////////////////////////////////////////////////////////////////////////////////////////////////////////////////////

**ONTWERPVERSLAG**

FOSB WG Metadata & Standardisation

////////////////////////////////////////////////////////////////////////////////////////////////////////////////////

Datum: 16/09/2020

1. Resume In-depth discussion of the draft metadata model
2. Discuss the use of the Flemish Research Discipline Standard to classify research data sets, discuss archiving and storage terms.

**Aanwezig**

Joke Meeus FWO

Niek Van Wettere VUB

Rudi Baccarne UAntwerpen

Sara De Coster KU Leuven (vervangt Roxanne Wyns)

Charlotte D’Eer FWO

Frederik Coppens VIB

Ils de Bal Departement EWI

Pascale Dengis Departement EWI

Sadia Vancauwenbergh ECOOM-UHasselt

Evy Neyens ECOOM-UHasselt

Dimitri Brosens INBO

Helga Deschrijver IMEC

Francisco Hernandez VLIZ

Katrien Desmet VLHORA

Quentin Groom Plantentuin Meise

Hans Polet ILVO

Rianldo Beck VIB

**Verslag Vergadering 16/09/2020**

Dries Moreels UGent

Alexander Botzki VIB

Steven Heylen FWO

Mia Vanstraelen Brain Group

/////////////////////////////////////////////////////////

Third in-depth discussion of the draft metadata model

**Research discipline**

EWI has asked the members whether it is possible to report research data at level L3 of the Flemish Research Discipline Standard, by analogy with projects. *Research discipline* refers to the classification of the dataset on the basis of the disciplines of the Flemish Research Standard.

Some members propose to make this field recommended instead of mandatory. We should strive to stay as close as possible to the DataCite/OpenAire model. The main reason is to strive for a "single-entry model", where it is sufficient for researchers to deliver the metadata for the dataset only once to a data repository (Zenodo etc.) where the data are archived. In the OpenAire guidelines the metadata field "subject" is labelled as "recommended" (see [*https://guidelines.openaire.eu/en/latest/data/field\_subject.html*](https://guidelines.openaire.eu/en/latest/data/field_subject.html)). Further, FRIS provision of discipline codes for datasets is not mentioned in the BOF decree, and this is not mandatory for publications either.

The ability to deliver discipline codes for datasets to FRIS also depends on the degree to which this information is already available within the institutions and it also depends on technical issues.

It is inconvenient to complement manual input of researchers with an institutional "imputation strategy" where default values are filled in where there is no "authentic value" available. Not all discipline codes of the publication are applicable to the associated datasets, which complicates the imputation strategy. It’s not technically possible to implement a mix of manual input and automated default imputation in some CRIS systems.

Can the FRIS search engine itself apply an imputation strategy to optimize the dataset search results on the FRIS website, e.g. based on links with persons and/or other research outputs if those links are available? At this point FRIS has never added data themselves, data is always provided by the data providers. This depends on the rules that are used within the institutions. At the FRIS level it is impossible to distinguish between disciplines that are added by the researcher and those that are imputed. A key FRIS principle is that EWI trusts the quality of the information that is sent by the providers. Hence, it is up to the institutions to create rules that enable them to impute discipline codes from an organisation to a person or vice versa. EWI can do this, but prefers information that is provided by the institutions. What about datasets that are created by authors that have multiple affiliations? If different institutions deliver info on the same person, all info is just added together for the same person, using ORCID.

EWI agreed it is possible to make this field recommended if providing this information is difficult for the institutions. However, FRIS wants to be part of a roadmap to open science and thus strives to be as open and FAIR as possible, hence, higher quality data is preferred when possible. It also benefits a higher discoverability rate. It was suggested to also keep the user-perspective in mind. How will people search and filter the data on the FRIS-portal? People will look for datasets based on research disciplines and keywords.

Finally, it was decided to make the field ‘Research discipline’ recommended while striving to make it mandatory in the long-run. Meanwhile, researchers who want to add discipline codes to their datasets in institutional repositories should be able to do so. Moreover, a long-term goal must be agreed upon. It can be considered a KPI for discoverability of metadata. It will also depend on how the EU repositories will evolve and which metadata are required there, because researchers will not want to do double work. In most institutions the metadata on datasets is already filled in twice (in IR and other (discipline-specific) repositories such as Zenodo, GBIF, Dryad). It was proposed to automate this as much as possible but how can this be managed when every repository uses slightly different metadata standards? It was agreed to propose a timeline for the implementation of discipline codes for datasets. A preliminary timeline was estimated at 2-3 years.

**Contributing person/organisation:**

Regarding the fields “Contributing person “and “contributing organisation” it was agreed to keep only one “contributor” field, with person types and organisation types. A contributor is defined as: “The institution or person responsible for collecting, managing, distributing, or otherwise contributing to the development of the resource.” It is possible to provide multiple contributors by repeating the property. However, it is only mandatory to provide one contributor type. It’s important to be able to contact someone who can give access to the dataset. Since people will not be available indefinitely, a generic address should also be provided as a fall back.

**Format**

Format is a mandatory field because the BOF-decree determines that it’s mandatory to be able to track how many datasets are delivered in an open format. It was noted that zip-files are often used.

**Open Data Status**

Open data status is a very important field because it’s used as a KPI for open data. An open data label and legitimate opt out will be mandatory for the second quarter of 2021. How are we going to define open data? Should we reduce the definition to “the status on the access possibilities of the dataset.”? Then it refers to whether or not the data are accessible. Access rights are specified using the OpenAire 16. Rights (MA). Within this property you have rightsURI (MA): Use terms from the info:eu-repo-Access-Terms vocabulary. The values are:

info:eu-repo/semantics/closedAccess

info:eu-repo/semantics/embargoedAccess

info:eu-repo/semantics/restrictedAccess

info:eu-repo/semantics/openAccess

If the status is embargoed: should you change the status after the embargo period, or can this be done automatically, for instance by providing the end date of the embargo period? Can the institutions overwrite this information themselves? If this needs to be changed manually, it is very unlikely that this will happen. Is it possible to use OAI protocols to harvest the fields from the IR’s and translate them into Openaire metadata standards?

However, access does not equal reuse. Reuse also requires that the data are downloadable in an open, machine-readable format and accompanied by an explicit open license that enables reuse by anyone without any restrictions.

According to the Open Definition: “Open means anyone can freely access, use, modify, and share for any purpose (subject, at most, to requirements that preserve provenance and openness).” [*https://opendefinition.org/*](https://opendefinition.org/)

When there are licenses specified that restrict commercial reuse such as CC-BY-NC, then the dataset is not open according to the open definition.

Is a dataset only open when it is: 1) openly accessible (immediate download), 2) in an open, machine-readable format, 3) with an open license? It was suggested to make “open data status” a multidimensional field with values YES/NO, meaning it is only open if all three requirements are fulfilled. Another option is to provide the degree of openness by analogy with the 5 stars of open data ([*https://5stardata.info/en/*](https://5stardata.info/en/)).

Other members suggested to indicate only whether the data are open or closed, and to further expand this indicator in the future. Should we add a field “restrictions” to indicate when a dataset is not open? For instance, when it concerns personal/sensitive data? This might be undesirable since then we add another field that doesn’t exist in DataCite. It was proposed to only use the field *IP rights* (url rights: creative commons) for our open data KPI since, if there’s no open license, the dataset is not open by default.

The discussion on ‘open data status’ will be resumed in the next meeting.

**Data link to projects**

Data-link to projects refers to DataCite RelatedIdentifier: funderID; RelationType: (1 relation, result of?)

It provides a list of the project/grant agreement identifier(s). It was agreed to make this field Mandatory if applicable, yet Mandatory for BOF, IOF, FWO and public funding as specified in covenants. Agreements should be made on the format of references to the funder. Relation Types: the direction of the relations is difficult. Using links to a project in FRIS makes sense, there are identifiers for projects and publications in FRIS. We can use relationship roles that already exist in FRIS and we can adapt them to make them in line with DataCite relation types. For example, this dataset is the result of, and an identifier of a project. It might not be easy for researchers to get the correct identifier. It is valuable to have the context of the dataset, link to project etc. Will we use only one relation type? It was decided to drop the link to funder since we have already the link to the project.

**Data link to DMP**

DMP is not included as a field in Datacite, only as a free text field of resource type, so we might have to do something new. It’s not necessary to put DMP’s on FRIS, but for new projects we should know whether they have a DMP. We should provide a link to the DMP and a link to the project in FRIS. DMP itself will not be open. Not every stakeholder uses DMPonline.be, so you don’t have a unique way to put identifiers there. The discussion where to get these identifiers is for the KPI group.

**Data link to datasets**

EWI proposes that if someone wants to give links to datasets it should be possible.

Link to dataset: Relatedidentfiertype: DOI, relation type: Datacite links

**Data link to publications**

EWI: Link to publication; should we use something more generic than ‘isreferencedby’ as a relationtype?

Either provide link publication - dataset(s) or indicate that no datasets are available for publication

**FAIR data label**

A FAIR data label has to be developed by FOSB. By 2024, 90% of all datasets should have a FAIR data label (second quarter of 2021), and 60% of those should meet a high FAIR standard (baseline measure in 2022). First the assessments will be based on self-evaluation, later automatization processes like AI will be developed. A Fair data label (about the data not metadata) is not included in Datacite, we tested whether the metadata fields are FAIR but what about the data? It was suggested to think of an algorithm.

|  |
| --- |
| A new meeting has been scheduled for September 29th, 15.30-17.30 pm. |