Extending the CMDI Universe: Metadata for Bioinformatics Data

Olaf Brandt, Holger Gauza, Steve Kaminski,
Mario Trojan, Thorsten Trippel, Johannes Werner
Eberhard Karls Universität
Tübingen, Germany
firstname.lastname@uni-tuebingen.de

Abstract

The Component Metadata Infrastructure (CMDI) is a discipline independent metadata framework, though it is currently mainly used within CLARIN and by initiatives in the humanities and social sciences. CMDI allows flexible modelling of metadata schemas that are adjusted to the type of data. The model has built in functionality for semantic interoperability based on inventories providing persistent identifiers for definitions. In this paper we investigate, if and how CMDI can be used in bioinformatics for metadata modelling and describing the research data. For this purpose we embed CMDI based metadata in METS containers. Two sample schemas are developed the first based on a bottom up process and the second one takes the requirements of data publishing portals as the starting point of development.

1 Introduction

Data management in bioinformatics projects requires a very diverse and flexible set of metadata to accommodate for different scientific, organisational, and technical needs. Data categories must provide for the workflows of various types of experiments in the field of OMICS research (gen*omics*, prote*omics* etc.), including workflows of researchers, laboratories archives, public repositories[1] and third party suppliers such as sequencing labs. Most laboratory working groups use individual, table based metadata for their projects, which are neither semantically described, nor interoperable with established workflows in data archival or data analysis. Within the project *BioDATEN*[2] funded by the state of Baden-Württemberg in Germany, subject matter experts meet to develop an environment that facilitates data storage and collaboration of different bioinformatics working groups and archives. BioDATEN combines expertise in data management, archiving, library science, bioinformatics and related scientific workflows.

Part of ensuring the interoperability and semantic interpretation of metadata is the discussion of a common description of metadata. Though there are specific metadata schemata in the bioinformatic community like the PRIDE schema for proteomics and approaches like qPortal[3] there is no recognized gold standard for metadata handling in this subfield of OMICS research let alone the broader field of bioinformatics. On the other hand, there are well established standards outside bioinformatics that are used in the archiving and library community, such as METS/MODS[4], PREMIS[5], MARC 21[6], etc. The variety of research data, research questions, methods and workflows require additional flexible and research specific schemata, that can be adjusted to the needs of the concrete projects' and working groups' context. Here, the ISO standardised ISO 24622-1 and -2 XML based CMDI framework is going to be explored as a candidate for representing the metadata in this project.

2 Motivation

Collaboration in bioinformatics is becoming increasingly important, by sharing information about genetic sequencing and data for reproducing results, applying different algorithms and workflows. Sharing primary data

[1] A prominent example of a public repository is the *National Center for Biotechnology Information* (NCBI, USA, https://www.ncbi.nlm.nih.gov/.

[2] http://www.biodaten.info

[3] Mohr et al. 2018.

[4] for example <METS>Version 1.6.

[5] Caplan 2009 and http://www.loc.gov/standards/premis/.

[6] https://www.loc.gov/marc/.

Olaf Brandt, Holger Gauza, Steve Kaminski, Mario Trojan, Thorsten Trippel and Johannes Werner 2021. Extending the CMDI Universe: Metadata for Bioinformatics Data. *Selected papers from the CLARIN Annual Conference 2020*. Linköping Electronic Conference Proceedings 180: 180 47–54.

becomes more and more widespread within the last 20+ years, but is still comparatively new to the field. This is contrasted by the fact that prices for genetic sequencing are constantly dropping, resulting in an exponential growth of data production and availability. Due to the increased amount of data, no single data centre is and will be able to store and provide access to all data, even if repositories specialize for species, etc. This results in the need for a distributed infrastructure in which research data is provided in a FAIR[7] way.

Distributed environments providing data require a clear idea of the required levels of descriptions of research data. In the context of the European infrastructure initiative CLARIN, this has a long tradition with metadata being available and searchable with tools such as the Virtual Language Observatory[8], though CLARIN addresses primarily research in the humanities and social sciences. In bioinformatics, similar methods and tools have been developed[9], accompanied by strong market influences of large archives and publishers. In this paper, we try to elaborate on the technology used within CLARIN to see if the methods applied there are applicable for the BioDATEN project in bioinformatics.

2.1 Structured documentation of research data

The idea of sharing research data implies the distributed nature of research. Often more than one working group is interested in specific research questions to be addressed with the help of specific data sets. The diversity of research questions, size of groups, and distribution of interested parties results in the need for detailed descriptions that are necessary to understand the data. The internal documentation of each group such as code books, Read-Me files, laboratory books etc. are part of this documentation. This is not only true for labs working in natural language processing (NLP), but independent of the research discipline.

Within the BioDATEN project, various partners provided samples of metadata they use in their respecive laboratories and for publication purposes. In bioinformatics, we noticed that there is metadata documentation available. However, to our knowledge there is no established and bioinformatics-wide schema. There are attempts to use representations based on `schema.org` within the `bioschemas.org` project, but these are not sufficient: First, there are no profiles provided that fit OMICS data. Second, those initiatives provide a number of data category definitions that could be utilized, but the categories extending schema.org do not have an identifier that can be used for references and it is unclear if they are stable enough for long time use. For example, the Gene Profile[10] is targeted at life sciences, including diseases, and omits processing information, while the Biosample type [11] does not provide identifiers for some data categories that could be used. However, where possible, the concepts used by bioschemas.org will be reused here as well.

As a science data centre (SDC) of the state of Baden-Württemberg in Germany, the BioDATEN project also has ties to other initiatives in the field, which includes contacts to ELIXIR, the German Network for Bioinformatics infrastructure (De.NBI), etc. Consequently, in the development of metadata schemas we monitor the developments, planning for interfaces between these state infrastructures and larger initiatives. This also includes the use of existing ontologies where they fit to the needs of the OMICS community, avoiding duplication of work. However, to date ontologies of these communities are - as for the bioschemas.org representations - not sufficient for the OMICS community that uses the services of BioDATEN.

The metadata provided by project partners can be subdivided according to (1) their descriptive function, (2) specific information for the community, (3) process oriented information, and (4) technical metadata. Descriptive metadata is used to describe the data in an archive for citation purposes, such as the DataCite standard[12], but some data categories are not applicable in the context of bioinformatics, e.g. the concept of *Author* of raw DNA sequencing data. Process and workflow-oriented information[13] provides the background and origin of data, as well as information about the tools and experimental techniques that have been used to generate the research data[14]. Technical information contains file information often provided in terms of PREMIS. Community specific information is often provided to allow specific keywords and structures in the search process.

When investigating the sample metadata provided by project partners, it became obvious that existing ontologies and taxonomies are often not applied in the concrete laboratory situation, where ad hoc or laboratory

[7] e.g. Wilkinson et al. 2016.
[8] van Uytvanck, Stehouwer and Lampen 2012.
[9] e.g. BLAST `https://blast.ncbi.nlm.nih.gov/Blast.cgi`
[10] `https://bioschemas.org/profiles/Gene/0.7-RELEASE/`
[11] `https://bioschemas.org/types/BioSample/0.1-RELEASE-2019_06_19/`
[12] *DataCite* 2019.
[13] for example De Nies 2013
[14] for an example of the workflow with its documentation, see Mohr et al. 2018.

specific spreadsheets are used to identify and document the research data. The lack of standard metadata formats thwarts the exchange and searchability of data with tools.

The interoperability of individual laboratory data with existing portals for sharing research data thus requires tailored solutions for each lab and portal despite the fact that labs often use similar testing machinery and testing procedures. Besides the rather informal definition of metadata, the public data repositories are often used for enhancing publications and sharing data. Data publication becomes even more imminent as publication of research data is mostly required for funding purposes and publication of scientific results. This is very similar to disciplines in the humanities such as (linguistic) annotation in fieldwork, corpus linguistics, etc.

2.2 The unit of description: the granularity of the research data

Infrastructures in the humanities and for OMICS research both show a great variety. The variety in the infrastructures for humanities results from different data types ranging from lexical resources, corpora, to data matrices. These data sets are serialized according to various conceptual models, such as graphs in RDF, XML annotations, table structures, etc.

In contrast to that, OMICS research operates on data sets that structurally have much more in common. However, the variety is still huge with for example different species, recurrence of analysis, size of cohorts, geographic variation, number of cells investigated, etc. Besides this variety, there are different workflows, also influenced by laboratory hardware. As the diverse data sets in the humanities require a flexible metadata model, the same applies to the area of bioinformatics. The archiving of research data, as well as searching and retrieving data units relies on descriptions by means of metadata and the assignment of a persistent reference to these units. For this reason, it is essential to have a solid understanding of the data unit to be described, usually termed the *granularity* of data. Granularity in this sense is the unit of data to be stored, archived and referenced in the research process.

ISO 24619 recommends on the granularity to use existing granularities, complete files, resource autonomy, and the requirement for a unit to be citable as criteria for selecting the underlying unit. This standard has been applied in CLARIN for assigning PIDs. In bioinformatics, there are some obvious candidates for archival objects. Inherent *atomic* units could be a base pair of nucleobases, a gene, a chromosome or an entire genome of an individual. From a computer science perspective, it could also be a single data file that is created in the process, such as FASTQ, FASTA, BAM, VCF, Excel or CSV files. Another natural unit would be a package of all files in an experiment, or all files that relate to a publication.

For bioinformatics applications it turned out that the granularity is implicitly given by the *sample*, i.e. the unit of a physically extracted sample of material, for example drawn with a needle. In bioinformatics workflows, these samples do only occur initially, afterwards other units will be referred to, such as sequencing information or experiments. For archiving, the sample often remains the common unit, but sometimes multiple samples are packaged into a study. It is noteworthy that raw data produced by a sequencing lab (DNA, RNA etc.) is nearly always transformed, trimmed, cleaned etc. This pre-processing is necessary to allow deeper analysis. The pre-processing is very similar to the processing and selection of corpus data in the humanities.

2.3 Automatic metadata extraction requirements during a data creation workflow

Metadata creation is often seen as a burden for researchers creating data. Due to the lack of standardised processes and project management software, archiving metadata is often created manually, based for example on the headers of TEI files[15], or partly automatized by language processing applications and workflow engines such as WebLICHT[16]. The quality and completeness of the metadata in the archiving process is a major issue, for which automatic metadata enrichment processes are seen as a major step forward. This could mean to enrich metadata by authority file references, keyword extraction from textual resources, technical information extraction such as file size, checksums, dates, etc.

In bioinformatics processes, samples are analysed and processed in complex workflows. Many of these workflows are run on high performance clusters (HPC) or cloud infrastructure, are automatized and require only little intervention, hence the manual creation of metadata is even more problematic. The creation of metadata, especially of process and technical metadata, can partly be automatized, as the workflow engines on the infrastructure use, collect and provide process information during the process. Additionally, the technical metadata can be generated easily with appropriate software tools. Larger parts of the descriptive

[15] *TEI P5* 2020.
[16] M. Hinrichs, Zastrow and E. Hinrichs 2010.

and community specific information tend to be very similar in specialized labs, working with specific species, controlled conditions, health environments, etc. These can partly be defined in templates to be post-edited by the researcher. Again, this is similar to fieldwork situation or within large annotation projects in the humanities, though here this is often a manual process. For large NLP tasks, the metadata related processes are comparable to the bioinformatics workflows.

3 Specification and serialization options

In bioinformatics, researchers have to adhere to requirements by sequencing labs and scientific publishers. For sequencing labs, metadata descriptions contain details about the arrangement and preparation of samples to attribute the reads to the samples, treatments etc. The information may be lost in the resulting raw DNA sequences. Researchers have to define their lab processes to ensure that the DNA sequences are attributed to the sample, and in turn to the treatment or experimental condition. For publishing articles, the publication of data sets is often a requirement, the publication portals requiring various bits of information about the underlying sample. Hence a metadata schema needs to cater for the third party and laboratory internal requirements. To avoid redundancy in the metadata, the metadata categories need to be mapped onto each other, identifying common concepts and allowing transformation.

BioDATEN interdisciplinarily explores options for serializing the metadata with the full flexibility of metadata schemas required. Options using different data models such as RDF are left out. However, converting the metadata into RDF and offering it, possibly enriched by ontologies and authority data is seen as a valid option for the integration into the linked data cloud. In the following we discuss PREMIS, METS and CMDI serialization.

3.1 PREMIS

Implemented and used by archives and libraries, the PREMIS standard[17] is meant to support the long-term preservation of digital objects via metadata. In the BioDATEN project, PREMIS will be primarily utilized for the storage of technical and rights metadata, as well as for the recording of events like data format conversion, checksum validation or changes in the related metadata records. The PREMIS data dictionary offers comprehensively controlled vocabularies allowing pointers with persistent identifiers. The description of scientific workflows denotes a clear limitation of the PREMIS standard. Hence PREMIS will be used for interoperability, but alone it is not sufficient for meeting all requirements.

3.2 METS

In order to manage the different metadata schemas used to describe research data, it is useful to collect them in a container format. Having multiple metadata records for one digital object should be avoided. One solution would be to use a container format such as the XML based METS format to combine different schemas. METS is described by an XML schema and is almost exclusively serialized as XML. As a container format it is able to integrate other XML schemas without loss of information via so called extension schemas. A decisive reason to choose METS is the integration with PREMIS, which is described in detail in the literature. The different building blocks offered by the METS standard can be used to store the variety of metadata schemas needed for research data. These schemas can be registered in METS profiles[18], which also allow for a comprehensive documentation and therefore re-usability of metadata in the METS container format. However, as a container format, METS does not provide the required metadata schemas in itself.

3.3 CMDI

Another option for modelling the metadata is by using the Component Metadata Infrastructure (CMDI, ISO 24622-1 and ISO 24622-2), which is applied in the CLARIN community. As an XML based serialization, many tools for editing and maintaining exist, archival systems implement ways of storing the data. Transformation into other XML based formats is easily conducted using XSLT or similar technologies. CMDI offers flexible modelling options. For example, each lab can create their own metadata profile, assembling all necessary data categories required in their respective workflow. At the same time, they can reuse parts of the metadata profiles that match the requirements of portals and service providers, archives and other partners. Using these common components, the target data format can easily be generated by a simple transformation. In

[17] for example Caplan 2009
[18] *METS Profiles* 2018

fact, due to the definition of the CMDI components with concept links, for example, referring to definitions in the CLARIN Concept registry or in persistent ontologies, a high degree of semantic interoperability is achieved. For the serialization, the envisaged problems are similar to those already known in the CLARIN community: if labs define their own profiles and components, a certain degree of fragmentation is bound to be the result. Additionally, there is currently no fixed set of data categories, and the CLARIN Concept Registry does not contain the required definitions for bioinformatics data sets beyond interdisciplinary metadata categories. Another potential problem is that neither the bioinformatics archives and portals nor the external service providers natively support CMDI, hence a transformation is required at each step in the workflow, if CMDI were used. However, metadata generated in an automatic workflow can successively be added to the metadata file, which supports the required flexiblity.

4 Implementation

Bearing in mind the established metadata workflows of the data centres and publishers based on METS, we envision the integration of CMDI in METS-containers, also integrating PREMIS metadata. Based on previous work within CLARIN we created two CMDI Profiles, the BioDATEN Profile (clarin.eu:cr1:p_1588142628378)[19] and the BioDATEN Minimal Profile (clarin.eu:cr1:p_1610707853515)[20]. Figure 1 shows the preliminary integration of the mentioned schemes in the form of a simplified section of a METS-XML file. [21]

Using this procedure, we can combine CMDI's flexiblity for modelling while keeping the data interoperable with the archives and service providers.

4.1 The BioDATEN profile

The CMDI profile reuses components previously defined in their newest developmental version, especially

- the GeneralInfo component for general information, which is Dublin Core inspired
- the optional Project component for information on the project
- the optional Publications component to provide information on associated publications
- the Creation component with information on the creation of the resource. This component was enriched by a new ethics component providing information on obligations by ethics commissions, etc. As this also becomes more relevant for other disciplines, this should be a general recommendation for future releases of the creation component.
- the optional Documentations component for available documentation that is not part of the publications
- the Access component to provide information on accessing the resource
- the ResourceProxyListInfo component providing information on each data stream, including checksums, size, and original file name.

The tailored component SequencingInfo provides specific information on OMICS data beyond the creation process. For selecting data categories here, we were able to use Excel files used for managing metadata and provided by some partner laboratories. The schema is defined with its extension in mind, especially during a consolidation phase in which the community tests the schema. By planning for the extension, it is possible to add fields requested by researchers. The intention was to provide a bottom-up design of a metadata profile.

Currently, we evaluate the mapping of laboratory internal metadata storage to this profile and assess if the integration of this metadata framework, including CMDI, METS and PREMIS in the Invenio[22] repository system, used within the BioDATEN project. However, to our knowledge even open repository systems such as Invenio or Fedora-Commons require additional work when used with tailored metadata schemas. Additionally, it is still essential to investigate, which transformations are required from a lab internal metadata set to interoperable metadata sets used by archives and portals.

The selection of data categories was bottom-up, starting with the partner laboratories and researchers of the project. The schema turned out to be very detailed with 99 fields, most of them optional. The interaction with public repository platforms prooved to be problematic, as they required other metadata fields. Additionally

[19] https://catalog.clarin.eu/ds/ComponentRegistry/rest/registry/1.x/profiles/clarin.eu:cr1:p_1588142628378/xsd

[20] https://catalog.clarin.eu/ds/ComponentRegistry/rest/registry/1.x/profiles/clarin.eu:cr1:p_1610707853515/xsd

[21] The complete sample metadata file and the schemas discussed in this paper have the DOI 10.5281/zenodo.4506354 and can be viewed and downloaded from the following url: https://doi.org/10.5281/zenodo.4506354.

[22] https://github.com/inveniosoftware/invenio

```
<mets:mets>
    <mets:metsHdr ID="e6879deaa-2f64-48d7-bfc9-21cd77fb9571"
        CREATEDATE="2020-08-13T11:28:51"
        LASTMODDATE="2020-08-13T11:28:51" RECORDSTATUS="NEW">
        <mets:metsDocumentID TYPE="UUID">
            e6879deaa-2f64-48d7-bfc9-21cd77fb9571</mets:metsDocumentID>
    </mets:metsHdr>
    <mets:dmdSec
        ID="dmdSecGeneralDataCite_6879deaa-2f64-48d7-bfc9-21cd77fb9571">
        <mets:mdWrap MDTYPE="OTHER">
            <mets:xmlData>
                <cmd:CMD CMDVersion="1.2">
                    ...
                  <cmdp:BioDatenProfile>
                     <cmdp:GeneralInfo>
                       <cmdp:ResourceName xml:lang="en">Sample data set</cmdp:ResourceName>
                       <cmdp:ResourceTitle xml:lang="en"/>
                       <cmdp:ResourceClass>OMICS data</cmdp:ResourceClass>
                       <cmdp:Version xml:lang="en"/>
                       <cmdp:LifeCycleStatus>archived</cmdp:LifeCycleStatus>
                       <cmdp:dateCreated>2020-08-08</cmdp:dateCreated>
                       <cmdp:LegalOwner xml:lang="en"/>
                       <cmdp:FieldOfResearch>Bioinformatics</cmdp:FieldOfResearch>
                               ...
                     </cmdp:GeneralInfo>
                            ...
                  </cmdp:BioDatenProfile>
                   ...
                </cmd:CMD>
            </mets:xmlData>
        </mets:mdWrap>
    </mets:dmdSec>
   ...
</mets:mets>
```

Figure 1: Simplified extract of the preliminary integration of the schemes CMDI and PREMIS into a METS-XML container.

by reusing preexisting components such as the ones for general information, the resulting metadata showed redundancies between the CMDI section and other parts of the metadata in the METS container.

4.2 The BioDATEN Minimal Profile

Based on the BioDATEN-Profile in conjunction with a survey of various archives and portals targeting OMICS data we created a second profile. In order to cover the variety of the OMICS research on the one hand and the demand for a minimal set of metadata to be supplied by the scientists on the other hand we based our profile on the *Minimum Information about any (x) Sequence* (MIxS) schema.[23]. At the same time we concluded that a perspective focused solely on samples (and environments) is too narrow and classified the subject-specific metadata in groups named study, experiment, sample, environment, run, and data. This process approach has been inspired mainly by the Extracellular RNA Atlas[24] and the MOD-CO schema[25]. The new schema comprises the absolutely required fields for data portals plus additional fields suggested by expert users, resulting in 21 OMICS specific fields. We expect to see a demand for more metadata fields beyond our minimal set. In order to guide this in a manageable way we plan to offer many optional fields from the MIxS schema, currently the schema offers about 30 of these. Furthermore additional information can be added without structural restrictions. By this procedure we hope to collect a feedback about necessary metadata apart from the minimal set. The current implementation does not implement the full set of vocabularies that are intended to be used in the schema. The integration of the vocabularies based on various ontologies is still pending.

For review of the schema, an HTML based input form was generated, using the Comedi editor[26]. Unfortunately, this editor still implements CMDI 1.1; integration into METS containers in our case requires CMDI 1.2. Different editor generation tools based on the XSchema are still being tested such as the open source tool XSD2HTML2XML[27], but this tool is of limited use for schemas embedded in containers and shows some usability issues when using controlled vocabularies in the schema.

5 Future Work

Automatic metadata retrieval from bioinformatics workflows still remain challenging. Several approaches have already been developed, however a standard independent of specific research disciplines is not yet existent. This requires APIs in the workflow engines to extract the appropriate metadata in the process where they are present. The adaptation of the enriched metadata to specific modelling environments such as CMDI would result from this.

Based on the automatic generation of an HTML form to edit the metadata instances, we plan to test the schema with researchers from OMICS research and test, if the information can be fed into data publication portals with standard APIs.

Within BioDATEN, the development points in a direction of not only supporting one metadata schema, but a variety of schemas that fit to the needs of the users in the specific domains. Initially, BioDATEN will support a limited number of schemas providing appropriate converters to DataCite, Dublin Core and other relevant formalisms. Users of the BioDATEN infrastructure will be required to restrict themselves to these schemas at first, but an extension to a general framework for registering metadata schemas might be required for scalability reasons. This is very similar to the development in CLARIN and the CMDI infrastructure with the component registry and a tool for mapping the data categories to a facetted search comparable to the VLO. However, as various schemas, also outside of the CMDI universe might be required, a generalized framework might be required that allows also the registration of other schemas, including schemas that are not defined for XML. Due to the support within the tools of BioDATEN, a proliferation can still be avoided, as no unsupervised schema development is part of the process. The modelling and implementation of this will be part of future developments.

Acknowledgements

The work reported here was funded by the Ministry of Science, Research and Art of the Federal State of Baden-Württemberg (MWK). We would also like to thank the anonymous reviewers for helpful comments.

[23] https://gensc.org/mixs/
[24] http://exrna-atlas.org
[25] https://www.mod-co.net
[26] Lyse, Meurer and De Smedt 2015
[27] http://www.linguadata.nl/ and https://github.com/MichielCM/xsd2html2xml

References

Priscilla Caplan. 2009. Understanding premis.

DataCite. 2019. DataCite Metadata Schema 4.3, 08.

Tom De Nies. 2013. Constraints of the prov data model. W3C Recommendation.

Marie Hinrichs, Thomas Zastrow, and Erhard Hinrichs. 2010. Weblicht: Web-based lrt services in a distributed escience infrastructure. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odijk, Stelios Piperidis, Mike Rosner, and Daniel Tapias, editors, *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta, 05. European Language Resources Association (ELRA).

ISO 24619:2011(E). 2011. Language resource management — Persistent identification and sustainable access (PISA). Standard, International Organization for Standardization, Geneva, CH, 01.

ISO 24622-1:2015(E). 2015. Language resource management — Component Metadata Infrastructure (CMDI) — Part 1: The Component Metadata Model. Standard, International Organization for Standardization, Geneva, CH.

ISO 24622-2:2019(E). 2019. Language resource management — Component metadata infrasctructure (CMDI) — Part 2: Component metadata specification language. Standard, International Organization for Standardization, Geneva, CH, 07.

Gunn Inger Lyse, Paul Meurer, and Koenraad De Smedt. 2015. COMEDI: A component metadata editor. *Selected Papers from the CLARIN 2014 Conference*, 8(116):82–98, 08.

METS 1.6. 2010. <METS> metadata encoding and transmission standard: Primer and reference manual. version 1.6 revised. Technical report.

METS Profiles. 2018. METS profiles. Technical report.

Christopher Mohr, Andreas Friedrich, David Wojnar, Erhan Kenar, Aydin Can Polatkan, Marius Cosmin Codrea, Stefan Czemmel, Oliver Kohlbacher, and Sven Nahnsen. 2018. qportal: A platform for data-driven biomedical research. *PLOS ONE*, 13(1):1–18, 01.

PRIDE. n.d. Guide to generate PRIDE XML files.

TEI P5. 2020. TEI P5: Guidelines for electronic text encoding and interchange. TEI Recommendation.

Dieter van Uytvanck, Herman Stehouwer, and Lari Lampen. 2012. Semantic metadata mapping in practice: the virtual language observatory. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Mehmet Uğur Doğan, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, pages 1029–1034, Istanbul, Turkey, 05. European Language Resources Association (ELRA).

Mark D. Wilkinson, Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, Jan-Willem Boiten, Luiz Bonino da Silva Santos, Philip E. Bourne, Jildau Bouwman, Anthony J. Brookes, Tim Clark, Mercè Crosas, Ingrid Dillo, Olivier Dumon, Scott Edmunds, Chris T. Evelo, Richard Finkers, Alejandra Gonzalez-Beltran, Alasdair J.G. Gray, Paul Groth, Carole Goble, Jeffrey S. Grethe, Jaap Heringa, Peter A.C 't Hoen, Rob Hooft, Tobias Kuhn, Ruben Kok, Joost Kok, Scott J. Lusher, Maryann E. Martone, Albert Mons, Abel L. Packer, Bengt Persson, Philippe Rocca-Serra, Marco Roos, Rene van Schaik, Susanna-Assunta Sansone, Erik Schultes, Thierry Sengstag, Ted Slater, George Strawn, Morris A. Swertz, Mark Thompson, Johan van der Lei, Erik van Mulligen, Jan Velterop, Andra Waagmeester, Peter Wittenburg, Katherine Wolstencroft, Jun Zhao, and Barend Mons. 2016. The fair guiding principles for scientific data management and stewardship. *Scientific Data*, 3(160018).