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

**ONTWERPVERSLAG**

FOSB WG Metadata & Standardisation

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

Datum: 04/09/2020

1. Resume In-depth discussion of the draft metadata model
2. Discuss implementation of open data indicators

**Aanwezig**

Joke Meeus FWO

Niek Van Wettere VUB

Rudi Baccarne UAntwerpen

Roxanne Wyns KU Leuven

Ils de Bal Departement EWI

Pascale Dengis Departement EWI

Sadia Vancauwenbergh ECOOM-Hasselt

Evy Neyens ECOOM-Hasselt

Dries Moreels U Gent

Dimitri Brosens INBO

Helga Deschrijver IMEC

Francisco Hernandez VLIZ

Katrien Desmet Vlhora, vervangt Ilse Beerland

Mia Vanstraelen Industrie; the Brain group

Quentin Groom Plantentuin Meise

Alexander Botzki VIB

Hans Polet ILVO

Rinaldo Beck VIB

**Verontschuldigd**

Frederik Coppens VIB -ELIXIR ESFRI

Steven Heylen FWO

**Verslag Vergadering 04/09/2020**

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

Second in-depth discussion of the draft metadata model

We resumed our discussion to formulate a generic metadata model for research datasets.

First, we discussed a few recent changes that were made after a discussion with the schools of Arts & Humanities. They suggested changing our definition of a dataset to include that the data would be created or generated in the course of researchers’ investigations to test a hypothesis. Hence, the definition of a dataset was adapted as follows: “*Data and objects generated or collected by researchers affiliated by a Flemish research institution in the course of their investigations, regardless of their form or method, that form the object on which researchers test a hypothesis. This includes the full range of data: raw, unprocessed datasets, proprietary generated and processed data and secondary data obtained from third parties*.”

Next, the definition of ‘identifier’ was adapted to include the landing page where the dataset is *downloadable*. The definition of identifier currently is: “*The Identifier is a unique string that identifies the landing page of a downloadable dataset, i.e. a concept DOI or a handle. Each metadata record has a minimum of 1 identifier. A concept DOI is preferred over a version DOI.”*

This adaption was made to accommodate the KPI’s that we should reach in order to obtain funding from the Flemish government. The Department EWI wants to be able to map how many datasets are FAIR and open. One important indicator of both FAIR and open data is whether the dataset is immediately downloadable via a link or not.

It was agreed to make “Embargo date” a mandatory field that can be answered with Yes/No. The subfield “Date type” will be Mandatory if applicable with the following Date type values:

- issued: date used to indicate when the dataset is made available, published or uploaded to a formal database.

- accepted: date used to indicate the start of an embargo period

- available: date used to indicate the end of an embargo period, sometimes specific conditions are applying (GDPR, ...)

The question was asked whether the status “available” refers to open data. This is not necessarily the case. The dataset is available when the endpoint of an embargo period is reached, but this does not specify whether the dataset is open or not. For example, the accessibility of a dataset that contains sensitive data may be closed or restricted after an embargo period, which means the (meta)data are FAIR but not open. According to the General Data Protection Regulation (GDPR), sensitive data always undergo an embargo period. The data can still be opened after a special request to open the data. With biomedical data samples, for instance, it is common practice to request the data among researchers, although this is not the same as a free, online, open download. This is in line with the standards of the EOSC stating that data should be “as open as possible, as closed as necessary”. Hence, it might happen that, when for example privacy concerns are valid, research data are only available under specific conditions, and thus not openly downloadable for everyone. Within the life sciences it is common practice to submit data that is under embargo until the publication is published.

Then we discussed the metadata field “Research discipline” which is a classification of the dataset on the basis of the disciplines of the Flemish Research Discipline Standard. It was agreed to make this metadata field mandatory. This classification is a detailed, elaborate list of research disciplines that is already used to classify Flemish researchers, publications and projects. Some members wondered whether the same research disciplines apply as to the publication that coincides with the dataset. This might not necessarily be the same since a publication starts with a hypothesis that corresponds to a specific research discipline, but the application field of the data can be broader. An important point of discussion was the level of granularity of the research discipline list that would be best to use to classify research datasets on the FRIS-portal. The Department EWI suggested using the Flemish research discipline list until the L3 level. However, it was suggested that this level would be too narrow for some disciplines and too broad for other disciplines.

It was observed that while the L2 level might be too broad to be valuable, the L3 level requires preliminary knowledge to know what you’re looking for. For the L3 level, additional effort is needed to include all disciplines also from the socs. It should be feasible to add more research disciplines given that the procedure to add new disciplines to the list is clearly laid out. The question was posed whether people really look for datasets based on the research discipline, how often is this used?

VUB is not in favour of a mandatory provision of discipline codes for datasets, and this for several reasons:Generally speaking, as indicated earlier in the FOSB WG, we believe it is best to remain as close as possible to the DataCite/OpenAire model. The main reason for this is to strive as much as possible for a "single-entry model", where it is sufficient for researchers to deliver the metadata to the data repository (Zenodo etc.) where the data are archived, without the researcher having to fill in additional metadata manually in the institutional CRIS system. The data repository is the central actor for the delivery of dataset metadata. DataCite offers a realistic view of which metadata elements are currently usually available through data repositories, and which are not (or only optional). In the OpenAire guidelines the metadata field "subject" (which seems to be quite close to the principle of discipline codes) is labelled as "recommended" (see https://guidelines.openaire.eu/en/latest/data/field\_subject.html).

FRIS provision of discipline codes for datasets is not mentioned in the BOF decree.

Also for publications, there is no mandatory FRIS provision of discipline codes.

Then the metadata field “Keywords” was discussed. Some members wondered about the added value of this field relative to the metadata field “Research discipline”. The Department EWI suggested that keywords are important since they can be much more specific than research disciplines which is a great help for discovery for people who come to the FRIS-portal looking for specific types of datasets. Moreover, search engines look into both the description and keywords. Keywords are additional information that have more weight than each word in a description in the FRIS algorithm for publications. Hence, keywords are more important than a description for discovery purposes.

It was noted that it is standard practice and mandatory to use keywords with scientific publications and projects, so why would this be any different for research data? But then again it is another effort to ask from researchers. It was argued that researchers who want to make their data discoverable, will add keywords, and using a free text field is even more simple than using a super granular code list.

Other members recommended to stick to the mandatory and recommended elements of DataCite since this is what most institutions will already have in place. Further, it was also mentioned that we should adopt the view that these metadata are in the first place submitted to a data repository (not always institutional) which might not have a good capture of keywords.

Further, it was proposed to make the keywords only mandatory for datasets that are funded via FWO, IOF or BOF funding. The question was raised on the alignment with the EOSC in this regard, however this is not known yet because the EOSC has not been fully developed.

Similar to the discussion on the research discipline classifications, the metadata field “Keywords” is also used with publications to highlight the research and to aid with discovery. The question was posed whether the keywords used to describe a dataset align with the keywords that are used to describe the corresponding publication? Some members wondered whether it would be possible to make an automatic link between keywords of a publication and the corresponding dataset. It was recommended to use controlled vocabularies for the keywords, a list is already available.

The members agreed to make the field “Keywords” mandatory and to make it a free text field since this requires less effort for researchers than looking into a code list. When using a free text field, researchers will use keywords that come immediately to mind and are the most valuable to describe the dataset. Hence, the main keywords describing the dataset should be provided using a free text field.

Regarding the field “Scientific contact person” it was insisted that this field is not a mandatory field in DataCite, but since it is required to provide the contact details of a “Person with knowledge of how to access, troubleshoot, or otherwise field issues related to the resource”, in order to make data FAIR, it might be wiser to choose someone who is sustainably connected to the institution. It was suggested that this person might also be the data manager of data curator. It was suggested to provide contact information at the level of the research group since sooner or later any given person will be gone.

The members decided to take the DataCite term and definition of “Contact person” and make it Mandatory : “Person with knowledge of how to access, troubleshoot, or otherwise field issues related to the resource. May also be “Point of Contact” in an organisation that controls access to the resource, if that organisation is different from Publisher, Distributor, Data Manager.” It is up to the knowledge institution to choose who will take on this role.

The contributing researcher is defined as: “Name of researchers, not included in “creator”, that were involved to some extent in producing the dataset.” This field is Mandatory if applicable and will be completed using a dropdown list using the DataCite contributor list:

Contributing organisation / contributing person:

ContactPerson; DataCollector; DataCurator; DataManager; Distributor; Editor; HostingInstitution; Producer; ProjectLeader; ProjectManager; ProjectMember; RegistrationAgency; RegistrationAuthority; RelatedPerson; Researcher; ResearchGroup; RightsHolder; Sponsor; Supervisor; WorkPackageLeader; Other

The subfield “Name type” specifies whether it’s a personal or organizational name.

This allows people to give credit to the right people. Only the used contributor roles will be highlighted on the FRIS-portal.

It was decided to make the fields *technical information, version, and size,* optional since this refers more to the metadata of the dataset itself, which the researcher will take up in the DMP. However, the field *Format* should be mandatory to be able to assess how many datasets are available in an open format, since this a KPI for reusability. It was agreed to make format information mandatory on the landing page itself but optional on FRIS. It was suggested to make this a free text field, however, EWI prefers to use a list with fixed values: XML, CSV, JSON, RDF, TXT.

IP rights are defined as: “Intellectual property rights for the dataset”. It was suggested to make this field mandatory since this can be used to assess how many datasets use an open license that enables free reuse because this is an important KPI for open data*.* According to the FAIR reusability indicator:[R1.1. (Meta)data should be released with a clear and accessible data usage license](https://www.go-fair.org/fair-principles/r1-1-metadata-released-clear-accessible-data-usage-license/). A dataset is only truly reusable when it has an open license. However, it was decided to make this field recommended. This field will be completed using free text or a link to a license (e.g. [*Http://creativecommons.org/licenses/by/3.0/de.deed.en*](http://creativecommons.org/licenses/by/3.0/de.deed.en)).

An agreement was reached to make the field “Open data status” mandatory. This field refers to the status of the access possibilities of the dataset. It was suggested to look at 3 dimensions of open data: open availability, open format, & open license. However, others argued this would be too complex and this is not the task of researchers to fill this in. There are also technological limitations: not always possible to change to an open format without loss of information. It was proposed to follow the OpenAire guidelines. OpenAire explicitly discloses the access right of the resource via the Rights property (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

It wasn't decided yet how we're going to fill in this field.

|  |
| --- |
| 1. A new meeting was scheduled for Wednesday, September 16th, 10am-12pm |