A list of recommendations to improve the workflow of dataset metadata in the Flemish research landscape.

Evy Neyens[a], Evelien Dhollander,[b] Dieuwertje Bloemen[c], Kevin Leonard[b], Werner Brebels[a], Tom Wuyts[d], Pascale Dengis[d], Marc Portier[e], Hanne Poelmans[a]

a. Directorate Research, Library and Internationalisation & Centre for Research & Development Monitoring (ECOOM-UH), Hasselt University, Belgium.
b. Research Department, University Library, Open Science Team, Ghent University, Belgium.
c. KU Leuven Libraries, LIBIS operations & relations, KU Leuven, Belgium.
d. Economy, Science & Innovation Department, Flemish Government, Brussels, Belgium.
e. Flanders Marine Institute, Open Science Team, Ostend, Belgium.

Background

Flemish research performing institutions are obligated by the Flemish government to provide descriptive metadata on research entities (researchers, organisations, projects, publications, datasets, infrastructure, and patents) to *Flanders Research Information Space* (FRIS), the research portal of the Flemish government. FRIS is a Current Research Information System (CRIS)[1] that collects and showcases information on publicly funded research carried out by Flemish research institutions. As a business information tool, FRIS reports research and development (R&D) and innovation indicators to the Flemish government and is used to monitor the Key Performance Indicators (KPIs) resulting from the Flemish Open Science Policy. FRIS also serves as a discovery hub connecting with the European Open Science Cloud[2] (EOSC) through OpenAire[3].

As part of the Open Science policy in Flanders, knowledge institutions have agreed to mandatorily register metadata of datasets that underlie publicly funded, peer-reviewed scientific articles and submit them to FRIS, commencing in 2019. To facilitate this process, the Flemish Open Science Board's (FOSB) working group Metadata & Standardization, chaired by ECOOM-UHasselt, developed an application profile specifically for dataset metadata in 2020 (Neyens & Vancauwenbergh, 2021). The FOSB Application profile is built upon the DataCite model but has been expanded in order to effectively monitor the metrics of Open Science. The FOSB Application profile for dataset metadata ensures that the right information is available in FRIS to calculate the Open Data KPI. The Open Data KPI requires that, from 2019 onwards, datasets underlying peer-reviewed, Flemish-funded scientific publications should be openly accessible as far as possible (excluding legitimate opt-outs). Hence, the dataset metadata application profile facilitates the assessment of the proportion of 'open' datasets in terms of access level and data usage licenses.

A next step was to map the current flow of dataset metadata in the Flemish research landscape and to identify challenges and business needs in order to optimise this workflow. To this end, a *project group (PG) Enriching metadata* was established under the *Flemish Research Data Network* (FRDN)[4], that is responsible for implementing and monitoring Open Science in Flanders.
The PG Enriching Metadata is situated within the FRDN network under the umbrella of the Working Group (WG) Architecture and ran from September 2022 to December 2023. The PG consisted of 9 members (metadata experts from different research performing organisations (RPOs)) and was

[1] https://researchportal.be/en. FRIS is integrated with the CRIS systems and research infrastructures of the Flemish RPOs where information is exchanged and revised using the CERIF 1.5 exchange format with specific FRIS extensions (in some cases manual input is provided).
[2] European Commission. (2017). European Open Science Cloud. New research and innovation opportunities. Retrieved from https://eosc- portal.eu/sites/default/files/eosc_declaration.pdf
[3] https://www.openaire.eu/
[4] https://www.frdn.be/

chaired by ECOOM-UHasselt[5] and co-chaired by Ghent University. The scope of the PG Enriching Metadata was twofold: scope 1 focused on optimising metadata gathering & publication processes within the Flemish research ecosystem and scope 2 aimed to develop an institution-wide approach for harvesting research dataset metadata from external repositories. This abstract will report on the main outcomes of scope 1 of the project group. The key deliverable for scope 1 was to a) create a Business Process Model and Notation (BPMN) visualising the current flow of research dataset metadata within the Flemish research landscape, and b) to develop a list of prioritised points of improvement.

Method

A survey was administered to obtain understanding into the current process flows related to dataset metadata in the Flemish research ecosystem, in order to identify areas for improvement, and to elaborate on an optimal flow of dataset metadata. The survey consisted of two sections. Questions in section 1 were centered on current approaches to dataset metadata registration and collection, curation, publication, and reporting. Section 2 focused on the desired situation as envisioned by the stakeholder respondents. The survey was followed up by bilateral structured interviews to delve deeper into desired, future scenarios. Twelve RPOs took part in the survey (consisting of 5 Flemish universities, 3 strategic research centres, and 4 Flemish scientific institutions), with nine of them also participating in the follow-up interviews.

Results

The responses to section 1 of the survey revealed that there is considerable variability in the use of CRIS-systems to record dataset metadata among institutions. Several institutions have deployed modules into their CRIS or research infrastructure to record dataset metadata or are in the process of developing them. Not all institutions have a CRIS in place, but they may record metadata in other ways. A few RPOs are currently not collecting any dataset metadata yet, although processes are being set in motion (25%). 75% of participating institutions do not have formal obligations regarding the registration of dataset metadata in their RDM policy. However, about one fourth do state a strong recommendation to submit dataset metadata. Half of the respondents use the FOSB application profile to register dataset metadata (including all 5 Flemish universities). Some of the surveyed institutions (25%) use multiple metadata standards to suit the diversity of the data they hold. These typically are strategic research centres and scientific institutions with highly-specialized data.

Based on section 1 of the survey, a business process model of the present flow of dataset metadata was created. This part will briefly describe the present flow (to view the BPMN and for more information read the full report on Zenodo[6]). Overall, the dataset metadata workflow starts with the researcher publishing a dataset in an external repository (e.g. Zenodo, Figshare, Dryad).
The dataset repositories request descriptive metadata to be registered and usually record a Persistent Identifier (PID; e.g. DataCite DOI) for the dataset included in the metadata. However, most repositories only ask for minimal mandatory fields (common use of generic metadata schemes such as DataCite & DublinCore). This poses challenges when, in a consequent step, researchers are expected to record the dataset metadata into the CRIS or institutional metadata repository of their affiliated institution(s). Because the FOSB application profile that is used to register dataset metadata in internal CRISs (or other infrastructures) contains more mandatory metadata fields than what is typically requested by repositories, researchers and research administrators often need to enrich the

[5] https://www.ecoom.be/en
[6] https://doi.org/10.5281/zenodo.10634664

metadata to meet the requirements for Open Science in Flanders. Moreover, registration and enrichment are performed manually in most cases (9 out of 12 participating institutions), leading to much administrative effort. To reduce regulatory burden, half of the participating RPOs (6/12) provide autocomplete for certain metadata fields during manual metadata registration, allowing automatic metadata retrieval by entering a PID (e.g. DOI, Handle, ORCID). At this stage, the pulled metadata need manual enrichment in order to meet the requirements of delivering to FRIS based on the FOSB application profile.

In a next step, the validation and curation of the metadata record occurs after it has been registered or imported and manually updated. Data stewards ensure the records are complete, they add business entities (employee database etc.), link the datasets to their corresponding projects and funding sources, and validate if the FRIS business rules are met. If any information is missing, researchers are contacted to provide it. Finally, in some cases, the metadata must be mapped to the FOSB application profile as expected by FRIS.

In a final step, metadata is transmitted to FRIS via SOAP webservices in most cases.
A number of institutions that are using Pure by Elsevier have their metadata pulled in by FRIS through a custom add-on. A few organisations send metadata through an XML-file or manually through the FRIS-UI. Most of the strategic research centres and Flemish scientific institutions are not currently sending dataset metadata to FRIS (dataset catalogues in development).

Section 2 of the survey inquired the RPOs about their business needs and points of improvement regarding the current flow of dataset metadata in the Flemish research ecosystem.
A thorough analysis of the interview responses yielded a list of 10 recommendations. Using a prioritisation exercise, these 10 recommendations were scored for implementation ease and degree of impact on the current research landscape. Considering recommendations with high impact and low implementation difficulty yielded a selection of the following 4 recommendations:

1. RPOs should utilise the regional research data portal FRIS more often to pull in, validate and enrich metadata previously provided by partner institutions (for example in the case of collaborations and multiple affiliations), and then feed it back to FRIS. This would greatly reduce administrative investments in metadata registering and publishing, and avoids duplication errors (e.g. different titles, order of authors/dataset creators etc.).
2. RPOs should consider building out integrations with external repositories and aggregators to collect dataset metadata from affiliated researchers (This is worked out as a proof of concept in key deliverable 2 in the full paper on Zenodo[7]).
3. RPOs should investigate the potential of integrating premium ORCID membership. Promoting proper ORCID usage among researchers ensures RPOs are aware of how and when to access dataset metadata.
4. The FOSB application profile for research datasets has more mandatory metadata fields than commonly used standards like DataCite and Dublin Core, resulting in additional manual work. Changing some metadata fields from 'Mandatory' to 'Recommended' or 'Optional' would reduce manual workload. This would allow more dataset metadata records to be submitted to FRIS. It remains important to encourage RPOs to provide as much metadata as possible, however, for example by flagging records that lack necessary metadata fields for the Open Science KPI calculations.

[7] https://doi.org/10.5281/zenodo.10634664

In addition, the project group formulated a general advice for a fifth recommendation that had the highest impact score yet also a low implementation ease: the implementation of funder grant IDs. During the interviews, respondents emphasized the need for Grant PIDs to unambiguously link datasets and publications to their respective funding grants. PIDs are globally unique and sustained digital research object references that offer stable identification and referencing of research entities (e.g. publications, datasets, persons) and their associated metadata (Brown, 2020). PIDs facilitate information exchange and enable automation of administrative workflows (Brown, 2020). Grant PIDs, such as the Crossref Grant ID, provide trusted and automated connections between grants and researchers, institutions and research output[8]. This facilitates workflows for reporting, evaluation, funder compliance, impact monitoring, and collaboration.

The general advice here is to establish an FRDN project group to explore the potential for a Flemish PID strategy focusing on Grant IDs. This project group should start by analysing the current European PID landscape, review PID roadmaps, policies, and projects from other countries, and explore the benefits and requirements of implementing Grant IDs in the Flemish academic context.

Conclusion

This abstract reports the main findings of part of a survey, conducted by the FRDN project group Enriching metadata, collecting the viewpoints of the participating stakeholders on their current and ideal dataset metadata flow. This resulted in a list of recommendations to improve the flow of dataset metadata in the Flemish research landscape. For more information, readers are referred to the full report on Zenodo[9].

Acknowledgements: the members of the Flemish Research Data Network (FRDN) Project Group Enriching Metadata. Co-chairs: Evy Neyens (ECOOM-UHasselt), Evelien Dhollander (UGent).
Members: Dieuwertje Bloemen (KU Leuven), Kevin Leonard (UGent), Werner Brebels (UHasselt), Pascale Dengis (EWI-FRIS), Tom Wuyts (EWI-FRIS), Marc Portier (VLIZ).

References

Brown, J. (2020). Developing a persistent identifier roadmap for open access to UK research. https://repository.jisc.ac.uk/7840/

Neyens, E., Dhollander, E., Bloemen, D., Leonard, K., Brebels, W., Wuyts, T., Dengis, P., & Portier, M. (2024). Dataset Metadata in the Flemish Research Landscape. Zenodo. https://doi.org/10.5281/zenodo.10634664

Neyens, E., & Vancauwenbergh, S. (2021). Towards a Semantic Interoperable Flemish Research Information Space: Development and Implementation of a Flemish Application Profile for Research Datasets. *International Journal of Digital Curation, 16(1)*. https://doi.org/10.2218/ijdc.v16i1.762

[8] https://scholarlycommunications.jiscinvolve.org/wp/2020/10/12/theres-a-pid-for-that-part-1-grants/
[9] https://doi.org/10.5281/zenodo.10634664