11306?detailsPage=pltci_1709154#, *accessed 19.6.2016*

PhenoMeNal website (2016) http://phenomenal-h2020.eu/home, *accessed 17.6.2016*

Haug K., Salek R., Conesa P., Hastings J., et al. (2013), MetaboLights—an open-access general-purpose repository for metabolomics studies and associated meta-data. *Nucl. Acids Res.* **41** (D1): D781-D786. doi: 10.1093/nar/gks1004

Sansone S.A., Rocca-Serra P., Field D., Maguire E., Taylor C., et al. (2012), Toward interoperable bioscience data. *Nature Genetics*, **44**,121–126 (2012), doi:10.1038/ng.1054

Schober D., Boeker M. (2010), Ontology Simplification: new buzzword or real need? *OBML 2010 Workshop Proceedings*. Edited by: Herre H, Hoehndorf R, Kelso J, Schulz S, Institut fuer Medizinische Informatik, Statistik und Epidemiologie (IMISE), Markus Loeffler. 2010, M1-5

Schober D., Smith B., Lewis S.E., Kusnierczyk W., Lomax J., Mungall C., Taylor C.F., Rocca-Serra P., Sansone S-A. (2009), Survey-based naming conventions for use in OBO Foundry ontology development. *BMC bioinformatics* 2009, 10:125.

Schymanski, E. L., Jeon, J., Gulde, R., Fenner, K., Ruff, M., Singer,H. P., et al. (2014), Identifying small molecules via high resolution mass spectrometry: communicating confidence. *Environmental Science and Technology*, 48 **(4)**, 2097–2098. doi:10.1021/es5002105

Strehmel N., Bottcher C., Schmidt S., Scheel D. (2014), Profiling of secondary metabolites in root exudates of Arabidopsis thaliana. *Phytochemistry* **108**:35–46

Sumner L.W., Amberg A., Barrett D., Beale M. et al. (2007), Proposed minimum reporting standards for chemical analysis. *Metabolomics*. 2007;**3**(3):211–221. doi: 10.1007/s11306-007-0082-2.

Sumner L. W., Lei Z., Nikolau B.J., Saito K., Roessner U., Trengove R. (2014): Proposed quantitative and alphanumeric metabolite identification metrics. *Metabolomics* **10**:1047–1049. doi:10.1007/s11306-014-0739-6.