A project-scale map of metadata to improve future data management

Mélanie Pétéra, Cecile Cabasson, Blandine Comte, Christophe Duperier, Olivier Filangi, Sylvain Prigent, Estelle Pujos-Guillot, Franck Giacomoni

A project-scale map of metadata to improve future data management

Mélanie PÉTÉRA[1,5]; Cécile CABASSON[2,3,5]; Blandine COMTE[1,5]; Christophe DUPÉRIER[1,5]; Olivier FILANGI[4,5]; Sylvain PRIGENT[2,3,5]; The MetaboHUB Consortium[5]; Estelle PUJOS-GUILLOT[1,5]; Franck GIACOMONI[1,5]

[1] Université Clermont Auvergne, INRAE, UNH, Plateforme d'Exploration du Métabolisme, MetaboHUB Clermont, F-63000 Clermont-Ferrand, France
[2] Univ. Bordeaux, INRAE, UMR1332 BFP, F-33882 Villenave d'Ornon, France
[3] Bordeaux Metabolome, MetaboHUB, PHENOME-EMPHASIS, F-33140 Villenave d'Ornon, France
[4] IGEPP, INRAE, Institut Agro, Université de Rennes, F-35653 Le Rheu, France
[5] ANR INBS-0010, Toulouse, France

Introduction

Today, the intra-lab application of best practices in the metabolomics field usually guarantees an adequate data exploitation within a single lab. However, the growing interest in multi-analyses designs (e.g. complementary analytical platforms, variety of matrices, multi-omics), as well as the need of data sharing and reuse, increase the difficulty of data management. Indeed, managing the multiplicity and the heterogeneity of information involved is required to achieve relevant knowledge extraction from metabolomics data.

Within the MetaboHUB national infrastructure (MTH), one objective is to optimize data handling, especially metadata, to facilitate large-scale analyses, multi-platforms studies, and data FAIRisation (Findability, Accessibility, Interoperability, Reusability). In particular, this fits in the MetaboHUB scientific roadmap that promotes the open science development in the field of metabolomics.

Material & Method

In the context of metabolomic and lipidomic studies, data production and analysis come along with a large diversity of metadata (data of the data). To identify clearly-defined bottlenecks and targets for future improvement in data management, the objective of this work was to build a metadata map at the scale of a scientific project. Aiming for completeness, this map was constructed in a collaborative and multidisciplinary way involving chemists, biologists, data stewards as well as computer scientists, combining their respective experience and knowledge.

Materials:

- MetaboHUB expertise combining a large variety of skills, profiles and fields
- Data repository resource examples where metadata are central (*e.g.* EBI MetaboLights[1])
- Publications on the data cycle (*e.g.* Savoi et al - Grapevine community[2])
- Online ontology resources (*e.g.* BioPortal[3], AgroPortal[4], OLS ontology search[5])

Methodology:

- Common meetings, including all area experts: ***topics borders, constitutive of the future map***
- Iterative map building through various methods: face to face, by groups, by fields and scales: ***we do not want to miss any metadata.***
- Work on information representation. How to represent the diversity and the complexity of different points of view and interests (*e.g.* chemistry *vs.* biology *vs.* bioinformatics)? Examples:
  - ✓ Which choice to make to group metadata in categories depending on how to use the mapping afterward?
  - ✓ How to combine multiple grouping, without loosing essential information as well as keeping the map easy-to-handle for all actors?
- Work on how to transform this preliminary map into active and productive initiatives, useful for FAIR data and the open science movement.

Results

We constructed a project-oriented map of metadata that sorts and labels metadata through various angles. This complete map can be adapted in refined maps depending on the objective of representation. In this communication, we present a specific implementation of our first built map (Figure 1) where we display seven general topics to address through three temporal steps.

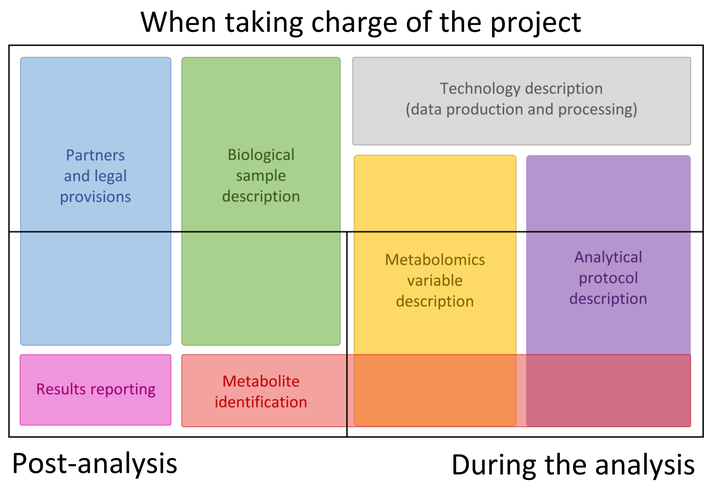

Figure 1: A map that links the timeline of metadata generation and the various types of metadata at the scale of a scientific project

Topics were also declined in more accurate aspects, illustrated in Figure 2. These aspects in each topic area, are source of metadata, usually difficult to store, retrieve and publish. Based on the resulting metadata map, targets (areas and topics) to be further investigated were identified, enabling the construction of transversal working groups at the MetaboHUB consortium scale. In particular, this work enables to focus efforts on clearly defined issues to improve standardisation of practices regarding data management and metadata documentation.

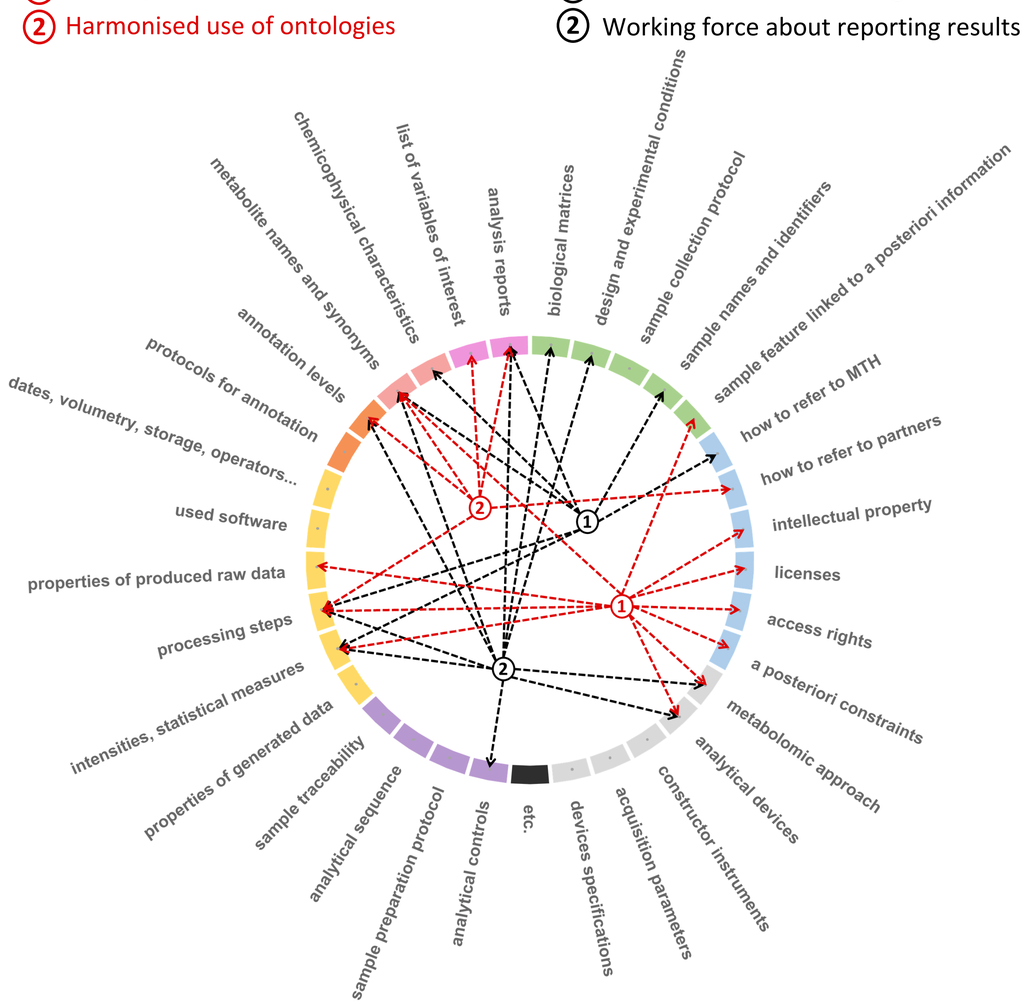

Figure 2: Examples of metadata types listed by fields identified in Figure 1 and examples of mapping with possible working topics and MetaboHUB task forces activities

Example of mapping of a concrete MTH development project:
The project Metabolomics Semantics Data Lake aims to build and provide a large-scale distributed infrastructure for data-processing and massive integration of semantic information about metabolomics data studies. This system is strongly structured from semantic web technologies to maximize the reuse of existent ontologies and knowledge (NCBI, EBI) and finally manage heterogeneous metabolomics content from the MetaboHUB consortium.

Interacting fields identified with the scope of routine usage:

- processing step
- used software
- dates, volumetry, storage
- analyses reports
- sample-collection processing
- sample names and identifiers
- sample feature linked to a posteriori information

Conclusion

In conclusion, this collaborative map construction has been shown to be an efficient tool to draw a clear « where do we stand / where do we go » picture inside a national infrastructure like **MetaboHUB** regarding project-scale metadata. This facilitates the definition of a precise data management. Such an approach could be translated within other infrastructures, consortia and/or communities.

SCAN ME

References:
[1] Kenneth Haug, Keeva Cochrane, Venkata Chandrasekhar Nainala, Mark Williams, Jiakang Chang, Kalai Vanii Jayaseelan, Claire O'Donovan. MetaboLights: a resource evolving in response to the needs of its scientific community. Nucleic Acids Research, gkz1019, doi:10.1093/nar/gkz1019, PMID:31691833.
S.; Schwander, F.; Gougeon, R.D.; Ferreira, A.C.S.; Theodoridis, G.; Töpfer, R.; Vrhovsek, U.; Adam-Blondon, A.-F.; Pezzotti, M.; Mattivi, F. Grapevine and Wine Metabolomics-Based Guidelines for FAIR Data and Metadata Management. Metabolites 2021, 11, 757. https://doi.org/10.3390/metabo11110757
MA. BioPortal: enhanced functionality via new Web services from the National Center for Biomedical Ontology to access and use ontologies in software applications. Nucleic Acids Res. 2011 Jul;39(Web Server issue):W541-5. Epub 2011 Jun 14.
vocabulary and ontology repository for agronomy. Comput. Electron. Agric., 144 (2018), pp. 126-143
[4] C. Jonquet, A. Toulet, E. Arnaud, S. Aubin, E.D. Yeumo, V. Emonet, J. Graybeal, M.-A. Laporte, M.A. Musen, V. Pesce, et al. AgroPortal: a
[5] Jupp S. et al. (2015) A new Ontology Lookup Service at EMBL-EBI. In: Malone, J. et al. (eds.) Proceedings of SWAT4LS International Conference 2015