Metadata standards and tools in Life Sciences – an overview

August 2020. Eliane Blumer, Sitthida Samath (EPFL Library)

Table of content

« Executive Summary » 3
1. Context of study 6
2. Disciplinary scope and definitions 6
3. Methods 7
Phase 1 : Literature review, searching for standards 7
Phase 2 : Survey in the EPFL Life Sciences community 7
Conducting the survey 8
Phase 3 : online survey follow-up with in-depth interviews 8
4. Results 8
Phase 1 : literature review results, searching for standards 8
Phase 2: survey results and analysis 9
a) Data cleaning and preprocessing 9
b) Participation 9
c) Results and analysis on standards 10
d) Results and analysis on tools 18
e) Contribution to standard development, global result 21
f) What are metadata standards good for, global results 22
g) Free answer field, global results 23
h) Conclusion 24
Phase 3: Interviews results and analysis 25
Conclusion for interviews 26
5. Overall and final reflections 26
Next steps 27
Annex 1 – Literature analysis working logs (in French) 28

Bases de données 16 et 18/10/2019 28
Implémentations outils et sites d'éditeurs 16/10/2019 28
Entrepôts disciplinaires à partir de re3data 16/10/2019 29
Répertoires de vocabulaires et standards 16/10/2019 29
RDA metadata standard directory 29
Base BARTOC 29
Répertoire LOV 29
Fairsharing 16/10/2019 29
Existant 16/10/2019 30
Réserves émises à l'issue de la revue de littérature 30
Annex 2 – Analysis survey additional methods (in French) 31
Préparation des données pour la question sur les standards 31
Annex 3 – Blank questionnaire 32

« Executive Summary »

In mid-2019, EPFL Library launched a survey on bioinformatics tools and training needs in the EPFL Life Sciences community, with few results (9 responses). At the same time, we acknowledged a lack of strong references on the use of metadata standards at EPFL whereas these could prove useful for the Research Data team (recommendations when reading Data Management Plans, during consulting or training appointments…).

As a result, we decided to resume the study with a new and broader scope: Tools AND Standards. By standard, we mean:

- terminological resources (vocabularies, terminologies, classifications, thesauri),
- formats and data models / schemas,
- structured knowledge bases (databases, reference databases, ontologies).

And by tools, we mean

- bioinformatics software (i.e. for sequence or molecular structure analysis of proteins and genes)
- databases from the Life Sciences field (i.e. genome databases)

Moreover, we adopted an express and minimalist approach, aiming at acquiring the maximum amount of knowledge with a minimum means involved, ideally in a 6 months' time. The new study consists of 3 phases:

1. Literature review of standards
2. Survey in EPFL Life Sciences Community
3. Follow-up interviews with interested participants

Starting the project in the fall of 2019, our goal was twofold: on the one hand, to gain new knowledge and insights, and on the other hand, to develop a reproducible survey methodology resolutely based on liaison librarian-data librarian collaboration.

Literature review – This part must answer the following question: what metadata standards exist in the Life Sciences community, especially in the fields of research represented at EPFL? Literature review for tools was done already. The search is carried out in :

- Bibliographic databases, PubMed, PLOS Biology etc.
- Tools or software editors, such as RedCap, Rspace, UniProt etc.
- Disciplinary repositories, such as DDI, Genome Metadata etc.

- Vocabularies and standards repositories, such as BARTOC or Fairsharing

At the end of the literature review, we got a set of 46 results meant to serve as a basis for the forthcoming questionnaire: 3 terminology resources, 31 formats and templates/schemas, 11 structured knowledge bases.

Survey – This part must answer the following question: which Life Sciences metadata standards are actually known and used at EPFL, in the School of Life Sciences? Targeted respondents are EPFL School of Life Sciences labs members and staff. We exclude undergraduate students. The questionnaire is launched online in early 2020, with one relaunch at mid-term, through EPFL mail groups. It includes a limited number of questions to not take more than 5 minutes to complete: one question about standards, one question about tools, and finally a few open questions to better identify the respondents and their relationship to the said tools and standards.

We received 51 complete answers (N.B: population of collaborators of EPFL in Life Sciences, which is roughly our target population, is 626 FTE in 2019), full with comments from the respondents, which, once analyzed, allowed us to get the following 3 main results:

- A "podium" of standards and a "podium" of the most used tools in Life Sciences community, in general and in the different groups in particular;
- The groups are profiled according to their degree of familiarity with the said tools and standards;
- A set of initiatives, achievements, insights and suggestions from the respondents.

Interviews – This part must answer the following question: do interviewees validate the results? And what are they expecting next? At the end of May 2020, we conducted interviews with four of the eight respondents who expressed the wish to be contacted again, a small group but with remarkably varied profiles (systems specialists, group leader, doctoral student).

- 1st observation: during interviews we highlighted the answers rather than "peer validated" the results of the survey.
- 2nd observation: the respondents had a better understanding of the generic term "standard" rather than the more specific term "metadata standard ".

- 3rd observation: the richness of the opinions expressed and the various actions suggestions provide abundant material for our conclusions and course of action after the study.

At the end of the study, seven main possible lines of action emerged, from raising awareness/information to lobbying, including also training and consulting (our two usual modes of action). Our immediate two goals are: promoting these results to the Life Sciences Community and taking up the study in a new discipline to extend our knowledge base and services about standards and tools at EPFL.

1. Context of study

The original goal of EPFL Library Research Data team is to find ways to offer to EPFL labs and researchers a more discipline-oriented support in the field of metadata standards and tools. This pilot study in the School of Life Science helps gaining disciplinary-oriented knowledge and defining a methodology which can later be reused in other disciplines.

2. Disciplinary scope and definitions

By metadata standards, we mean three categories of standard resources:

- terminological resources (vocabularies, terminologies, classifications, thesauri),
- formats and data models / schemas,
- structured knowledge bases (databases, reference databases, ontologies).

By tool, we mean:

- bioinformatics software (i.e. for sequence or molecular structure analysis of proteins and genes)
- databases from the Life Sciences field (i.e. genome databases)

Our working scope is as follows:

- EPFL Life Sciences Community = labs members and staff in the School of Life Sciences. We exclude undergraduate students. We deliberately choose a very local scope in order to gain accurate results, as we know that many metadata standards are already used and implemented. The scope also bridges liaison/disciplinary expertise and metadata expertise.
- Therefore, the study sometimes focuses on the research institutes of the School of Life Sciences at EPFL, on their commonalities and differences: Institute of Bioengineering, Swiss Institute for Experimental Cancer Research, Global Health Institute, Neuroscience – Brain Mind Institute, Center for Neuroprosthetics, Blue Brain Project. In the survey, respondents were to choose themselves the research group/institute that fitted them the best (self-declarative belonging to the group).

Above all, we want to achieve standards and tools practice assessment:

- Which ones are widespread?
- Which ones are used in the researcher's surroundings?

- Which ones are applicable in one's research environment?

3. Methods

Phase 1 : Literature review, searching for standards

We focused on searching for standards. Tools were already identified during previous study (July 2019).

The main sources were as follows:

1. Searching grey literature, professional and scientific literature, with keyword "Biology", "Lifesciences", "metadata standard", "metadata schema", "database model" in the reference bases Pubmed, Scopus (Life sciences domain), WOS (Life sciences domain), PLOS Biology, PLOS Neuroscience, PLOS Medicine, PLOS Computational Biology, Bio Arxiv.
2. Examining implementations of the standards in tools or softwares as publishers state it, such as Redcap, Bioformats, Rspace, openBis, Uniprot, Ensembl, i2b2, European genome phenome archive, Protein databank, etc.
3. Examining implementations in disciplinary data repositories, with re3data.org as a starting point.
4. Exploring vocabulary and standards knowledge base, including RDA metadata standard directory, LOV, BARTOC base, Fairsharing base, Open metadata directory (deprecated, see Annex 1 for reasons).

For reasons of time and efficiency, and because other sources have already provided usable results, we have decided not to extend the search to recommendations from publishers or journals in our review. We also chose not to sift through LibGuides or guidelines from reference institutions, such as NIH, Wellcome trust, INSERM, CHUV, ETHZ, etc.

As a result, our search is not a comprehensive one, rather a compromise between investment and expected result. We primarily aim at sketching the metadata standards and tools landscape at EPFL.

Phase 2 : Survey in the EPFL Life Sciences community

Based on the results of the literature review, the structure and content of the survey is twofold:

- What standards and tools do you practice? (collection of knowledge)
- Who are you? (assessing the answers)

Conducting the survey

For the survey, an online questionnaire implementation with Survey Hero survey engine was chosen, displaying standards by research area for relevancy reason (see blank questionnaire in Annex 3). The questionnaire was proofread and commented by colleagues in EPFL Library and the School of Life Sciences before launch. Targeted respondents were EPFL School of Life Sciences labs members and staff. We excluded undergraduate students. The emailing invitation was sent to Life Science Community, through the corresponding mailing groups with the help of EPFL Library communication officer via proofreading, comments, layout in the Mailchimp environment with the EPFL corporate identity, preparation of the sending, w/ additional reach out on Twitter. The questionnaire opened on 13/02/2020 11:45, was relaunched at mid-term on 18/02/2020, and finally closed on 29/02/2020.s

After collection, data was prepared and reprocessed to allow proper analysis. More specifically, apart cleaning and reformatting, we generated radar graphs based on tools usage data and applied a hot-cold colour pattern to metadata usage data (see Annex 2).

Phase 3 : online survey follow-up with in-depth interviews

Our main goal in this third step is to assess/peer-review the results and discuss insights and possible outcomes with interviewees.

We devised short sessions (15-30 minutes max) for semi-structured interviews with 3 main questions to ask:

1. Assess results and consolidate knowledge
2. Discuss the various outcomes and suggestions
3. Hook up to the research process in general and metadata activities in particular

This qualitative data was then summed-up in a synthetic chart.

4. Results

Phase 1 : literature review results, searching for standards

Final selection 22/10/2019 includes 46 results of which:

- 3 terminology resources
- 31 formats and templates/layouts
- 11 structured knowledge bases

For more details, see Annex 1.

At the end of this phase, we had few standard references specifically relevant for Cancer research and Neurosciences. We decided that the second phase of the study (survey) would help confirm this. If needed, respondents would help source missing standards.

Phase 2: survey results and analysis

a) Data cleaning and preprocessing

- Results are exported from survey platform in excel format.
- Empty answers and test answers are discarded. There are no incomplete answers, only wholly empty or wholly complete answers. Among a total of 74 answers, 51 complete answers are kept for analysis.
- One file contains answers per respondent group: one tab per group with all questions for this group
- One file contains answers per question irrespective of group: one tab per question with all responses
- Data undergoes a “count” transformation processing: counts the number of answer values "very familiar", "quite", "little", "not", "not at all", empty (counted as "not at all"), yes/no, "efficiency", "compliance", "sharing", "don't know"...
- See dataset and complete description on Zenodo:
  Blumer, Eliane, & Samath, Sitthida. (2020). Metadata standards and tools practice at EPFL School of Life Sciences 2020 Survey [Data set]. Zenodo. http://doi.org/10.5281/zenodo.4003720

b) Participation

- 51 complete responses. No incomplete response, only wholly empty or wholly complete answers.
- In the absence of exact figures concerning the number of people actually reached by the questionnaire, let us note as an indication that the population of collaborators of EPFL in Life Sciences (which is roughly our target population) is 626 FTE in 2019 (according to institutional statistics 2019 https://www.epfl.ch/about/overview/fr/statistiques-institutionnelles/statistiques-personnel/).
- 8 respondents left their contact information for follow-up interviews, that is to say about 1/6 respondents is motivated to participate.

It is noticeable that Bioengineering (14 respondents) and "neuro-x" (Brain Mind Institute, Center for Neuroprosthetics and Blue Brain Project grouped together for analysis, 20 respondents) are the two largest groups (see Table 1 and Figure 1).

Research group, research institute	Blue Brain Project	Center for Neuroprosthetics	Global health Institute	Institute of Bioengineering	Brain Mind Institute	Swiss Institute for Experimental Cancer Research
Number of respondents	1	5	9	14	14	8

Table 1 Respondents per institutes

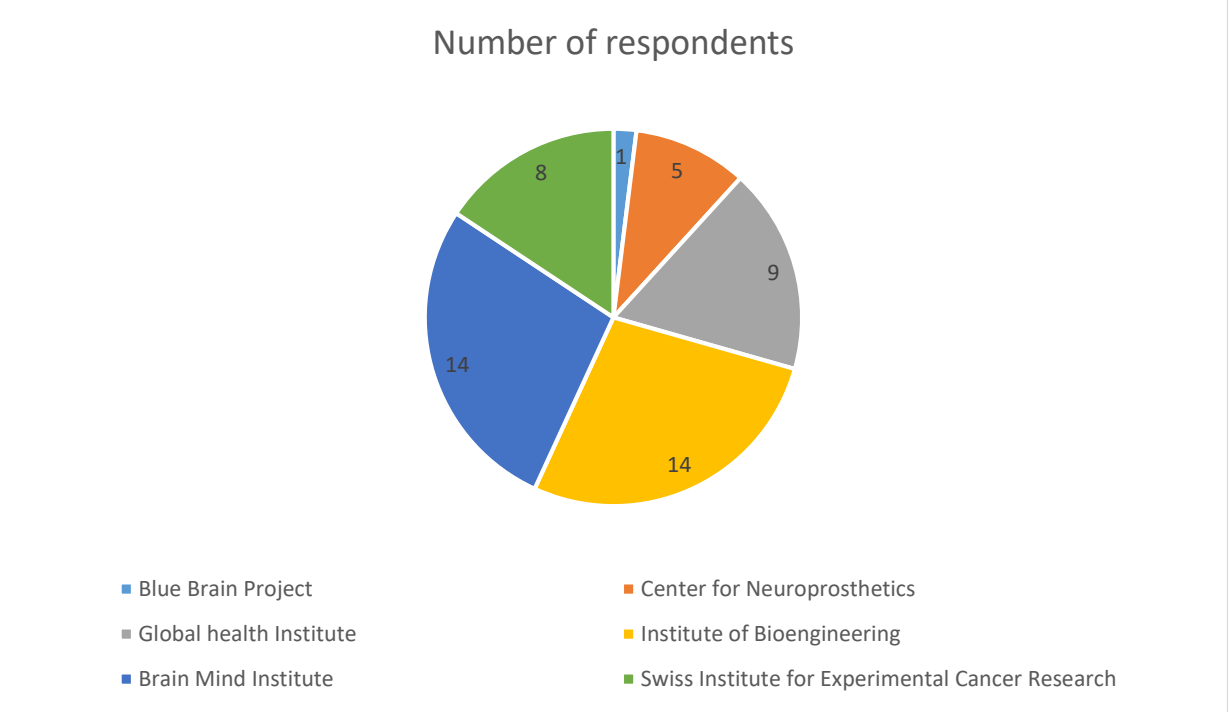

Figure 1 Respondents per research group/institutes, graphic

c) Results and analysis on standards

Standards in Bioengineering

Summary: Bioengineering is found to be a community that uses a large number of standards, often with daily or frequent use (see colour distribution in Figure 2). Table 2 summarizes the main results for this group.

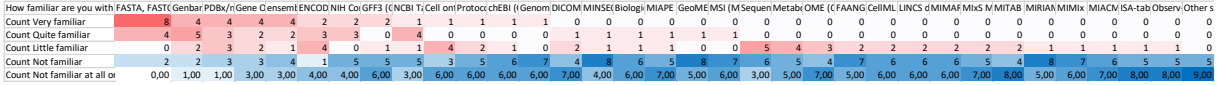

Figure 2 Bioengineering standard familiarity

1. Overall group profile	Pretty familiar with standards. Nice dispersion with a wide gradation of red and pink over all the standards proposed in the survey. NB: Bioengineering is the field where we identified the largest number of standards during the literature review phase
2. Best known standards	Top seven seems quite different from that of other groups (i.e. genomics) Top seven standards are: - FASTA, FASTQ - Genbank sequence format - PDBx/mmCIF Dictionary (Protein Data Bank, Crystallographic Information File) - Gene Ontology - ensembl - ENCODE (Encyclopedia of DNA Elements) - NIH Common data elements
3. Standards that we did not identify or wrongfully removed during prior preparation steps	NA
4. Identified standards falling within the group's area of interest	Good relevance, red-coloured or pastel areas

Table 2 Bioengineering summary

Standards in Global Health

Summary : Global Health is a respondent group with seemingly light standards usage (few standards are cited, usage seems infrequent, see Figure 3 and Table 3), very different from the Bioengineering group above.

How familiar are you wit	NIH (	PDBx	SNOI	Sequ	ODM	OME	Biolo	Biotc	Protc	MIBE	CellN	COM	GSCI	ISA-t	Med	MIAC	MIRI	DDI (	DICO	MITA	Othe
Count Very familiar	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Count Quite familiar	2	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Count Little familiar	0	2	0	3	2	2	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0
Count Not familiar	5	2	3	4	3	3	4	3	2	5	4	4	4	4	4	4	4	3	3	3	2
Count Not familiar at all	2	4	5	2	4	4	4	5	6	4	5	5	5	5	5	5	5	6	6	6	7

Figure 3 Global Health Standard Familiarity

1. Overall group profile	A group different from the others, less familiar with standards. Color pattern is mostly white and blue. The “very familiar” value was never ticked (see Figure 3).
2. Best known standards	Top seven is quite similar with that of the two other analyzed communities below. Top seven standards: NIH Common data elements PDBx/mmCIF Dictionary (Protein Data Bank, Crystallographic Information File) SNOMED-CT Sequence Ontology (SO) ODM, ODM-XML (CDISC Clinical data interchange standards consortium CDISC operational data model) OME (Open Microscopy Environment Data model, OME-XML, OME-TIFF) Biological Pathway Exchange (BioPAX)
3. Standards that we did not identify or wrongfully removed during prior preparation steps	NA
4. Standards falling within the group's area of interest (broad confidence zone)	Selection seems relevant but usage stays low.

Table 3 Global Health Summary

Standards in Cancer research

Summary : a group with moderate use of standards (limited set of standards, frequent to infrequent usage, see Figure 4 and Table 4).

It is noteworthy that respondents named several standards that were not mentioned in the survey list.

How familiar are you wit	OME (	Other	PDBx/	NIH Cc	Biolog	Seque	CellMl	DICOM	Metab	ISA-tal	LINCS	MIACM	MIRIA	MITAB	ODM,	Protoc
Count Very familiar	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Count Quite familiar	0	2	1	1	0	0	0	0	0	0	0	0	0	0	0	0
Count Little familiar	1	0	2	1	2	2	1	1	0	0	0	0	0	0	0	0
Count Not familiar	3	2	3	2	3	3	3	3	4	3	3	3	3	3	2	2
Count Not familiar at all	3	4	2	4	3	3	4	4	4	5	5	5	5	5	6	6

Figure 4 Cancer Research Standard Familiarity

1. Overall group profile	Profile seems similar with the one from Global Health. The group seems little familiar with standards. Color pattern is mostly white and blue (see Figure 4).
2. Best known standards	Top seven is quite similar with that of the other following groups. Top seven standards are: - #2 is « Others » , otherwise : - OME (Open Microscopy Environment Data model, OME-XML, OME-TIFF) - PDBx/mmCIF Dictionary (Protein Data Bank, Crystallographic Information File) - NIH Common data elements - Biological Pathway Exchange (BioPAX) - Sequence Ontology (SO) - CellML metadata - DICOM
3. Standards that we did not identify or wrongfully removed during prior preparation steps	« Others » is #2 in the top seven. This comes as a confirmation of our observation that results of the literature

	review phase was lacking Cancer research references.
4. Standards falling within the group's area of interest (broad confidence zone)	Selection is relevant but seems little used nonetheless

Table 4 Cancer Research Summary

Standards in Neurosciences ("Neuro-x": Brain Mind Institute, Center for Neuroprosthetics and Blue Brain Project grouped together)

Summary : a group that uses most of the standards mentioned in the survey, with a seldom to very frequent usage (see Figure 5 and Table 5).

How familiar are yc	Other	DICOM	OME (	Seque	NIH C	NINDS	PDBx/	Protoc	MIRIA	ISA-ta	Biolog	MITAE	CellMI	MIACM
Count Very familiar	3	2	1	1	0	0	0	0	0	0	0	0	0	0
Count Quite familia	3	4	1	0	1	1	1	1	1	0	0	0	0	0
Count Little familia	1	1	1	1	2	2	1	0	0	1	1	1	1	1
Count Not familiar	3	4	5	4	8	6	4	6	5	5	4	4	3	2
Count Not familiar	10	9	12	14	9	11	14	13	14	14	15	15	16	17

Figure 5 Neuro-X standard familiarity

In this respondent group, « Others » is the number one answer. This comes as a confirmation of our observation that results of the literature review phase was lacking Neuroscience research references. Respondents declared the following standards as missing:

- Neurodata without borders (NWB) x2,
- BIDS (Brain Imaging Data Structure),
- Neuroimaging Informatics Technology Initiative (.nfti),
- BrainVision Core Data Format (.vhdr, .vmrk, .eeg, adopted in BIDS in 2019),
- European data format (.edf, for medical or biological time series data),
- MINI (Mini International Neuropsychiatric Interview),
- FAIR,
- Flybase.

As a result, the list of most relevant standards for this group has considerably expanded.

1. Overall group profile	The group seems rather familiar with standards but not necessarily with the standards mentioned in the survey. Wide dispersion, but the number of standards is smaller than that of all the other groups (see Figure 5)
2. Best known standards	Top seven is quite similar with that of the other previous groups, Bioengineering group excepted.

	Top seven: - « Others » is #1 (see above) otherwise - DICOM - OME (Open Microscopy Environment Data model, OME-XML, OME-TIFF) - Sequence Ontology (SO) - NIH Common data elements - NINDS Common Data Elements - PDBx/mmCIF Dictionary (Protein Data Bank, - Crystallographic Information File) - Protocol Data Element Definition
3. Standards that we did not identify or wrongfully removed during prior preparation steps	« Others » is number one in the top seven. This comes as a confirmation of our observation that results of the literature review phase was lacking Neuroscience research references.
4. Standards falling within the group's area of interest (broad confidence zone)	The “not familiar” line is mostly filled by a high number of respondents. A clear majority of respondents has no concern at all for the standards mentioned in the survey.

Table 5 Neuro-X summary

Overall EPFL Life Science Community results : standards « podium »

Standard (belonging to disciplinary top 7s)	Number of groups using this standard
NIH Common data elements	4
PDBx/mmCIF Dictionary (Protein Data Bank, Crystallographic Information File)	4
OME (Open Microscopy Environment Data model, OME-XML, OME-TIFF)	3
Sequence Ontology (SO)	3
Biological Pathway Exchange (BioPAX)	2
DICOM	2
CellML metadata	1
ENCODE (Encyclopedia of DNA Elements)	1
ensembl	1
FASTA, FASTQ	1
Genbank sequence format	1
Gene Ontology	1
NINDS Common Data Elements	1
ODM, ODM-XML (CDISC Clinical data interchange standards consortium CDISC operational data model)	1
Protocol Data Element Definition	1
SNOMED-CT	1

Table 6 Standards « Podium »

d) Results and analysis on tools

Tools in Bio Engineering

In general, rather unknown, with a peak in very familiar in the following tools :

- 6 occurrences : Protein Sequence databases (e.g. Uniprot, RefSeq, InterPro, PROSITE)
- 5 : Primary nucleotide sequence databases (e.g. GenBank, European Nucleotide Archive)
- 5 : Genome databases (e.g. Ensembl, UCSC Genome Browser)
- 4 : Signal transduction/Metabolic pathway databases (e.g. Reactome, KEGG)
- 4 : Survey, database and data management tools (i.e. RedCap, SLIMS)

In the quite familiar, the following categories of tool adds:

- 5 occurrences: Protein structure/Model databases (e.g. PDB, SWISS-MODEL)
- 5 : Gene expression/Microarray databases (e.g. ArrayExpress, Gene Expression Omnibus)

Little familiar, participants report :

- 4 occurences: Signal transduction/Metabolic pathway databases (e.g. Reactome, KEGG)
- 4 : Referencing of data (i.e. Identifiers.org, RRID portal)
- 4 : Model organism databases (e.g. WormBase, Mouse Genome Informatics)

Not familiar, but interested in :

- Quite a lot of interested in different tools (nowhere no answer at all, the minimum answer quote is 3) with the most in (only up to six mentions highlighted here):
    - 9 occurences: Protein Annotation (i.e. HAMAP)
    - 8 : Mutation databases (e.g. OMIM, HGMD, dbSNP)
    - 6 : Public protein data repository (i.e. PRIDE Archive)
    - 6 : Gene ontology annotation (i.e. GOA)
    - 6 : Protein-protein and other molecular interaction (e.g. Intact, String)

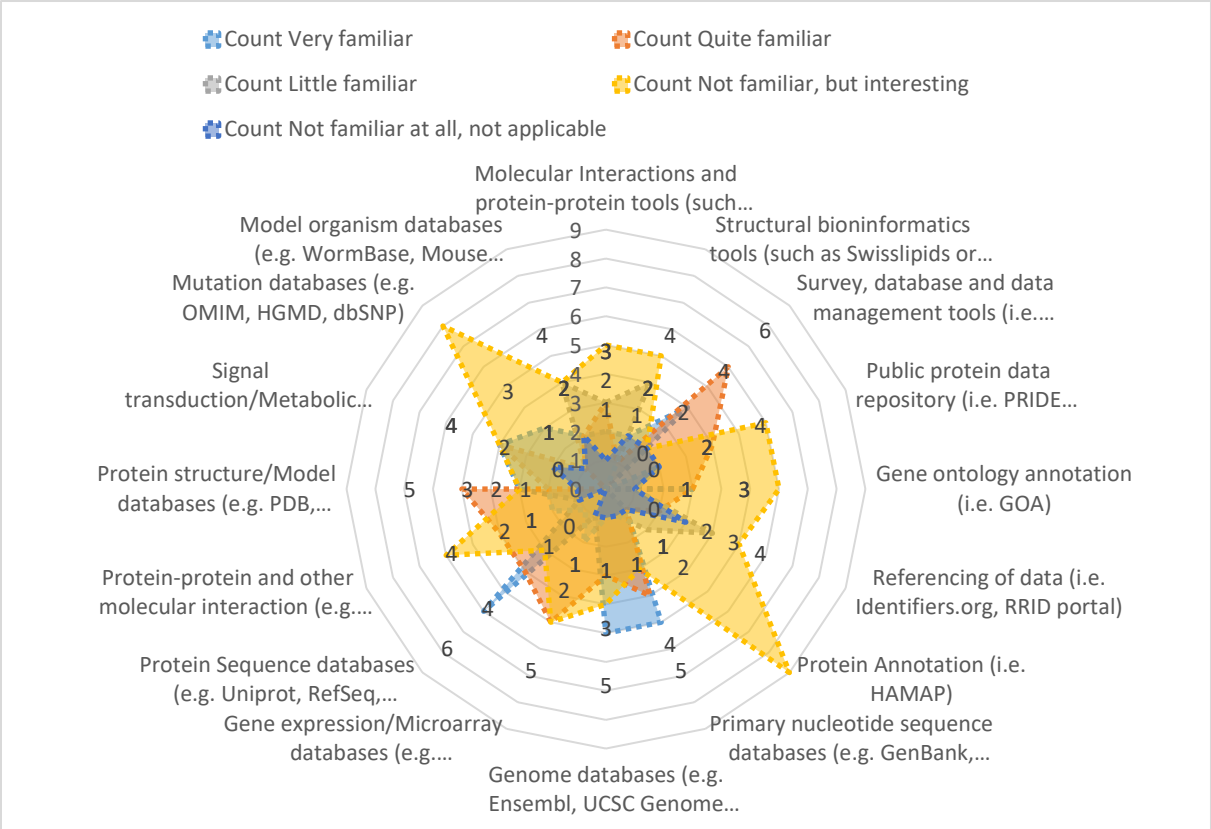

Figure 6 Bioengineering tools familiarity

Tools in Neuroscience

This seems not to be the respondent group with the most of familiarity in tools mentioned in our survey, it might however also be that we did not mention the tools that are useful to the community. They mentioned in the "other" section: ModelIDB and Flybase .

There are a few tool families known by a few:

- 6 occurences in little familiar: Protein structure/Model databases (e.g. PDB, SWISS-MODEL)
- 4 occurrences in little familiar: Model organism databases (e.g. WormBase, Mouse Genome Informatics)
- 4 occurrences in little familiar: Survey, database and data management tools (i.e. RedCap, SLIMS)

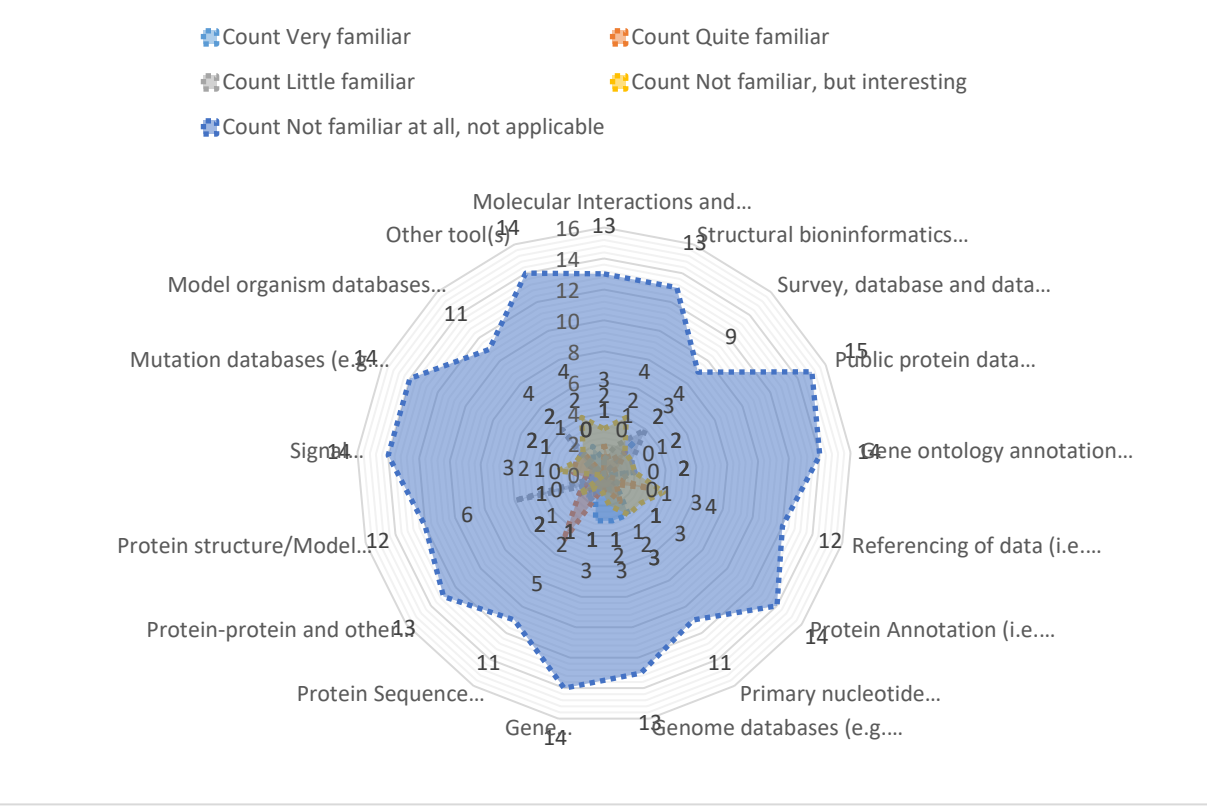

Figure 7 Neuroscience tools familiarity

Tools in Cancer

- Quite distributed among the familiarity
- But very low answer rate, not more than four as a maximum of answer

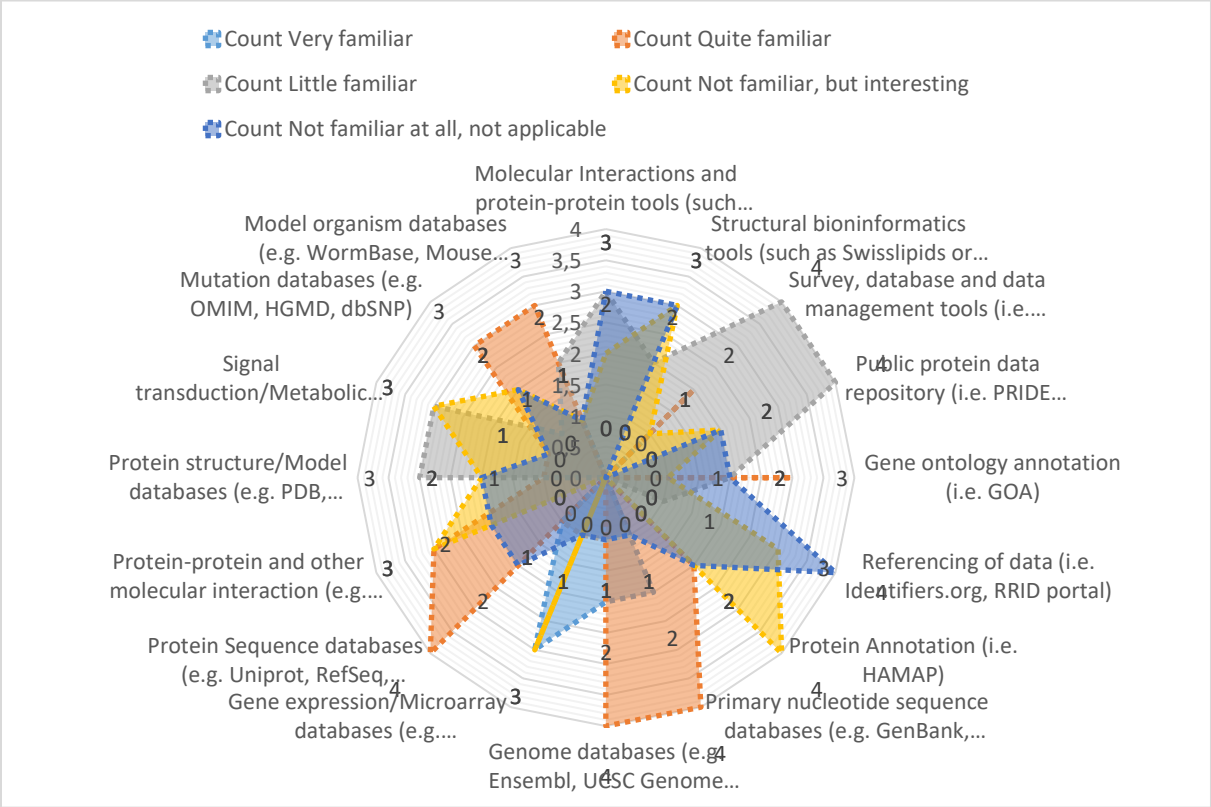

Figure 8 Cancer tools familiarity

e) Contribution to standard development, global result

13/51, which simply means about 1 in 4 respondents claim to be a contributor to a metadata standard effort or a tool effort!

By "effort", in the original question, we meant : design, development, promotion, teaching... Here are the examples of contributions reported in free text:

Activities

- Asking each lab member to **fill the metadata** at least before leaving the lab
- **Pushing OME** for microscopy data
- **Promoting OME** for image processing and analysis + **developing bridges**
- **Metadata documentation** of animals, brain slices, electrophysiology experiments, morphological reconstructions of neurons

(Local !) realizations

- Mass Genome Annotation repository (ndla https://ccg.epfl.ch/mga/ ?)
- Flygut (ndla https://flygut.epfl.ch/ ?)
- Gene Regulation Ensemble Effort for the Knowledge Commons (http://greekc.org/)

f) What are metadata standards good for, global results

Top three with collaboration and self-efficiency (by a narrow margin, so self-discipline wins out over constraint). Moreover, reassuringly few "never thought about it" (see Figure 9 below).

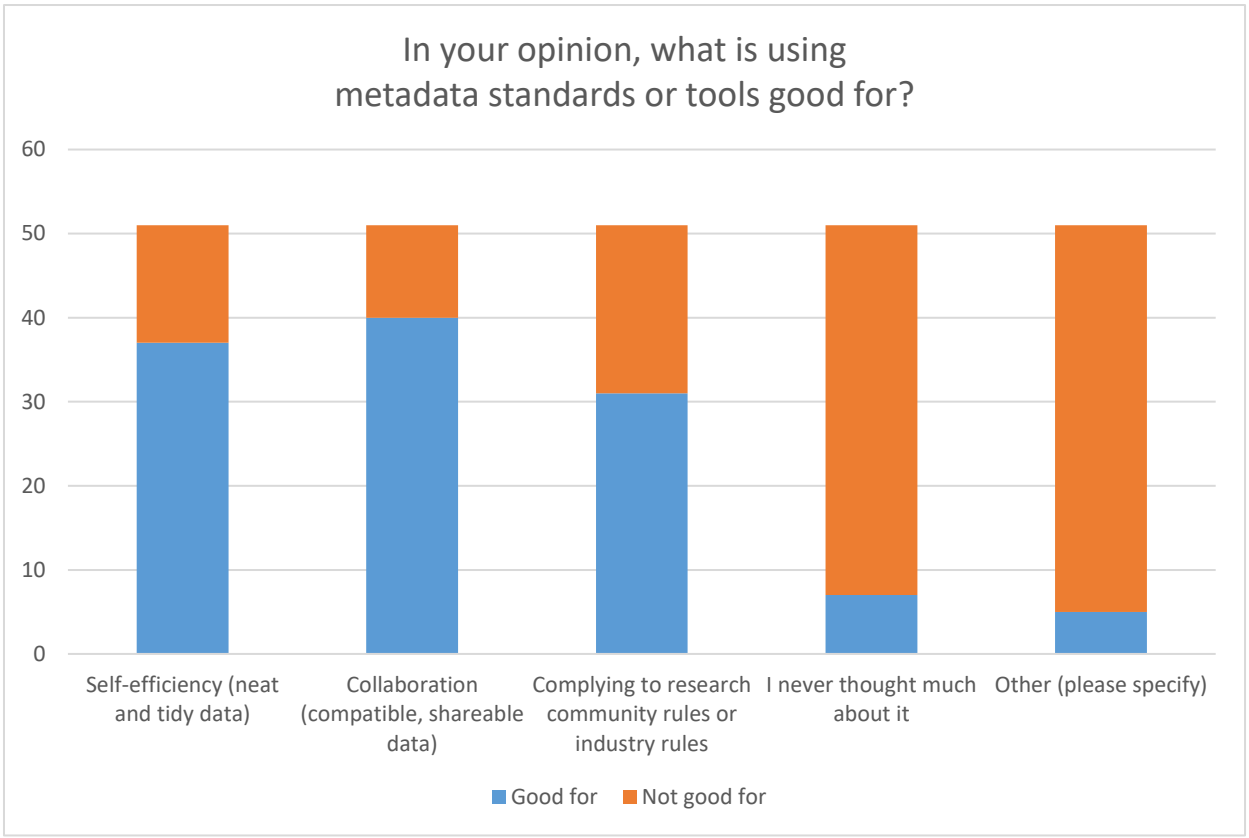

Figure 9 Use of standards

Respondents broadened the perspectives with their own reflections (other, specify) in free text :

- Helping for open data, which allows transparency
- Allowing readers to access published raw data without difficulty. Enabling transparency and reproducibility in research.
- Reproducibility
- Documentation

- Usability of public data. The inexistence, incompleteness, or inaccuracy metadata is the major obstacle to the usability of public data. The lack of standards is an obstacle to data integration and interoperability.

g) Free answer field, global results

- 8 comments / 51 responses
- Q : suggestions, ideas
    - 4 comments: more information, more EPFL knowledge sharing, benchmark, dedicated information support
    - 2 comments: negative aspects, incentives, investment in bio curation
    - 1 comment: annotation of proteins and genes
    - 1 comment: DICOM absence (related to a gap in the survey)
    - Either 6/51, or about 1/9 of the respondents who ask for information and incentives for use.
    - Summary of the profiles of respondents who left their contact information: varied; prof, admin and tech staff, post doc, doc; beginner and advanced; male and female

Figure 10 Free comments

h) Conclusion

There's an abundance of standards and tools existing and also being used within the EPFL School of Life Sciences community.

In general, it seems that a person only has specific knowledge for one type of tool, for example FASTA, probably related to his or her research work. The tendency is not that a person knows all the tools, regardless of their research work.

For standards, the Bioengineering respondent group stands out clearly from the other three respondent groups (Global Health, Cancer research, “Neuro-x”). It seems that it is the discipline that has the most know-how for various tools. This could be linked to their research, or to their training, which tends more towards engineering than biology.

Cancer and “Neuro-x” respondent groups have very little knowledge of tools, based on our results. It is however possible that we have not given the right ones for their research, or that the research is generally less tool oriented.

Phase 3: Interviews results and analysis

ss	Participant 1	Participant 2	Participant 3	Participant 4
Discipline	“Neuro-x”	Cancer research	Bioengineering	Bioengineering
Position	System specialist	Group leader and lecturer	Doctoral assistant	Post-doc, lab engineer with informal data curator tasks
Background	In the process of standardizing use of digital data for the lab, experienced with standard development	Years of working practice in bioinformatics and related standards	Highly interdisciplinary lab different view on data management and standards	Works on data acquisition and processing, methods side and uses platforms. Striving to keep record of data produced in lab, with accurate metadata.
1. Assessment of survey results	NO	YES, good, but usage needs enforcement	NO	NO
2. Perspectives	N/A	Involve reviewers for metadata Train young people to do it and standards that are excepted	Have a volunteer in the lab, develop common guidelines and make people follow the guidelines	N/A lab is well aware
3. Link with research process and tools	N/A	This should be the « common » way because those tools are successful in promoting a standard	N/A	Initiate it at the beginning of the process or just at the start of sequencing . Incentives
4. Necessity of Community	N/A	Difficult to say what the community actually needs	N/A	Community standard looks more like a burden
5. Necessity of Training	N/A	Group "data champions" could help to provide proper metadata. Course could be provided with practical examples:	Training would be really good option, also a mandatory training Recommendations are really nice, but not enough	Support, not training

		Publications with describe data, one could ask students to look at the public data and write a critical report about the metadata and share it At EPFL: Could enforce that the PHDs are checked for interoperability, such as with plagiarism	Timing: Definitively useful when you are having results. When already publishing it is too late, also when writing the paper What about the FOBS but for data management?	
6. Necessity of Guidelines	N/A	N/A	Sharing seems to be a good driver for taking good care of data Basic guidelines for different kind of data would be useful	Still trying to find the drivers apart publication for non-published data and early data

Table 7 Interview results summary

Conclusion for interviews

In this last phase of the study, we got an interesting balance between the various professional and personal profiles of the 4 interviewees: two scientific engineers, one group leader, one post-doc, and thus a complementary plurality of points of view.

Some interviewees have some difficulty to understand what a metadata standard is. This seems to be linked to the stage of career, the younger ones were more interested in general data management issues than metadata standards.

Semi-structured interviews seem to be an useful approach, but the survey assessment part is not as thorough as wanted, therefore not really valid regarding authority of interviewees and number of interviewees. Explanatory part of the answers and development part based on needs and perspectives are much more enriching.

5. Overall and final reflections

This three methods study has shown that there remains a strong urge to further develop and promote the "culture of metadata and documentation" in EPFL Shool of Life Sciences audience, and beyond. The focus of future sensitizing work should be on standard implementation and practical use of such standards, whether with a tool or without.

Only few standards are among the top ten, they seem however to be spread among different research communities.

The abundance of tools and standards is not helping to come up with simple solutions, but asks for a discipline-specific and personalized approach. However, the general overview is currently as well lacking. The results of this study is a first step in this direction.

Next steps

- An EPFL-relevant, exploitable metadata standards, schemas and models memo/knowledge base (planned)
- A status report of metadata standards, schemas and models used in EPFL School of Life Sciences community
- A metadata engineering cost database
- Actual metadata engineering tools and systems, provided by the library or others

Annex 1 – Literature analysis working logs (in French)

Bases de données 16 et 18/10/2019

Source	Date	No de résultats
Pubmed : entre guillemets « metadata standard »	2019/10/21	24 résultats
Bio Arxiv : entre guillemets « metadata standard »	2019/10/18	41 résultats
PLOS Biology : entre guillemets « metadata standard »	2019/10/18	7 résultats
PLOS One Neuroscience : entre guillemets « metadata standard »	2019/10/18	33 résultats
PLOS Medicine : entre guillemets « metadata standard »	2019/10/18	3 résultats
PLOS Computational Biology : entre guillemets « metadata standard »	2019/10/18	8 résultats

- ➔ Sélection de standards parmi les résultats bruts : GeOme, COMPARE data hubs, MIACME, GSCID/BRC Project and Sample Application Standard, cell ontology, ENCODE project, Clinical Data Interchange Standards Consortium (CDISC) Operational Data Model (ODM) standard, LINCS data, CellML metadata
- ➔ Sélection de standards parmi les articles de synthèse (minimum information in life sciences) : MINSEQE, MIxS MIGS/MIMS, MIMARKS, MIMIx, MIAPE, Metabolomics standard Initiative MSI, MIRIAM, MIAPPE, MDM, FAANG, SNOMED-CT, FAST.
  Référence: P. C. Griffin et al., “Best Practice Data Life Cycle Approaches for the Life Sciences,” bioRxiv, p. 167619, Jul. 2017, doi: 10.1101/167619.

Implémentations outils et sites d’éditeurs 16/10/2019

Nous avons passé en revue les outils et bases de données préalablement identifiés et voulu répertorier quels standards ils disent implémenter.

Liste des outils et bases de données passés en revue : Redcap, bioformats, RSpace, openBIS, Uniprot, ensembl, i2b2, European genome phenome archive et Rcsb protein databank.

➔ Sélection: OME, Uniprot, ensembl, i2b2, european genome archive schema, PDBS/mmcif

Entrepôts disciplinaires à partir de re3data 16/10/2019

Dans la liste globale des résultats, il y a une liste de valeurs pour la facette « metadata standard ». La sélection comporte les standards parmi cette liste qui concernent la biologie et les domaines de recherche de l'EPFL

➔ Sélection : DDI, Genome metadata, ISA-tab, MIBBI.

Répertoires de vocabulaires et standards 16/10/2019

RDA metadata standard directory

La sélection comporte les ressources pertinentes en biologie ET en rapport pour les domaines de recherche EPFL présentes dans le répertoire.

➔ Sélection: Darwin Core, Ecological Metadata language, Genome Metadata, ISA-Tab, MIBBI, Observ-OM, OME-XML, PDBx/mmCIF, Protocol Data Element Definitions, Repository-Developed Metadata Schemas (dbEST Expressed Sequence Tag Database, Marine Geoscience Data System)

Base BARTOC

117 résultats pour 500 Pure science> 570 Biology. 160 résultats in 600 Technology > 610 Medicine and Health. Le tri par "rating" serait une fonctionnalité intéressante à exploiter, mais dû à des restrictions techniques (grande précision de la classification) et le grand nombre de résultats, l'étude de la source Bartoc a finalement été écartée (temps d'analyse trop important par rapport à l'objectif de l'étude).

Répertoire LOV

Recherche par tag : Biology

➔ Sélection : 14 resultats dont: boil biological taxonomy vocabulary 0.2 (Core), biopax ontology, biotop, medred ontology, obo ontology, uniprot

Fairsharing 16/10/2019

775 résultats in Life sciences Subject (le plus développé), ainsi que des pages de recommandations. Vu le grand nombre de résultats, l'étude de la source FAIRSharing a finalement été réduite (temps d'analyse trop important par rapport à l'objectif de l'étude) et la sélection s'appuie finalement sur un article de synthèse du contenu de FAIRSharing.

Référence: FAIRsharing Community et al., “FAIRsharing as a community approach to standards, repositories and policies,” Nat Biotechnol, vol. 37, no. 4, pp. 358–367, Apr. 2019, doi: 10.1038/s41587-019-0080-8.

➔ Sélection avec article de synthèse top 10 : FASTA, Gene Ontoloy, PDB, GFF3, ChEBI, NCBI Taxon, Genbank sequence format, schema.org, sequence onotlogy, MITAB

Existant 16/10/2019

➔ Sélection: DICOM, IsaTab, FAIRsharing, OME , RRID portal, NIH Common data elements, NINDS common data elements.

Réserves émises à l’issue de la revue de littérature

Nous n’avons que peu de résultats concernant plus spécifiquement la recherche sur le cancer et les neurosciences → à vérifier et confirmer/infirmer lors des phases suivantes de l’étude.

Annex 2 – Analysis survey additional methods (in French)

Préparation des données pour la question sur les standards

Les résultats sont présentés triés par ordre décroissant sur les valeurs de la 1ère ligne, puis les valeurs de la 2e, puis de la 3e etc. (familiarité décroissante du groupe). Les standards présentés dans les colonnes sont donc triés du plus familier au moins familier pour le groupe, de gauche à droite.

On ajoute sur les tables de résultats une couche coloration en « heatmap » : coloration en rose-rouge pour les 3 premières lignes (familiarité), coloration en bleu ciel-bleu pour les 2 dernières lignes (non familiarité). L'intensité de la coloration est déterminée entre les bornes des valeurs min et max pour chaque groupe. Ainsi, plus le groupe compte de répondants familiarisés avec un standard, plus la coloration sera chaude. Moins plus il compte de répondants non familiarisés, plus la coloration sera froide.

Cela nous permet d'appliquer une lecture synthétique par température et d'identifier :

- Le profil général du groupe : suivant la répartition entre rouge, blanc et bleu. Du rouge et du rose dans la zone supérieure (les 3 premières lignes) laissent supposer une bonne diffusion des standards dans le groupe. Plus de blanc dans cette zone témoigne d'une moindre familiarisation ;
- Les standards les mieux connus dans le groupe : la zone chaude en haut à gauche correspond aux valeurs hautes dans les deux premières lignes very familiar et quite familiar ;
- Les standards les moins connus dans le groupe : la zone froide en bas à droite correspond aux valeurs hautes dans la dernière ligne not familiar at all (ou NA). Cela permet aussi de vérifier la pertinence de la liste de standards proposée dans le questionnaire en comparant avec avec les réponses pour le standard others ;
- Les standards que nous n'avons pas identifiés ou indument retirés lors de la préparation disciplinaire du questionnaire : zone chaude dans la modalité other. à comparer avec les réponses dans la ligne not familiar at all) ;
- Les standards qui entrent dans la zone d'intérêt du groupe (zone de confiance pour nous) : zones pastel dans dans little familiar et not familiar maybe interesting.

Annex 3 – Blank questionnaire

20191104 (SSA) - SV Tools & Standards

Metadata standards and tools practice at EPFL School of Life Sciences

Isa-Tab, DICOM, DDI, Uniprot, Genbank... are some well-known metadata standards and tools in the field of Life Sciences. Do you use them?

This EPFL Library survey is aimed at EPFL Life Sciences faculty and labs staff to ascertain community practices around metadata standards and tools. It is a follow-up survey on the one conducted in 2019 entirely around tools.

The survey should not take you more than 5 minutes to complete and is open until 29.02.2020. You can share your contact details il you wish to be informed about the study next steps. Answers will be anonymized for the analysis.

If you have any questions or comment about the survey please email me at eliane.blumer@epfl.ch.

Which EPFL Life Sciences research group are you the closest with? *

Please choose...

Best known standards

How familiar are you with the following standards? (Institute of Bioengineering respondents)

By "standard" or "metadata standard", we mean: vocabularies, terminologies, data formats, data models and schemas, annotations formats, ontologies...

Just skip the lines you don't feel like answering.

	Very familiar (daily practice)	Quite familiar (occasional practice)	Little familiar (no practice)	Not familiar (never heard about it, but maybe interesting)	Not familiar at all (never heard about it, not applicable)
Biological Pathway Exchange (BioPAX)	○	○	○	○	○
Cell ontology	○	○	○	○	○
CellML metadata	○	○	○	○	○
chEBI (Chemical Entities of Biological Interest)	○	○	○	○	○
DICOM	○	○	○	○	○
ENCODE (Encyclopedia of DNA Elements)	○	○	○	○	○
ensembl	○	○	○	○	○
FAANG (Functional Annotation of ANimal Genomes)	○	○	○	○	○
FASTA, FASTQ	○	○	○	○	○
Genbank sequence format	○	○	○	○	○
Gene Ontology	○	○	○	○	○
Genome metadata from PATRIC (bacterial Bioinformatics Resource Center)	○	○	○	○	○
GeoME (Genomic Observatories MetaDatabase)	○	○	○	○	○

GFF3 (General Feature Format)	○	○	○	○	○
ISA-tab	○	○	○	○	○
LINCS data (Library of Integrated Network-Based Cellular Signature)	○	○	○	○	○
Metabolomics Standards Initiative (MSI) and Core Information for Metabolomics Reporting (CIMR)	○	○	○	○	○
MIACME (Minimum Information About Cell Migration Experiments)	○	○	○	○	○
MIAPE (Minimum information about a proteomics experiment)	○	○	○	○	○
MIMARKS (Minimum information about a marker gene sequence)	○	○	○	○	○
MIMIx (Minimum Information about a Molecular Interaction eXperiment)	○	○	○	○	○
MINSEQE (Minimum Information about a high-throughput SEQuencing Experiment)	○	○	○	○	○
MIRIAM (Minimal Information Required In the Annotation of Models)	○	○	○	○	○
MITAB (PSI-MI TAB format)	○	○	○	○	○
MIxS MIGS/MIMS	○	○	○	○	○

MSI (Metabolomics standard Initiative)	◯	◯	◯	◯	◯
NCBI Taxon	◯	◯	◯	◯	◯
NIH Common data elements	◯	◯	◯	◯	◯
Observ-OM and Observ-TAB	◯	◯	◯	◯	◯
OME (Open Microscopy Environment Data model, OME-XML, OME-TIFF)	◯	◯	◯	◯	◯
PDBx/mmCIF Dictionary (Protein Data Bank, Crystallographic Information File)	◯	◯	◯	◯	◯
Protocol Data Element Definition	◯	◯	◯	◯	◯
Sequence Ontology (SO)	◯	◯	◯	◯	◯
Other standard(s)	◯	◯	◯	◯	◯

If you ticked "Other standard(s)", please specify which one(s)

How familiar are you with the following standards? (Swiss Institute for Experimental Cancer Research respondents)

By "standard" or "metadata standard", we mean: vocabularies, terminologies, data formats, data models and schemas, annotations formats, ontologies...

Just skip the lines you don't feel like answering.

	Very familiar (daily practice)	Quite familiar (occasional practice)	Little familiar (no practice)	Not familiar (never heard about it, but maybe interesting)	Not familiar at all (never heard about it, not applicable)
Biological Pathway Exchange (BioPAX)	◯	◯	◯	◯	◯

CellML metadata	○	○	○	○	○
ODM, ODM-XML (CDISC Clinical data interchange standards consortium CDISC operational data model)	○	○	○	○	○
DICOM	○	○	○	○	○
ISA-tab	○	○	○	○	○
LINCS data (Library of Integrated Network-Based Cellular Signature)	○	○	○	○	○
Metabolomics Standards Initiative (MSI) and Core Information for Metabolomics Reporting (CIMR)	○	○	○	○	○
MIACME (Minimum Information About Cell Migration Experiments)	○	○	○	○	○
MIRIAM (Minimal Information Required In the Annotation of Models)	○	○	○	○	○
MITAB (PSI-MI TAB format)	○	○	○	○	○
NIH Common data elements	○	○	○	○	○
OME (Open Microscopy Environment Data model, OME-XML, OME-TIFF)	○	○	○	○	○
PDBx/mmCIF Dictionary (Protein Data Bank, Crystallographic Information File)	○	○	○	○	○

Protocol Data Element Definition	◯	◯	◯	◯	◯
Sequence Ontology (SO)	◯	◯	◯	◯	◯
Other standard(s)	◯	◯	◯	◯	◯

If you ticked "Other standard(s)", please specify which one(s)

How familiar are you with the following standards? (Global Health Institute respondents)

By "standard" or "metadata standard", we mean: vocabularies, terminologies, data formats, data models and schemas, annotations formats, ontologies...

Just skip the lines you don't feel like answering.

	Very familiar (daily practice)	Quite familiar (occasional practice)	Little familiar (no practice)	Not familiar (never heard about it, but maybe interesting)	Not familiar at all (never heard about it, not applicable)
Biological Pathway Exchange (BioPAX)	◯	◯	◯	◯	◯
Biotop	◯	◯	◯	◯	◯
CellML metadata	◯	◯	◯	◯	◯
ODM, ODM-XML (CDISC Clinical data interchange standards consortium CDISC operational data model)	◯	◯	◯	◯	◯
COMPARE data hubs standards checklists	◯	◯	◯	◯	◯
DDI (Data Documentation Initiative)	◯	◯	◯	◯	◯
DICOM	◯	◯	◯	◯	◯

GSCID BRC project and sample application standard	○	○	○	○	○
ISA-tab	○	○	○	○	○
Medred ontology	○	○	○	○	○
MIACME (Minimum Information About Cell Migration Experiments)	○	○	○	○	○
MIBBI (Minimum Information for Biological and Biomedical Investigations)	○	○	○	○	○
MIRIAM (Minimal Information Required In the Annotation of Models)	○	○	○	○	○
MITAB (PSI-MI TAB format)	○	○	○	○	○
NIH Common data elements	○	○	○	○	○
OME (Open Microscopy Environment Data model, OME-XML, OME-TIFF)	○	○	○	○	○
PDBx/mmCIF Dictionary (Protein Data Bank, Crystallographic Information File)	○	○	○	○	○
Protocol Data Element Definition	○	○	○	○	○
Sequence Ontology (SO)	○	○	○	○	○
SNOMED-CT	○	○	○	○	○
Other standard(s)	○	○	○	○	○

If you ticked "Other standard(s)", please specify which one(s)

How familiar are you with the following standards? (Neuroscience — Brain Mind Institute respondents)

By "standard" or "metadata standard", we mean: vocabularies, terminologies, data formats, data models and schemas, annotations formats, ontologies...

Just skip the lines you don't feel like answering.

	Very familiar (daily practice)	Quite familiar (occasional practice)	Little familiar (no practice)	Not familiar (never heard about it, but maybe interesting)	Not familiar at all (never heard about it, not applicable)
Biological Pathway Exchange (BioPAX)	○	○	○	○	○
CellML metadata	○	○	○	○	○
DICOM	○	○	○	○	○
ISA-tab	○	○	○	○	○
MIACME (Minimum Information About Cell Migration Experiments)	○	○	○	○	○
MIRIAM (Minimal Information Required In the Annotation of Models)	○	○	○	○	○
MITAB (PSI-MI TAB format)	○	○	○	○	○
NIH Common data elements	○	○	○	○	○
NINDS Common Data Elements	○	○	○	○	○
OME (Open Microscopy Environment Data model, OME-XML, OME-TIFF)	○	○	○	○	○

PDBx/mmCIF Dictionary (Protein Data Bank, Crystallographic Information File)	◯	◯	◯	◯	◯
Protocol Data Element Definition	◯	◯	◯	◯	◯
Sequence Ontology (SO)	◯	◯	◯	◯	◯
Other standard(s)	◯	◯	◯	◯	◯

If you ticked "Other standard(s)", please specify which one(s)

How familiar are you with the following standards? (Center for Neuroprosthetics respondents)

By "standard" or "metadata standard", we mean: vocabularies, terminologies, data formats, data models and schemas, annotations formats, ontologies...

Just skip the lines you don't feel like answering.

	Very familiar (daily practice)	Quite familiar (occasional practice)	Little familiar (no practice)	Not familiar (never heard about it, but maybe interesting)	Not familiar at all (never heard about it, not applicable)
Biological Pathway Exchange (BioPAX)	◯	◯	◯	◯	◯
CellML metadata	◯	◯	◯	◯	◯
DICOM	◯	◯	◯	◯	◯
ISA-tab	◯	◯	◯	◯	◯
MIACME (Minimum Information About Cell Migration Experiments)	◯	◯	◯	◯	◯
MIRIAM (Minimal Information Required In the Annotation of Models)	◯	◯	◯	◯	◯

MITAB (PSI-MI TAB format)	○	○	○	○	○
NIH Common data elements	○	○	○	○	○
NINDS Common Data Elements	○	○	○	○	○
OME (Open Microscopy Environment Data model, OME-XML, OME-TIFF)	○	○	○	○	○
PDBx/mmCIF Dictionary (Protein Data Bank, Crystallographic Information File)	○	○	○	○	○
Protocol Data Element Definition	○	○	○	○	○
Sequence Ontology (SO)	○	○	○	○	○
Other standard(s)	○	○	○	○	○

If you ticked "Other standard(s)", please specify which one(s)

How familiar are you with the following standards? (Blue Brain Project respondents)

By "standard" or "metadata standard", we mean: vocabularies, terminologies, data formats, data models and schemas, annotations formats, ontologies...

Just skip the lines you don't feel like answering.

	Very familiar (daily practice)	Quite familiar (occasional practice)	Little familiar (no practice)	Not familiar (never heard about it, but maybe interesting)	Not familiar at all (never heard about it, not applicable)
Biological Pathway Exchange (BioPAX)	○	○	○	○	○
CellML metadata	○	○	○	○	○

DICOM	○	○	○	○	○
ISA-tab	○	○	○	○	○
MIACME (Minimum Information About Cell Migration Experiments)	○	○	○	○	○
MIRIAM (Minimal Information Required In the Annotation of Models)	○	○	○	○	○
MITAB (PSI-MI TAB format)	○	○	○	○	○
NIH Common data elements	○	○	○	○	○
NINDS Common Data Elements	○	○	○	○	○
OME (Open Microscopy Environment Data model, OME-XML, OME-TIFF)	○	○	○	○	○
PDBx/mmCIF Dictionary (Protein Data Bank, Crystallographic Information File)	○	○	○	○	○
Protocol Data Element Definition	○	○	○	○	○
Sequence Ontology (SO)	○	○	○	○	○
Other standard(s)	○	○	○	○	○

If you ticked "Other standard(s)", please specify which one(s)

Favourite tools

How familiar are you with the following tools families?

By "tool" we mean : scientific software, facility, platform, reference database...

Just skip the lines you don't feel like answering.

	Very familiar (daily practice)	Quite familiar (occasional practice)	Little familiar (no practice)	Not familiar (never heard about it, but maybe interesting)	Not familiar at all (never heard about it, not applicable)
Molecular Interactions and protein-protein tools (such as PathBLAST or HADDOCK)	○	○	○	○	○
Structural bioninformatics tools (such as Swisslipids or Swissdock)	○	○	○	○	○
Survey, database and data management tools (i.e. RedCap, SLIMS)	○	○	○	○	○
Public protein data repository (i.e. PRIDE Archive)	○	○	○	○	○
Gene ontology annotation (i.e. GOA)	○	○	○	○	○
Referencing of data (i.e. Identifiers.org, RRID portal)	○	○	○	○	○
Protein Annotation (i.e. HAMAP)	○	○	○	○	○
Primary nucleotide sequence databases (e.g. GenBank, European Nucleotide Archive)	○	○	○	○	○
Genome databases (e.g. Ensembl, UCSC Genome Browser)	○	○	○	○	○

Gene expression/Microarray databases (e.g. ArrayExpress, Gene Expression Omnibus)	○	○	○	○	○
Protein Sequence databases (e.g. Uniprot, RefSeq, InterPro, PROSITE)	○	○	○	○	○
Protein-protein and other molecular interaction (e.g. Intact, String)	○	○	○	○	○
Protein structure/Model databases (e.g. PDB, SWISS-MODEL)	○	○	○	○	○
Signal transduction/Metabolic pathway databases (e.g. Reactome, KEGG)	○	○	○	○	○
Mutation databases (e.g. OMIM, HGMD, dbSNP)	○	○	○	○	○
Model organism databases (e.g. WormBase, Mouse Genome Informatics)	○	○	○	○	○
Other tool(s)	○	○	○	○	○

If you ticked "Other tool(s)", please specify which one(s)

What about you?

Did you ever contribute to a metadata standard effort or a tool effort? *

By "effort" we mean : design, development, promotion, teaching...

- ◯ No
- ◯ Yes (please specify)

In your opinion, what is using metadata standards or tools good for? *

You can select multiple options.

- ☐ Self-efficiency (neat and tidy data)
- ☐ Collaboration (compatible, shareable data)
- ☐ Complying to research community rules or industry rules
- ☐ I never thought much about it
- ☐ Other (please specify)

Almost done...

Do you have any comment, any suggestion?

Share your thoughts about tools and metadata standards and what support you expect from the Library.

Would you like to be informed about the next steps after this survey (additional interviews, results)? If yes, please share your mail.

name.surname@epfl.ch

Don't forget to click on Finish!