Dataset Metadata in the Flemish Research Landscape.

FOSB Report 2023

Flemish Research Data Network: 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).
Acknowledgement: Robert van der Vooren (Flanders Make).

November 2023

Executive summary

The scope of this project group (PG) was to optimize process flows and systems related to research (meta)data gathering and publishing for Flemish research performing institutions (RPOs).
The PG aimed to reduce monetary and labour costs of managing shared research data for all partners involved, while improving the discoverability and quality of research metadata.

A questionnaire was distributed and followed up by structured interviews with research support staff from various Flemish research performing organisations (RPOs) to understand the current ecosystem and identify areas for improvement. The scope of key Deliverable 1 was to visualise the current flow of research dataset metadata within the Flemish research landscape using a business process model (BPM), and to produce a list of recommendations to improve this flow.
After mapping the current flow based on part 1 of the questionnaire, part 2 consulted the institutions on their business needs, which areas for improvement they considered necessary, and what the ideal flow would look like. 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 impact, after which the following 4 recommendations were selected for further elaboration:

1. RPOs should make greater use of the regional research data portal (Flanders Research Information Space, FRIS[1]) to pull in, validate and enrich metadata previously provided by partner institutions (for example in the case of collaborations, and researchers with multiple affiliations), and then feed it back to FRIS. This would greatly reduce administrative investments in metadata registering and publishing.
2. RPOs should explore opportunities to build out integrations with external repositories and aggregators to ingest dataset metadata from affiliated researchers (This is worked out as a proof of concept in key deliverable 2, pp.25).
3. RPOs should explore the opportunities of premium ORCID integration. When correct ORCID usage is promoted with researchers, this system ensures that RPOs know where and when to retrieve dataset metadata.
4. The FOSB application profile for research datasets contains more mandatory metadata fields compared to commonly used standards such as DataCite and Dublin Core, creating extra work due to the need for manual enrichments. Changing the status of some of the metadata fields of the FOSB application profile for research datasets from 'Mandatory' to 'Recommended' or 'Optional' would significantly reduce the workload of manually enriching fields with missing information. As a consequence, more dataset metadata records can also be submitted to FRIS. Here, it remains important to stimulate RPOs to provide as much metadata as possible, for instance by flagging records that do not contain all necessary metadata fields to be included in the Open Science KPI calculation.

[1] https://researchportal.be/nl

Based on the interviews, it was found that RPOs encounter many challenges when trying to retrieve and capture dataset metadata (often manual entry by researchers or research staff).

RPOs generally have a very limited overview of their own research data outputs, so a necessary first step is to find their affiliated datasets. The scope of Key Deliverable 2 was to develop a method to find and ingest dataset metadata associated with specific knowledge institutions, that can be implemented by all RPOs. To define this method, business needs were collected and prioritised, limitations of available technology were identified and a proof of concept (POC) was implemented.

In this POC two new methods were explored and compared to the current flow. A first candidate is 'The Modular Approach' in which various sources (aggregators, repositories,...) are iteratively queried by the RPOs to find the metadata related to their organisation.

The second candidate, 'The Registration Approach', is already being implemented by several RPOs (e.g., UGent, KU Leuven), and involves researchers entering DOIs (or other PIDs) into a system which then automatically harvests the relevant metadata from the appropriate repository.

The main output of Key Deliverable 2 is a description of a two-pronged approach for registering more datasets within the Flemish ecosystem: a combination of a modular and a registration approach. The modular approach is based on an analysis of the coverage of existing dataset aggregator services and an examination of which sources could be consulted to find the greatest number of datasets affiliated with Flemish institutions. The analyses showed that the best method to find the highest number of datasets was to use the APIs of various data repositories to search directly for affiliated datasets. The proof-of-concept demonstrates that the modular approach was able to find over 800 datasets published in 2022, significantly more than the 320 currently registered in FRIS. We describe the methods used to search several of the main data repositories used by Flemish researchers, and we also describe the strategy for how the modular approach could be maintained and extended. We augment this with a discussion of the registration approach, whereby researchers could quickly and easily register their published datasets with their institutional CRIS and FRIS, using many of the same structural components as the modular approach.

The registration method has the benefit of immediate manual enrichment by the researcher and precision, while the harvesting approach only pulls in whatever metadata is available via the API's of the aggregators and repositories and will likely always include some false positives. The two methods together are complementary in nature, i.e. the registration method can fill the gaps in coverage and precision of the harvesting method.

Based on the POC, this project group advises to establish a formal group or community within the FRDN existing of research data staff (RDM support staff and technical staff) to develop and maintain a platform to exchange knowledge regarding metadata harvesting procedures (codes, queries, best practices etc.) based on the POC that resulted from this project group. This will benefit both dataset metadata registration and harvesting. Starting from the scripts for the most common repositories and aggregators, new scripts for other sources can be developed, shared and used by all RPOs. The RPOs can develop their own way of importing the output from the scripts into their CRIS-systems. This collaborative approach will ease the burden for all RPOs. However, in order for this approach to work, there is a need for long-term investments of time and (human) resources.

Introduction

This report aims to map the current flow of research dataset metadata in the Flemish research landscape and to identify possible points of improvement and business needs of Flemish Research Performing Organisations (RPOs). It also explores solutions for capturing research dataset metadata from external dataset repositories. This report has been prepared on behalf of the *Flemish Research Data Network* (FRDN) and the *Flemish Open Science Board* (FOSB).

The Flemish coalition agreement 2019-2024 decided to invest in Open Science in line with European guidelines and initiatives (e.g. EOSC[2]). At the European level, it was agreed that the outputs (publications, datasets, software,…) of publicly funded scientific research should be made openly accessible as soon as possible according to the principle, 'as open as possible, as closed as necessary'[3]. In December 2019, the *Flemish Open Science Board* (FOSB) was established, giving a mandate to the knowledge institutions to implement an Open Science policy in Flanders.

The implementation is coordinated via the *Flemish Research Data Network* (FRDN) - Coordination hub, led by Research Foundation Flanders (FWO). Representatives of research performing- and funding organisations (universities, universities of applied sciences, post-initial education, strategic research institutions, Flemish scientific institutions, and funders) actively participate in two working groups (WG Architecture and WG Research data management & Open Science) and several project groups to implement, track and monitor Open Science practices in Flanders.

The *Project Group (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. Both the 'FOSB KPI memo'[4] and the 'remediation action plan' described the need to: "*optimise process flows and systems related to the gathering and publishing of research metadata".* There is a need for a well-thought-out ecosystem that will improve the quality and discoverability of metadata and that will reduce costs substantially for all partners involved.

Hence, the scope of the PG Enriching Metadata was twofold. A first aim was to optimise metadata gathering & publication processes within the Flemish research ecosystem in order to save time & money, and increase the quality and findability of research dataset metadata. The key deliverables for scope 1 consisted of: a) a Business Process Model (BPM) visualising the current flow of research dataset metadata within the Flemish research ecosystem, and b) a list of prioritised points of improvement. A second aim was to find an institution-wide solution for collecting research dataset metadata from external repositories. Here, the key deliverable was to develop a proof-of-concept (POC) for harvesting metadata addressing the prioritised business needs.

Method

As a first step in accomplishing these goals, the FRDN project group *Enriching Metadata* distributed a questionnaire to gain insight into the current process flows related to dataset metadata in the Flemish research ecosystem, to discover points of improvement and business needs, and to elaborate on an ideal flow of dataset metadata. The questionnaire consisted of two parts.

Questions in part 1 focused on the current situation. Part 2 focused on the ideal situation as envisioned by the stakeholder respondents. Participating RPOs prepared and forwarded the answers to the questions in part 1 in advance, after which they were invited to participate in bilateral follow-up interviews focusing on part 2 (while elaborating on part 1 where necessary).

[2] https://eosc.eu/

[3] European Commission. (2017). European Open Science Cloud. New research and innovation opportunities. Retrieved from https://eosc- portal.eu/sites/default/files/eosc_declaration.pdf

[4] https://www.ewi-vlaanderen.be/sites/default/files/downloads/bestanden/5fc5f512b328e9000c0007f3.pdf

Twelve RPOs participated in the survey (5 Flemish universities, 3 strategic research centres, & 4 Flemish scientific institutions), nine of which also participated in the follow-up interviews.

Part 1. The current flow of research dataset metadata in the Flemish research ecosystem.

Flemish research performing and funding institutions (RPOs & RFOs) are legally required to submit descriptive research information on research objects (persons, organisations, projects, publications, datasets, infrastructure, and patents) to *Flanders Research Information Space* (FRIS), the research platform of the Flemish government. This is regulated under special and industrial research funding (BOF & IOF) decrees, decrees of Science & Innovation, subsidy agreements from various governmental funders, and the Flemish coalition agreement regarding Open Science whose KPIs are monitored through FRIS[5]. FRIS is a *Current Research Information System* (CRIS) that collects information on publicly funded research performed by Flemish research institutions. It functions as a business information tool to report R&D and innovation indicators to the Flemish government, as a window on research in Flanders for industry and research community within Flanders and abroad, and as a discovery hub that connects with the *European Open Science Cloud* (EOSC) via OpenAire[6].

To date, there are 36 institutions that provide metadata within the framework of monitoring Open Science indicators. FRIS is connected to the CRIS systems and research infrastructures of the Flemish RPOs (and for some institutions manual input is offered), information is exchanged and updated using the CERIF 1.5 exchange format with specific FRIS extensions. The automatic connection between FRIS and the CRIS systems is made in two different manners: metadata can be transferred to FRIS by a SOAP XML-upload or by an automated connection for Pure[7] CRIS systems. There are wide variations among institutions in the use and maturity of CRIS systems and other infrastructures to exchange research information. Most universities have their own CRIS systems up and running, while some other institutions (mostly strategic research organisations and scientific institutions) are still in the process of developing their CRIS system. There are also differences in the information that is structurally available and exchangeable in the internal systems.

A core principle of FRIS is that *ownership, validation and quality control* lies with the institutions as the authoritative source of the (meta)data. Next to that, stakeholders and FRIS agreed upon a high quality standard, translated into business rules. Most research performing organisations are using a push mechanism to provide validated metadata to FRIS. This approach in which data providers agree on standards and formats to exchange research information is unique within Europe.

However, it has been noted that the current information flow often involves *duplication* in data delivery from the institutions, and researchers may have to enter the same information multiple times (e.g. both in an external repository and in the affiliated institution's CRIS). This section outlines the current flow of research dataset metadata within the entire Flemish research ecosystem and identifies points of improvement. Rich and discoverable metadata in the external repositories are key to opening up, sharing, and reusing research outputs. However, this needs to be backed up and facilitated by infrastructures and information flows that reduce the time and effort that is needed to achieve this. In part 1 of the questionnaire, we asked the stakeholders about current approaches to *dataset metadata registration and collection, curation, publication* and *reporting*. A business process model of the present flow of dataset metadata was created based on the process visuals and responses we received from the institutions.

[5] https://www.ewi-vlaanderen.be/nieuws/flemish-open-science-board-fosb-opgericht

[6] https://www.openaire.eu/

[7] https://www.elsevier.com/products/pure

1.1 Institutional differences in the registration of dataset metadata: CRIS-systems and catalogues, internal policies and metadata standards.

In the Flemish research landscape, there are legal obligations for research institutions to register and submit dataset metadata to FRIS (e.g. BOF/IOF decrees, decree of Science & Innovation, subsidy agreements from various governmental funders, and the Flemish coalition agreement regarding Open Science). Within the framework of the Flemish Open Science policy, the institutions agreed to mandatory registration of dataset metadata and transmission to FRIS for all datasets underlying peer-reviewed, Flemish-funded publications from 2019 onwards. An application profile for dataset metadata was developed for this purpose by the FOSB working group Metadata & Standardization in 2020[8]. The FOSB Application profile is based on the DataCite model yet has been extended in order to monitor the Open Science KPIs. It consists of 10 mandatory fields, 9 'mandatory if applicable' fields, 4 recommended fields, and 10 optional fields.

The survey showed that there is quite some variability among institutions in the use of CRIS-systems to record dataset metadata. Several of the participating research institutions have deployed modules in their CRIS-system, academic bibliography or metadata catalogue in order to record dataset metadata, or they are in the process of developing this. However, not all institutions already have a CRIS in place, although they might record metadata in other ways.

About 75% of the participating institutions (9 out of 12) do not currently have any obligations regarding dataset metadata incorporated in their internal RDM policy. Two universities do have an obligation to register dataset metadata for datasets resulting from Flemish-funded research.
Although most RPOs do not have a formal obligation to register dataset metadata in their RDM policy, about one fourth do state a strong recommendation to submit dataset metadata. Two universities expect their researchers to submit dataset metadata (not limited to Flemish-funded research) to the institutional catalogue or CRIS. Half of all respondents (6 out of 12) use the FOSB Application profile to register dataset metadata (Figure 1, pp. 6). This group includes the 5 Flemish universities that all map their metadata to the FOSB metadata model. A number of institutions use multiple metadata standards to accommodate the diversity of the data they hold (25%). A few RPOs are currently not collecting any dataset metadata yet, although processes are being set in motion (25%, 2 SOCs & 1 scientific institution).

Furthermore, registration of rich dataset metadata in the repository chosen by the researcher is not a common practice yet among researchers. Many researchers are not aware of the requirements and benefits to opening up their data. According to the stakeholder respondents from the RPOs, researchers from their institution who are aware of these requirements and benefits typically register dataset metadata immediately after publication of the dataset (33%) in a repository, or after publication of the article resulting from that data (13%).

[8] 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

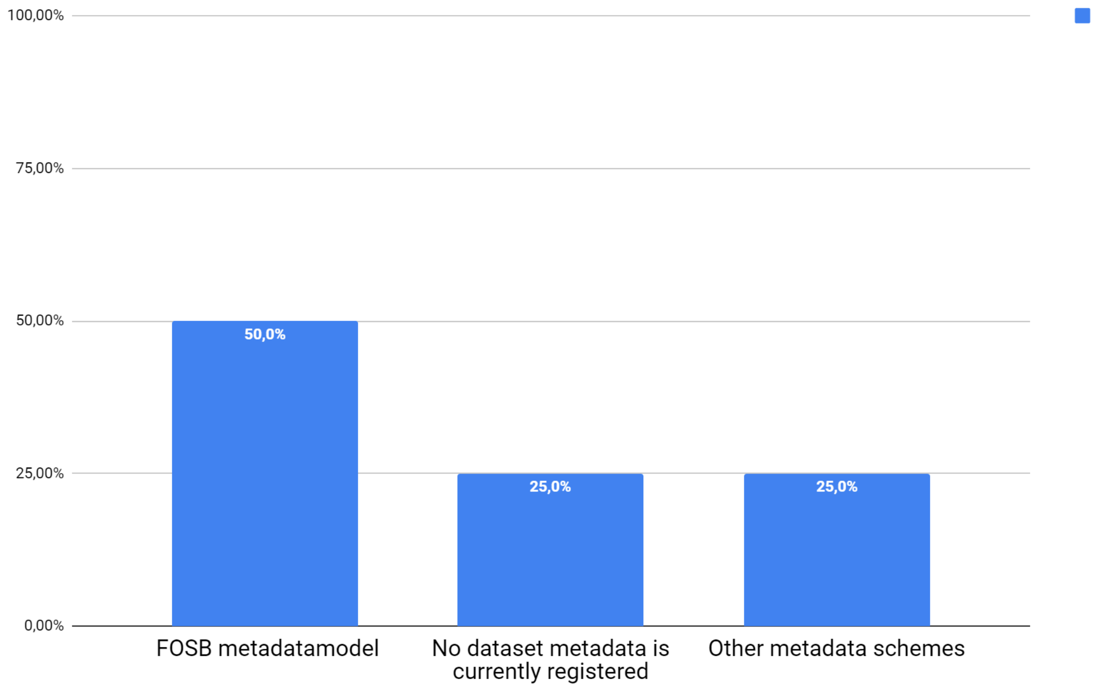


Figure 1. Usage of dataset metadata standards.

1.2 Dataset metadata business process flow: current flow to CRISs and FRIS.

Part 1 of the questionnaire aimed to map the current flow of dataset metadata by asking targeted questions to the participating RPOs as well as requesting business process flows visualising these processes. This resulted in a *Dataset Metadata Business Process Flow Visual* (Figure 2., p. 9).
In this section the different steps (metadata registration and collection, curation, flow to FRIS) will be described and complemented with data from the questionnaire.

Metadata registration.

In a first step, researchers typically publish a dataset into an *external dataset repository* (e.g. Zenodo, figshare, Dryad etc.), or in the institution's own dedicated dataset repository, if available. During this process, descriptive information about the dataset is requested by the repository and completed by the researcher. Most repositories, however, only require minimal mandatory fields (widespread use of minimal, generic metadata schemes such as DataCite, DublinCore etc.) resulting in an additional step where research administrators and/or researchers need to enrich the metadata to conform to the requirements of the FOSB application profile for dataset metadata (see next step: dataset metadata registration in institutional CRIS or repository).
If supported by the dataset repository, a connection with a number of PID services, like a DOI provisioning service or ORCID, is made during the publication process, resulting in the recording of PIDs as part of the metadata. Most of the time, the process for the researcher just ends after the publication of the dataset. Sometimes researchers are asked by their affiliated institution to send a *notification* after the publication of a dataset to ensure that the institution is aware of its existence.

In a next step, researchers are expected to *register the dataset metadata into the CRIS*

system[9] of their affiliated institution(s) or in a separate institutional metadata repository (MDR).
At 9 out of the 12 participating institutions, metadata enters the CRIS-system or institutional repository primarily via *manual registration* by the researcher. The researcher provides the required metadata in accordance with the FOSB Application profile - for half of the participating institutions- or by using another metadata standard, by completing a template or form. To reduce the administrative burden that comes along with manually completing metadata registration forms, institutions employ various methods. Half of the participating RPOs (6/12) offer an *autocomplete feature* for certain metadata fields (e.g. main title, authors…) during the manual metadata registration phase, to pull in metadata automatically by entering a PID (DOI, Handle, ORCID etc.). Some RPOs that work with Pure[10] are exploring the functionality of a *Data monitor* that allows them to semi-automatically import dataset metadata. It is at this stage that the metadata pulled in often needs to be supplemented with additional metadata fields that are required to be delivered to FRIS according to the FOSB application profile. Other examples to facilitate input include drop-down menus with controlled vocabularies or predefined lists for specific metadata fields, such as formats, licences, and access level, and automatic suggestions for affiliations.

Retroactive harvesting of dataset metadata from external sources.
Because of the legal obligations to register dataset metadata resulting from Flemish-funded, peer-reviewed articles as of 2019, data stewards from many institutions are actively seeking, collecting and pulling in this metadata into their CRISs or metadata catalogues. The primary approach used by the RPOs is to *query external sources via standard operating procedures (SoP)*
(e.g., OpenAIRE EXPLORE, DataCite, Zenodo, Dryad, Pangaea, Scholix, Google Dataset Search, SODHA). A second approach is to use the *DataCite and OpenAIRE APIs* pulling in metadata of published datasets. Some RPOs use the paid version of *Data search/Monitor*[11].

3 out of the 5 Flemish universities have a procedure in place to add datasets underlying publications retroactively from 2019 onward. The RDM team actively searches for these datasets using APIs and specific DOIs after which researchers are contacted to complete the records.
Some universities limit this effort by periodically scanning Zenodo and DataCite. One university has their own RDR Dataverse API to harvest metadata. Either way, if a DOI is available or known to the finder after having found a dataset, most (or all) of the metadata is first being fetched or imported from services like DataCite or Crossref or by integrating with an aggregator like OpenAIRE. If no DOI exists, then it's up to the organisation to add the metadata manually and optionally register a DOI afterwards. The strategic research centres (SOCs) and Flemish scientific institutions (FSI) are not yet prioritising retroactive retrieval of dataset metadata as only two out of the seven participating SOCs/FSIs do some manual harvesting. 5 out of the 12 participating institutions do not consult any external sources for dataset metadata harvesting yet.

Metadata curation. In the metadata curation phase, validation and curation of the metadata record typically happens a first time after the metadata record has been imported and/or manually updated (e.g. deduplication, completion,…). Once available in the institutional CRIS system
(or repository), the metadata often is enriched with other business entities that are available within the institution (employee database, academic file etc.), the business entities are mostly uploaded to the CRIS system by means of a batch process. Since not all information is linked yet, some business objects (e.g. projects and funding information) must be linked manually. Data stewards enrich the records with the necessary internal project and dataset creator information, they check the completeness (mandatory fields FOSB) of the metadata and validate if the FRIS business rules are met

[9] A current research information system (CRIS) is a database or other information system to store, manage and exchange contextual metadata for the research activity funded by a research funder or conducted at a research-performing organisation (or aggregation thereof).

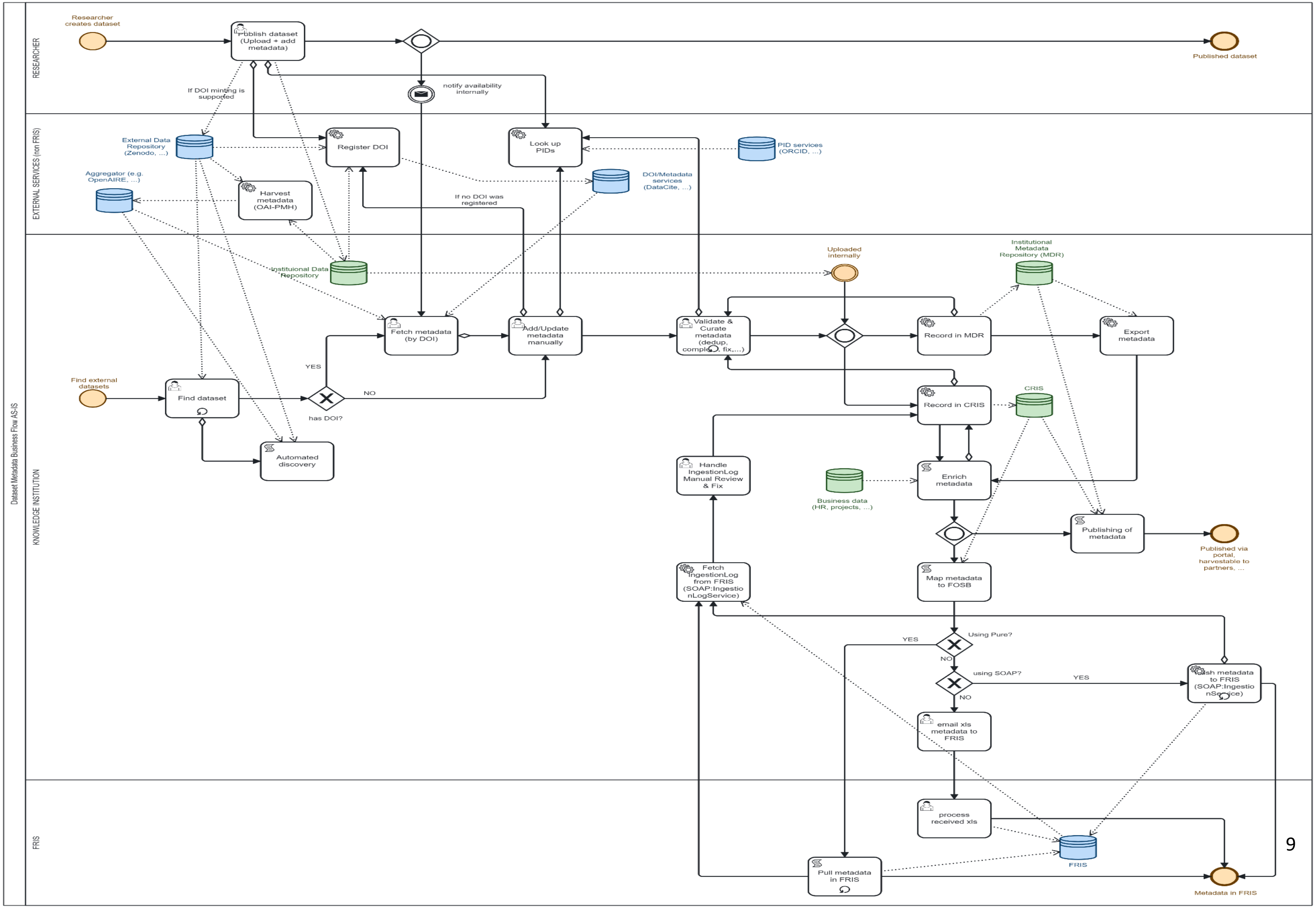
[10] https://www.elsevier.com/products/pure

[11] https://www.elsevier.com/products/data-monitor

(4 out of 5 universities). If any information is missing, the researchers are contacted to complete the missing fields. From there, the metadata must be mapped to the FOSB metadata model since this is what FRIS expects as input model. However, mapping isn’t always necessary since many RPOs (all of the 5 universities) already use the FOSB application profile as their standard metadata registration template.

Flow to FRIS. Metadata is sent to FRIS either via SOAP API, XML-upload (e.g. XML import) or via manual registration. If an organisation uses Pure (by Elsevier) as their CRIS system, they can have their metadata pulled in by FRIS. This integration has been made available as a custom add-on in Pure. Non-Pure organisations however are asked to push their metadata to FRIS by integrating and invoking SOAP web services. A minority of organisations don't have an integration and are required to send the metadata by means of an xml-file or by entering manually through the FRIS-UI. Most of the strategic research centres and Flemish scientific institutions are currently not sending any dataset metadata to FRIS yet (dataset catalogues in development). FRIS allows the retrieval of ingestion log files after having delivered the data to FRIS allowing organisations to manually fix and review metadata fields in their own system.

Figure 2. Dataset Metadata Business Process Flow


Dataset Metadata Business Flow AS-IS
RESEARCHER
EXTERNAL SERVICES (non FRIS)
KNOWLEDGE INSTITUTION
FRIS
Researcher creates dataset
Publish dataset (Upload + add metadata)
Published dataset
notify availability internally
If DOI minting is supported
External Data Repository (Zenodo, ...)
Register DOI
Look up PIDs
PID services (ORCID, ...)
Aggregator (e.g. OpenAIRE, ...)
Harvest metadata (OAI-PMH)
DOI/Metadata services (DataCite, ...)
If no DOI was registered
Instituional Data Repository
Uploaded internally
Institutional Metadata Repository (MDR)
Fetch metadata (by DOI)
Add/Update metadata manually
Validate & Curate metadata (dedup, complete, fix,...)
Record in MDR
Export metadata
YES
Find external datasets
Find dataset
has DOI?
NO
Record in CRIS
CRIS
Automated discovery
Handle IngestionLog Manual Review & Fix
Business data (HR, projects, ...)
Enrich metadata
Publishing of metadata
Published via portal, harvestable to partners, ...
Fetch IngestionLog from FRIS (SOAP:IngestionLogService)
Map metadata to FOSB
Using Pure?
YES
NO
using SOAP?
YES
Push metadata to FRIS (SOAP:IngestionService)
NO
email xls metadata to FRIS
process received xls
FRIS
Pull metadata in FRIS
Metadata in FRIS

1.3 Flow of dataset metadata to FRIS: technical, legal and practical issues.

Part 1 of the questionnaire also inquired the institutions about the challenges that they encounter with the flow of dataset metadata to FRIS. In this section, a distinction is made between technical, legal and practical issues. A few *technical issues* that were mentioned include *errors in the ingest to FRIS* leading to extra workload debugging and resubmitting. In line with this, many institutions stated that SOAP XML is considered an outdated technology and should be completed by preferably REST JSON or OAI-PMH. Another challenge regards *the multiple records of the same dataset.* This means that the same information object can have multiple records in FRIS
(e.g. differences in titles of the same dataset, carried out by complementary universities).
Dataset metadata that are affiliated with multiple institutions are submitted to FRIS multiple times because the metadata need to be registered in the repositories of all the affiliated institutions.
Because researchers with multiple affiliations can have multiple and separate metadata records in FRIS, dataset creators cannot always be resolved unambiguously. This can be solved by adequate merging of those records using ORCIDs (golden record). Next, respondents mentioned *legal issues* such as the question of ownership of the metadata, which party is responsible for quality control and validation and whether the metadata needs a licence.

A number of *practical challenges* were disclosed as well. The absence of *project and funding grant PIDs* to unambiguously link projects and funding to datasets was most often cited as a barrier for data sharing between local Flemish stakeholders. Currently, the information on the links between a project and its research outputs lies solely with the researchers themselves. Due to the overall lack of PID usage there are missing links between research information objects (people, projects, organisations, publications, datasets).

A second oft-cited challenge relates to the extra mandatory attributes in the FOSB metadata scheme (e.g., access rights, licences,...). All institutions mentioned that the FOSB metadata scheme is not aligned with current international standards and requires many additional, mandatory fields.
As a consequence, much time goes to manual enrichments that need to be performed when harvesting dataset metadata from external repositories.

It was also noted that the FRIS Business rules are sometimes quite strict, for instance for the measurement of the Open Science KPIs there are mandatory fields, which leads to only a subset of research datasets to be available in FRIS. In other words, only datasets that have all the mandatory fields completed are currently flowing to FRIS. The Flemish RPOs that participated in this survey do not consult FRIS as a source of metadata. However, some institutions use FRIS to check their own records, to get an overview of other research institutes' activities (e.g. number of projects in specific funding programs) or as a source for project partners.

Part 2. Business needs and points of improvement.

2.1 Key deliverable 1: List of prioritised recommendations.

The scope of Key deliverable 1 was to create a prioritised list of recommendations to improve the current flow of dataset metadata. Part 2 of the questionnaire inquired knowledge institutions about their business needs and points of improvement regarding the current flow of dataset metadata in the Flemish research ecosystem. Based on the interview responses to these questions, and the received "As-is" visuals, a list of points of improvements was made (Key deliverable 1). These points of improvement were then translated into recommendations from which a selection of ten recommendations was made, in random order:

1. **Change the status of some of the metadata fields in the FOSB application profile for datasets from 'Mandatory' to 'Recommended' or 'Optional', where appropriate.**
   Description: The FOSB metadata scheme requires more mandatory fields than what is commonly available in most data repositories, this includes metadata fields such as "access rights", "licences" etc. Making some of these fields optional would reduce work and would enable more dataset metadata to be sent to FRIS.
   Impacted organisations: RPOs, Government.
   Drawbacks: Too little info for qualitative monitoring; datasets less findable/usable.
   Here, alignment with the quality requirements of OpenAire and EOSC needs to be kept in mind.

2. **Implement Funder Grant IDs.**
   Description: Introduce universal and persistent grant IDs that can be obtained from a provisioning service. Funders would be able to 'request' such an ID. Once obtained and known, this ID can then be attached to the project and every research entity that has a link with the project, making it possible to uniquely identify the project (and funder).
   Impacted organisations: FWO, VLAIO, FRIS, RPOs[12].

3. **Monitor international developments regarding project PIDs and explore the opportunity to implement Project IDs.**
   Description: Introduce universal and persistent project IDs that can be obtained from a provisioning service. Funders would be able to 'request' such an ID. Once obtained and known, this ID can then be attached to the project and every research entity that has a link with the project, making it possible to uniquely identify the project (and funder).
   Impacted organisations: FWO, VLAIO, FRIS, RPOs[13].
   Drawbacks: Who will govern project IDs given that research projects often concern multiple partners?

[12] This includes the special and industrial research funding (BOF & IOF) managed by the RPOs.
[13] This includes the special and industrial research funding (BOF & IOF) managed by the RPOs.

4. **Use REST or OAI-PMH in addition to SOAP-XML.**
   Description: More easy to use for sending or harvesting metadata.
   Impacted organisations: RPOs, FRIS.
   Drawbacks: Technical effort.

5. **Integrate with external repositories.**
   Description: Institutional repositories should be able to harvest data from external data repositories like Datacite, OpenAire, genomics repositories, Data monitor (PURE), Dataset claimer (RDR), ORCID, ROR, Zenodo, figshare, CORDIS, InChi (chemical compounds), GBIF etc. Additionally, the question is raised whether economies of scale can be obtained if the integrations are clustered by an aggregator.
   Impacted organisations: RPOs.
   Drawbacks: The mapping to those repositories is difficult.

6. **Use FRIS for harvesting.**
   Description: If information on a dataset where several RPOs are involved, is already provided to FRIS by one RPO, the other RPOs can harvest that info from FRIS, enrich it with info that is specific for their institution, and send it back to FRIS. FRIS should then build a single golden record with all compiled info.
   Impacted organisations: RPOs, FRIS.

7. **ORCID integration by RPOs.**
   Description: RPOs can write to and pull from ORCID records that are automatically updated when new publications and datasets are published. This can make several workflows more efficient.
   Impacted organisations: RPOs.
   Drawbacks: costs, requires correct use of ORCID by researchers and integrations with dataset repositories. Legal doubts regarding ORCID and the General data protection regulation (GDPR).

8. **Deduplication of effort/ Use FRIS for harvesting (this was merged with point 6 because of the similar content).**
   Description: Metadata are now registered at different universities and sent to FRIS. Given that researchers in Flanders (often) work together this means that several researchers need to make this effort, while it is aggregated at a later stage anyway. Being able to import relevant metadata (and validate them) would mean a deduplication of efforts.
   Impacted organisations: RPOs, FRIS.

9. **FRIS should pull in more information.**
   Description: FRIS should already collect data which is entered in other systems (OpenAire, ORCID, project IDs, DOI, …), then have research institutes enrich it.
   Impacted organisations: FRIS, RPOs.
   Drawbacks: RPOs still need to (manually) enrich, FRIS system needs to be adapted and extra development required.

10. **Make FRIS Business rules less restrictive (overlap with point 1).**
    Description: Currently, dataset records are not being accepted by FRIS when they fail some of FRIS's business rules (e.g., it is mandatory that records have keywords and abstracts, and that those keywords and abstracts have associated language tags). These metadata fields are not always present or mandatory in the DataCite/Dublin Core metadata. Instead of rejecting such records, they could be accepted and perhaps flagged to indicate that they are not eligible for FRIS's KPI reporting.
    Impacted organisations: RPOs, Government.
    Drawbacks: datasets might no longer be accepted by OpenAire when missing fields result in a lack of quality.

These ten recommendations were then prioritised using a voting exercise that rated the recommendations on two parameters: the *level of easiness* to be implemented and the *degree of impact* on the current ecosystem. In the figures below, easiness is portrayed along the x-axis, with actions that are considered easier to implement receiving a higher easiness score (range: 1-3). Similarly, impact is portrayed along the y-axis, with actions that were considered to have a high positive impact receiving a higher impact score (range 1-3). Taken together, we can easily identify and follow-through with actions which are both relatively easy to implement and likely to have a highly positive impact (i.e., items falling within the green box), while avoiding actions which are likely to be both difficult to implement and have minimal benefit (i.e., items falling within the red box).
Only those recommendations with high impact and low implementation difficulty were considered. According to this analysis the following three recommendations are situated in the green quadrant:

1. Deduplication of effort/use FRIS for harvesting.
2. Integrate with external repositories.
3. ORCID integration by RPOs.

A fourth recommendation was selected as a quick win:

4. Change the status of some of the metadata fields in the FOSB application profile for datasets from 'Mandatory' to 'Recommended' or 'Optional', where appropriate.

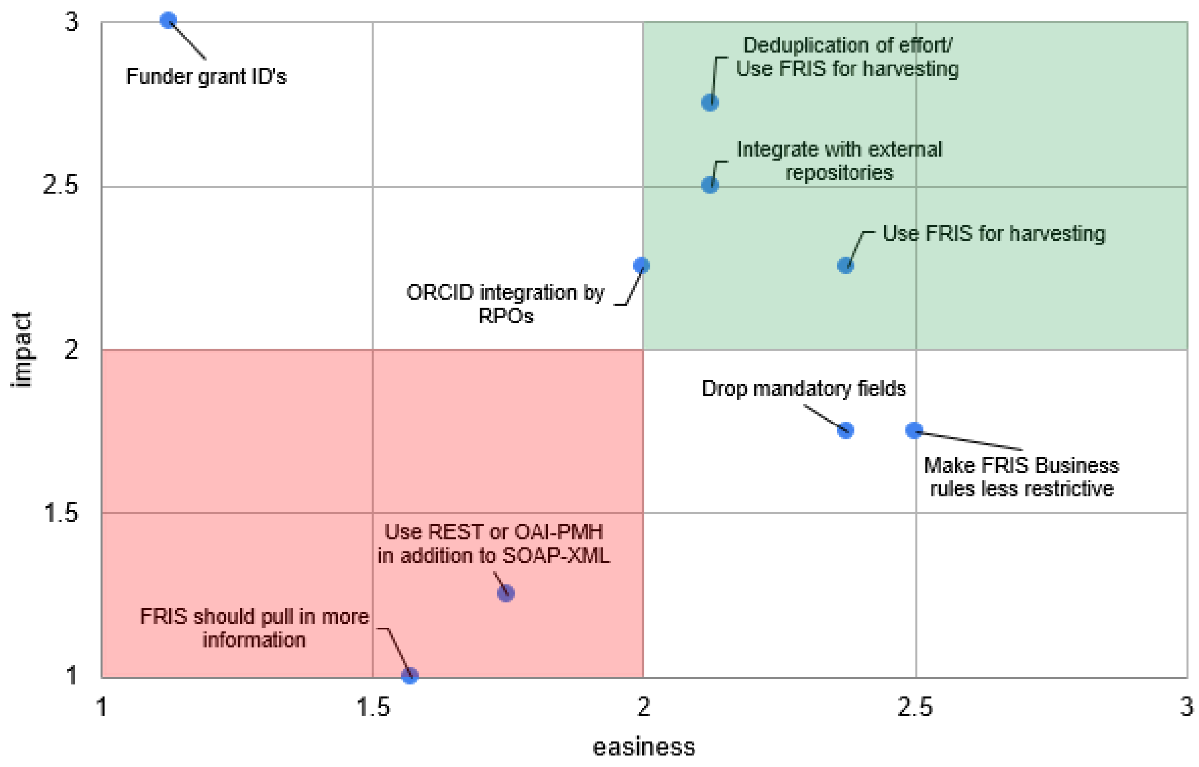


Figure 3.1. Impact versus easiness Matrix: PG Enriching Metadata

It should be mentioned that the highest impact score went to implementing funder grant IDs, though it was not retained in the matrix because of the low score on implementation easiness. However, because of the high impact of this measure, it was decided to formulate a general advice for this recommendation.

The matrix exercise was submitted to the FRDN WG Architecture who repeated the same exercise (Figure 3.2, pp. 15). Here, there were only three recommendations that qualified according to the matrix method:

1. Deduplication of effort/use FRIS for harvesting.
2. Make FRIS Business rules less restrictive.
3. Use FRIS for harvesting (merged with point 1 due to similar content).

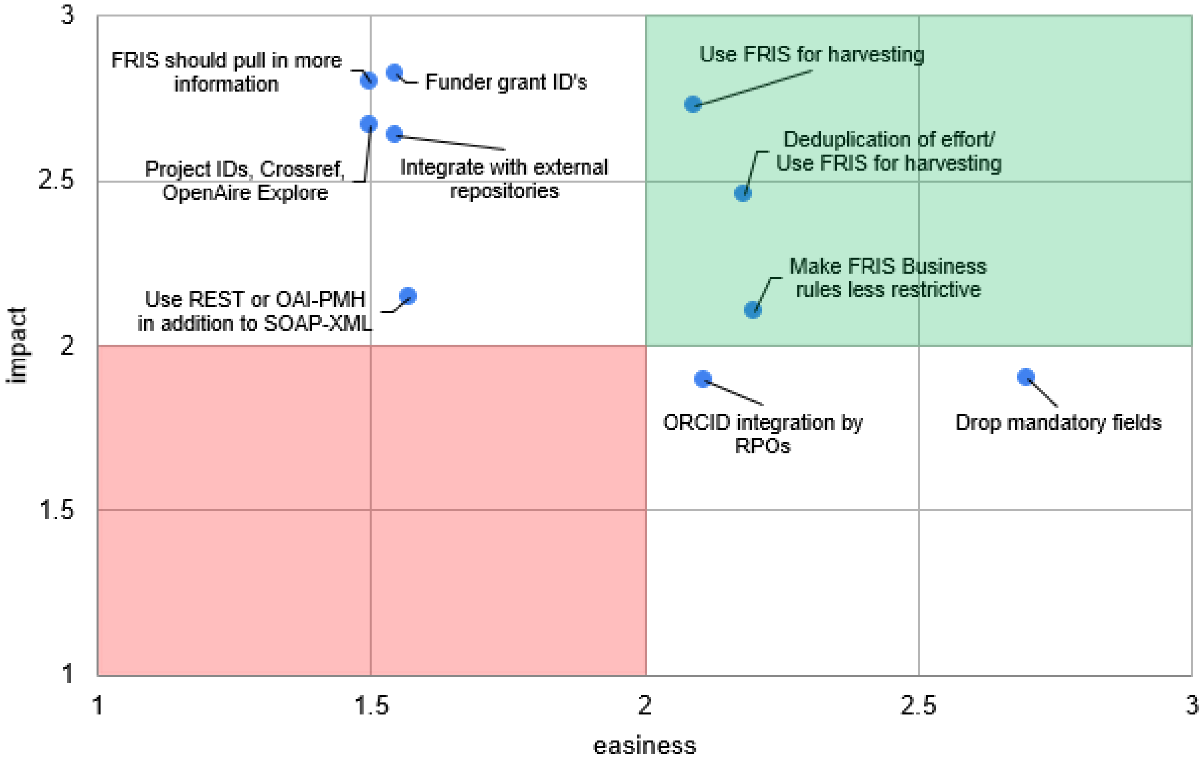


Figure 3.2. Impact versus easiness Matrix: WG Architecture

Here, the recommendation to make the FRIS business rules less restrictive is an extra recommendation discussed and voted by the WG Architecture. However, as for the project group, working group members gave the highest impact score to implementing Funder grant IDs.
Its downside is the similarly high implementation difficulty. Because of the overlap in results between the project group and the working group, we decided to go with the 4 priorities that resulted from the project group's voting exercise (pp.13), while also briefly addressing priorities that have high impact yet low ease of implementation. Below, these recommendations will be further discussed and elaborated.

1. <u>Deduplication of effort/ Use FRIS for harvesting.</u>

There are a few scenarios in which the same metadata are being collected multiple times: first, in the case where a dataset is being created in a collaboration between researchers affiliated with different institutions; second, in the case where a researcher creating a dataset is affiliated with more than one institution at the same time. Or there can be a combination of both scenarios where researchers with multiple affiliations collaborate with researchers from other institutions.

In the case of a researcher having multiple affiliations, the dataset metadata currently has to be registered at each of these knowledge institutions separately and delivered to FRIS multiple times. Afterwards, FRIS will merge the information into a golden record. The effort to collect the metadata could be reduced if institutions that are searching for metadata of a dataset, would check in FRIS if that dataset has already been delivered by another institution. If so, they can harvest all the metadata of that dataset from FRIS, validate it, enrich it with institute-specific details, and send it back to FRIS. This means that it would suffice if the researcher enters the metadata of the dataset in only one of the institutions he/she is affiliated to, and his/her other affiliations could retrieve the metadata via FRIS. The advantage of retrieving this information from FRIS is that it will already contain all the mandatory info that is required to send it to FRIS. For researchers that work at multiple institutions, this way of gathering dataset metadata can significantly reduce the burden of having to input the same metadata manually in the repositories or CRIS-systems of each of these institutions.

In the case of datasets generated by researchers working in multiple institutions, the same holds. Today each institution involved in the creation of the dataset is registering it in his/her own repository and/or CRIS. In order to fetch the metadata or complete missing information, the institutions could harvest it from FRIS when it is already there. Then again, the institutions can validate it, enrich it with institute-specific information, send it back to FRIS, and FRIS should then update the golden record with all compiled metadata. The flow would thus be much more efficient if, for instance, only the main dataset creator would register all of the metadata and provide it to FRIS. Then, the other institutions can harvest that information back from FRIS, enrich it and send it back again to FRIS. This approach would also result in a better flow so that the metadata isn't already duplicated at the start by multiple researchers working at different institutions (differences in the title, order of creators etc.). This flow would also be interesting to import the DMP label for joint projects. The issue of deduplication is not specific to the case of metadata of datasets only. The same issues arise when gathering information on other objects like publications, projects, patents and infrastructure.

2. Integrate with external repositories.

The "enter once, reuse often" mantra, first formulated by ORCID[14], is a principle which both researchers and research support staff advocate to reduce administrative overhead and increase an institution's awareness of its own research outputs. The process of researchers, or research support staff, re-entering the metadata of their datasets and software into their institutional CRIS is tedious and can be subject to error if changes are made between the metadata associated with the original record and those entered into the institutional CRIS. Therefore, in accordance with the "enter once, reuse often" mantra, it would be worthwhile for RPOs, insofar as they are able, to harvest metadata for their associated dataset and software records from the one place where those metadata must necessarily be entered: the datasets and software repositories themselves.

Harvesting the metadata of datasets and software from external repositories is no simple task, its difficulty largely coming from the sheer number of repositories which can be queried (at the time of writing, re3data.org[15] lists more than 3000 different data repositories). Therefore, we cannot advocate that RPOs integrate with available repositories; that would be impossible. Options to overcome this impossibility and improve the metadata collection by RPOs from external repositories are further explored in key deliverable 2. One option that is explored for key deliverable 2 includes that RPOs build integrations with external repositories in a modular fashion. The specific priority of the various repositories can be determined based on a variety of factors, including: frequency of use, availability of APIs, completeness of metadata (and the subsequent likelihood of finding the desired dataset and software records). Furthermore, these integrations would be largely identical between institutions for any given repository, with differences coming only in the precise queries necessary to extract records relevant to the institution in question (see Key Deliverable 2 for more information on fruitful query selection). Therefore integrations can and should be easily shared between institutions, thereby distributing the work necessary and providing further opportunity for support staff from various RPOs to share knowledge on repositories, harvesting best practices, dataset findability, etc.

There are many benefits to a modular approach to repository integration. First, by abandoning an "all-or-none" mentality regarding the harvesting of institutional records, it allows RPOs to immediately make tangible advances in their overview of their own research outputs by

[14] https://www.google.com/url?q=https://info.orcid.org/enter-once-reuse-often/&sa=D&source=docs&ust=1700475282205457&usg=AOvVaw0ZsE4odUMEzTApPgCJ6pM4

[15] https://www.re3data.org/

starting with so-called “easy wins”: repositories whose metadata are easily harvested. Indeed, the code provided as a part of Key Deliverable 2 (which focuses on popular repositories like Zenodo, Figshare, and Dryad) represents such a starting point by which many RPOs could quickly gain a greater insight on their research data and software outputs. Second, a modular approach is not incompatible with other popular methods for identifying institutional data and software outputs, such as using existing aggregator services (e.g., OpenAire, DataCite, Data Monitor). Based on our observations in Key Deliverable 2, these aggregation services often lack completeness, either because known records found through searching the repositories are not present in these aggregators or they are present but lack any metadata linking them to their associated institutions. Therefore, from a modular perspective, these aggregator services can be incorporated as independent modules whose results can be compared/combined with the results of the external repository searches.

Here it is useful to distinguish between two aspects of harvesting: the finding of relevant (i.e., institution-affiliated) records and the subsequent extraction of the metadata of those records. These aspects differ in their difficulties and the considerations necessary for their implementation. Finding relevant records within a given repository can be very difficult and can be very repository- and query-dependent, whereas extracting the metadata of those records (once found) is quite simple. However, issues can arise when a record is found in various locations (e.g., both on the repository page directly and in one or more of the aggregation services) with differences in the metadata. In such cases, how do you decide which set of metadata to take as correct? This is a difficult question, but generally the recommendation of this Project Group is that in such conflicts the metadata from the repository should take precedence as it is the location where researcher(s) would have directly contributed the most metadata and is also the place to which the PIDs will resolve.

3. ORCID integration by RPOs.

Most CRIS systems approach research output (datasets and software) from a publication or record centred point of view. Another option is to approach research output from a researcher centred point of view. Since names cannot be used to uniquely identify researchers (or any person), they are not suitable for dataset discovery. ORCID provides a persistent code for researchers so authors and contributors can be uniquely identified. ORCID also offers users the possibility to maintain an updated digital CV, providing an overview of their contributions to science beyond a simple publications list.

When ORCID records are kept up to date, they can be used for discovery of research output of researchers in RPOs. This does imply that (1) all researchers of RPOs have an ORCID and (2) they keep their ORCID records up to date. If these conditions are met, RPOs can use the ORCID API[16] to find research output linked to their institution. Even when these conditions are met, the implementation is not as straightforward as getting a DOI-list linked to your institution. ORCID expects to be queried based on ORCIDs of individual researchers and returns PIDs of works by the researcher. These PIDs can in turn be queried for metadata on this specific research object in their respective repositories.

Although this method of dataset and software detection is theoretically watertight, assuming accurate ORCID usage by researchers, the conditions that must be met are limiting the use of ORCIDs for discovery. KPIs of ORCID usage in Flanders are on the rise for all RPOs, but the KPI only takes into account whether a researcher has an ORCID or not, it doesn’t say anything about the quality and presence of a complete list of works by a researcher. To incentivize researchers to use ORCID, RPOs could implement a mechanism that feeds the ORCID-record of their researchers with works that are already in the CRIS systems.

[16] https://info.orcid.org/documentation/

4. Change the status of some of the metadata fields of the FOSB application profile for datasets from 'Mandatory' to 'Recommended' or 'Optional', where appropriate. Make FRIS business rules less restrictive.

The interviews revealed that many institutions believe that the FOSB application model for dataset metadata contains too many mandatory fields and that the metadata that is mandatory in the FOSB metadata scheme can be difficult to obtain. There are certain fields that are often not present in repositories or unable to retrieve using the API or another machine-readable interface. This causes issues when trying to automate and streamline the registration of datasets' metadata for FRIS, as it often requires manual enrichment to supply all mandatory fields. Some examples of frequently missing or unavailable metadata fields are; the composite *access rights field* (especially the legitimate opt-out and embargo date if applicable are rarely available), the possible *links to project and/or publications* are also frequently missing from machine-readable end-points. Limiting mandatory fields to those of the DataCite model would already significantly reduce the workload of manual curation and enrichment.

It's important to note, however, that the mandatory fields in the FOSB metadata model for datasets is so expansive due to them being used to calculate the open data KPI within the framework of the FOSB Open Science monitoring. Therefore a simple change of the status to optional or recommended would mean that measurements are reduced to those datasets that do have this information. A more general drawback is that less information means less quality and less findability.

A proposed suggestion is to put the responsibility with the institutions to make their open data KPI as complete as possible by supplying the KPI-relevant fields if they were changed to no longer be mandatory for all datasets delivered to FRIS. In other words, a dataset would only be included in the RPOs open data KPI calculation if the RPO has delivered the necessary metadata fields. In general, to promote the FAIR principles, more metadata is better, as it makes the data more findable in search indexes and portals such as FRIS. So, delivering as much metadata as possible would continue to be encouraged, with the idea being that it's better to have some metadata by decreasing the number of mandatory fields than having no metadata at all, but more metadata is always preferred.

Flemish strategy for Funder Grant Identifiers and/or Project identifiers.

During the interviews on points of improvement and business needs, it was clearly stated by all respondents that there is a need for a coherent PID strategy (funder grant IDs, project IDs,...) in the Flemish research landscape. In order to unambiguously link research outputs such as datasets with the projects they arise from, it is required to integrate funder grant, and project PIDs into existing processes and workflows. Establishing a Flemish research PID-policy in line with what is internationally developed will help to identify and manage duplicates, make harvesting more efficient, and improve existing workflows and data exchange. Capturing persistent identifiers as of the first deposit in repositories has many advantages such as avoiding double effort and duplicates, more efficient harvesting of metadata, improvement of existing workflows and automated data exchange.

Persistent identifiers (PIDs) have become an integral part of the digital open science research landscape since aggregated PID links form scientific knowledge graphs etc.[17]. PIDs are globally unique and sustained references to digital research objects (publications, datasets, persons etc., UK PID

[17] French National Plan for Open Science, 2018, p8. Available online at: https://libereurope.eu/wp-content/uploads/2018/07/SO_A4_2018_05-EN_print.pdf

Policy Roadmap)[18]. PIDs offer unambiguous and stable ways of identifying and referencing research entities and their associated (meta)data, and thus provide trusted connections between researchers, research institutions, and outputs (UK PID Policy Roadmap). PIDs contribute to the sustainability and trustworthiness of research information systems, facilitate interoperable exchange of information, help to prevent data loss, and provide the building blocks for knowledge graphs. PID integration into existing workflows, systems and services also allows for automation of administrative workflows, for instance for reporting-and evaluation purposes (PID-optimised workflows, UK PID Policy Roadmap, pp. 3). The JISC report 'Developing a persistent identifier roadmap for open access to UK research'[19] recommends the use of the following 5 key PIDs: ORCID IDs for researchers, Crossref and DataCite DOIs for publications and datasets, Crossref grant DOIs for grants, ROR IDs for organisations, and RAiDs for projects. This report identified project and grant PIDs as high-priority missing identifiers. Subsequently, a JISC project 'UK PIDs for Open Access'[20] identified that the use of grant IDs and project IDs ensures that the process from application to evaluation of an award is optimised, with automatic exchange of PIDs and related metadata between funders, RPOs and service providers/registries (ORCID etc.)[21].

Funder IDs.

CrossRef offers an Open Funder Registry[22] where research funding organisations can register a PID for their organisation (a DOI) with associated metadata that is publicly discoverable. These PIDs are used in funding acknowledgements in the metadata of articles and in ORCID funding sections. The use of trusted identifiers for research funding organisations makes it possible to unambiguously link funders to the projects they fund and the outputs that flow from them. This enables funders to pull in information for reporting and evaluation purposes, and track the impact of their funding. For RPOs the use of funder IDs facilitates data exchange and the management of duplicates.

Grant IDs.

Funders have long been using internal grant numbers for administrative purposes, however, local identifiers are not globally exchangeable since they are not persistent nor unique[23]. Using PIDs for grants offers trusted and automated links between grants and their associated researchers, research institutions, outputs etc. Grant PIDs facilitate reporting and evaluation workflows such as the tracking and monitoring of compliance with funder policies, the impact of funded projects and collaborations. As of 2019, CrossRef provides a Grant ID service enabling funders to register metadata for their grants. Funders can become a member of CrossRef to register new grants (Crossref Grant ID) using an API or web form (UK PID Policy roadmap pp. 12). CrossRef Grant IDs consist of a DOI and attached metadata that are publicly discoverable via APIs, search interfaces and tools.

Project IDs.

Currently, there are no globally unique trusted identifiers for research projects that are widely adopted. Project PIDs identify research projects and have associated metadata to describe their attributes and relationships[24]. One example is the *Research Activity Identifier* (RAiD), maintained by

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

[19] https://repository.jisc.ac.uk/7840/2/PID_roadmap_for_open_access_to_UK_research.pdf

[20] https://scholarlykitchen.sspnet.org/2020/06/29/the-uk-national-pid-consortium-a-pathway-to-increased-adoption/

[21] https://resources.morebrains.coop/jisc-workflows/jisc-funding-workflow/

[22] https://www.crossref.org/services/funder-registry/

[23] https://scholarlycommunications.jiscinvolve.org/wp/2020/10/12/theres-a-pid-for-that-part-1-grants/

[24]https://scholarlycommunications.jiscinvolve.org/wp/2020/10/13/theres-a-pid-for-that-part-2-projects/

the Australian Research Data Commons (ARDC)[25], that captures research objects, relations and roles connected to research projects. In the UK, a Project ID focus group was formed (consisting of university librarians, PID experts, research managers, and infrastructure providers) who recommended the use of RAiD for project identifiers[26].

The conclusion here is that the option should be explored to establish an FRDN working group to recommend a Flemish PID strategy. This working group can start with desk research to map the current European PID landscape by reviewing the PID roadmaps, policies and projects from other countries: for example the JISC project UK PIDs for OA[27], the Dutch NWO PID guide[28], the Irish National Open Research Forum (NORF) PID project[29] and roadmap, the PID Network Germany[30], the Finland PID roadmap[31], and the Austrian FWF notes on project data[32]. A good starting point would be to check out the outputs of the Research Data Alliance (RDA) National PID Strategies Working Group that has made a comparative analysis of national PID strategies resulting in a comparison guide and checklist, the RDA guide for national PID roadmaps[33], that can guide users in developing national PID strategies. A project group analysing a Flemish PID strategy should investigate the difference between funder grant IDs and project IDs, the possible merit of using both, and for project IDs specifically: the challenge of ownership (Who are the owners of project information and what are their incentives for managing RAiDs?).

Enhancing interoperability : introducing REST-JSON or OAI-PMH in addition to SOAP-XML.

In the current framework, all services within FRIS are developed using the SOAP-XML protocol. Both data delivery to FRIS and data retrieval from FRIS are executed through SOAP-XML, adhering to the CERIF 1.5[34] standard in XML format. Recognizing the suggestions given, there is a case to diversify and supplement SOAP-XML with REST-JSON or OAI-PMH.

1. REST Implementation for efficiency and flexibility:

- Given that SOAP-XML is an older standard, introducing a new version utilising a REST service with a JSON format has the potential to significantly reduce development time for both data providers and consumers interfacing with FRIS.
- The CERIF 1.5 standard, rooted in XML, may necessitate reconsideration if a REST interface with JSON is adopted. Exploration of a later version of CERIF, currently under development to incorporate a JSON format, could be a viable solution.

2. OAI-PMH for widened interoperability:
   Despite being an established technology, the adoption of OAI-PMH offers an additional avenue for interoperability. This recognized technology for information exchange between repositories, utilising XML, is integrated into FRIS. OpenAire and EOSC currently utilise the OAI-PMH endpoint for harvesting datasets from FRIS.

[25] https://ardc.edu.au/

[26] https://scholarlycommunications.jiscinvolve.org/wp/2020/10/13/theres-a-pid-for-that-part-2-projects/

[27] https://scholarlykitchen.sspnet.org/2020/06/29/the-uk-national-pid-consortium-a-pathway-to-increased-adoption/

[28] https://zenodo.org/records/4674513

[29] https://norf.ie/pid-roadmap/

[30] https://www.pid-network.de/en/news/blog/use-cases-nationalen-pid-strategien

[31] https://zenodo.org/records/5024228

[32] https://www.fwf.ac.at/en/discover/notes-on-project-data

[33] https://www.rd-alliance.org/group/national-pid-strategies-wg/outcomes/rda-national-pid-strategies-guide-and-checklist

[34] https://eurocris.org/cerif/feature-tour/cerif-15

In conclusion, diversifying the service protocols within FRIS to include REST and OAI-PMH, in addition to the existing SOAP-XML, can enhance the platform's compatibility with diverse data providers and consumers.

3.1 To-be Business Flow.

To design the ideal business flow, respondents indicated that the following points must be met to achieve the ideal flow:

- The flow is invisible to researchers, they just see their research output automatically distributed to relevant sources without requiring any action from the researcher.
- Only once principle, the researcher enters every piece of information only once.
- A trust between parties is established to guarantee the metadata are correct.

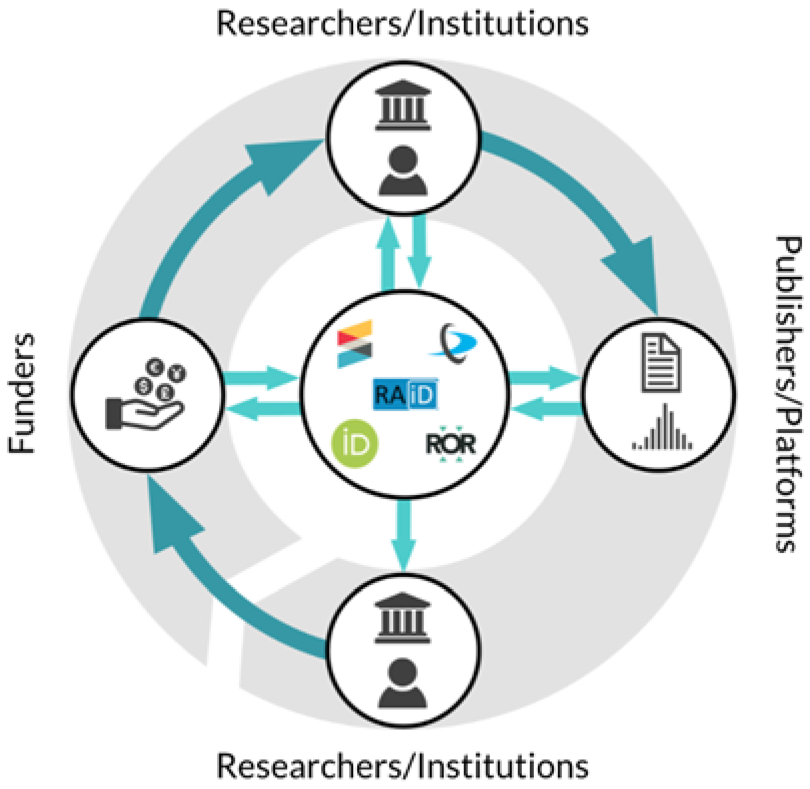


Figure 4. MoreBrains Visualisation of the PID-optimised research cycle.[35]

But to realise this ideal data flow, a number of conditions must be met:

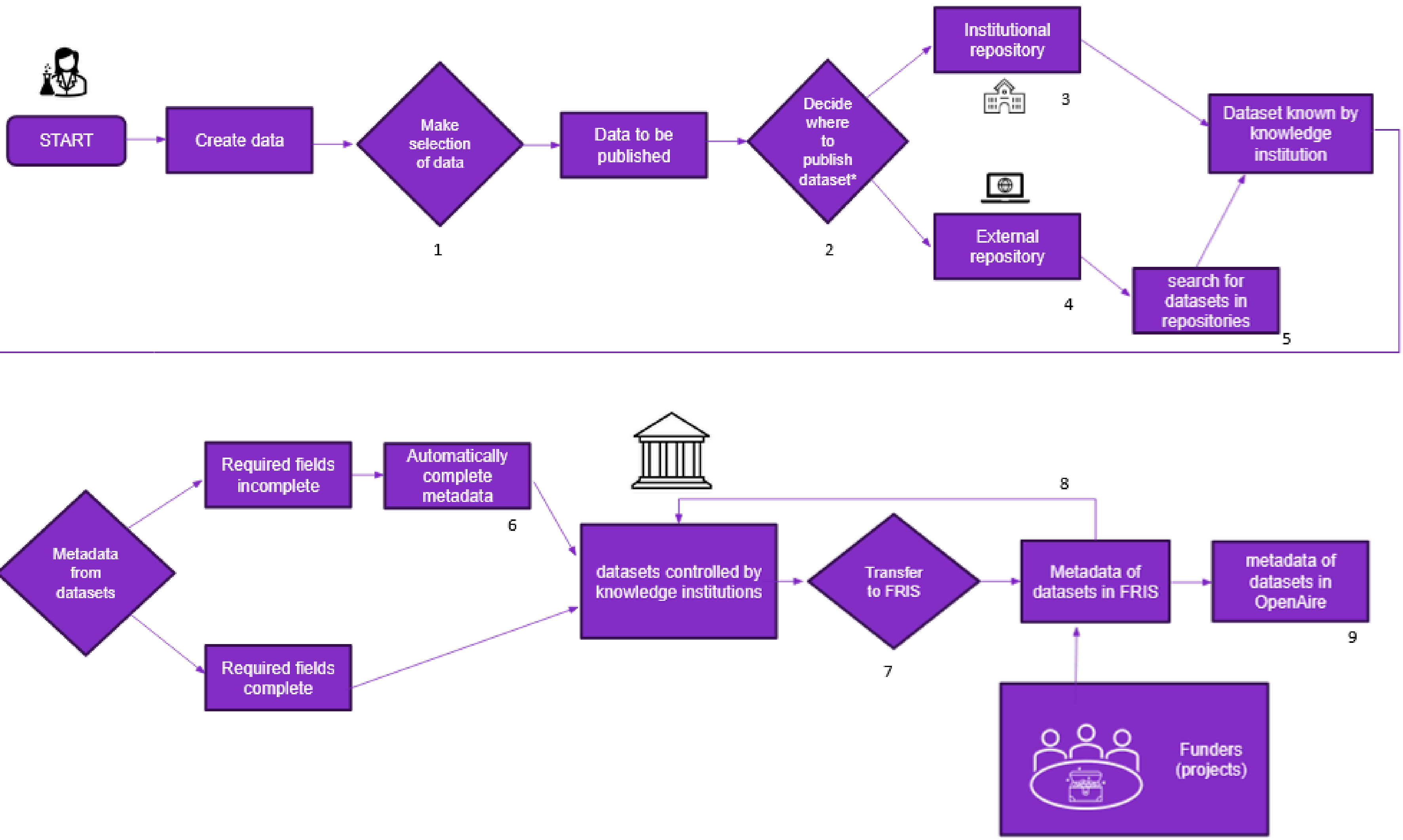
- All parties must use trusted and agreed upon PIDs.
- A RACI of all actors must be made and agreed upon.
- Masterdata must be determined (if a metadata field has two different values, what is the correct value?).
- Who is responsible for merging different records to one? (current problem in FRIS: two RPOs enter the same data for a project they are working on together).
- All researchers must register their datasets using ORCIDs and other PIDs.

[35] https://doi.org/10.5281/zenodo.4991733

- Validation and enrichment of externally discovered dataset metadata by researchers (furthermore, this information should flow through to the external sources).
- Metadata schemes of repositories should be aligned.

Since prerequisites of the ideal flow are not yet met, the project group decided to work on a flow that can be realised within the current landscape. A business version of this flow can be found on the next page (figure 5).

Figure 5. To-be business flow


START
Create data
Make selection of data
1
Data to be published
Decide where to publish dataset*
2
Institutional repository
3
External repository
4
Dataset known by knowledge institution
search for datasets in repositories
5
Metadata from datasets
Required fields incomplete
Automatically complete metadata
6
Required fields complete
datasets controlled by knowledge institutions
Transfer to FRIS
7
8
Metadata of datasets in FRIS
metadata of datasets in OpenAire
9
Funders (projects)

Comments on the business flow:

1. Researchers might not want to publish all data gathered (e.g. raw data) during their research.
2. Dataset = data + metadata + documentation.
3. Controlled environment (institutional repository/FAIR vault) with possible necessary restrictions and ideally curated before publication (curation activities can happen in the repository).
4. Accessible environment (external repository) with possible necessary restrictions and ideally curated before publication (curation activities can happen in the repository).
    a. Which repositories or aggregators to search? There are over 3000 repositories registered in re3data covering all academic disciplines. Possible strategy to cope with this, will be tackled in Deliverable 2.
5. Ideally, incomplete metadata is automatically enriched with information from other sources. Often this does not suffice and manual enrichment is needed.
6. Metadata delivery to FRIS requires datasets to be linked to publications and projects.
   This information can only be provided by the researcher.
7. When RPOs collaborate they both want the datasets linked to the collaborative research projects in their CRIS system. When the metadata is added in FRIS, RPOs can enrich their metadata by ingesting missing fields from FRIS. There is however an issue here: what if both RPOs have different sets of metadata, who is the master?
8. Although most universities already send their metadata to OpenAire, reporting and registering in OpenAire will be required more often by European funders, this part will become more relevant in the future. FRIS already functions as a hub towards OpenAire.

3.2 Analysis of AS-IS and TO-BE Business Flow: Key Differences.

The current representation of the AS-IS and TO-BE business flows in separate diagrams has impeded a clear identification of crucial disparities. Highlighted below are the most significant distinctions and remarks in no particular order:

1. Enhanced Project Information Integration from Funders into FRIS:
   The proposed TO-BE business flow emphasises the imperative of incorporating project and funding information from more funders into FRIS.
2. Establishment of Feedback Loop from FRIS to Institutions:
   Presently, FRIS is underutilised as a source of information for institutions. The TO-BE model introduces a feedback loop to ensure a continuous exchange of information between FRIS and the participating institutions.
3. Heterogeneous Implementation Across Institutions:
   The participating institutions exhibit a high degree of heterogeneity in adopting the TO-BE business flow. Each institution faces unique challenges, resulting in varied levels of progress.
    a. Initiation of Systems/Protocols from scratch: Certain institutions have no systems or protocols in place.
    b. Partial implementation across institutions: While some institutions have implemented specific steps of the TO-BE business flow, none have executed all the prescribed measures. This diversity underscores the ongoing evolution and adaptation within the participating entities.

Note: The TO-BE business flow presented herein is not perceived as an ideal state but rather as a practical and achievable framework to aspire towards.

Key deliverable 2: Solution for collecting metadata.

4.1 Harvesting dataset metadata from external repositories: challenges.

In part 1 of the questionnaire, the knowledge institutions were asked about any problems they encounter with harvesting dataset metadata from external repositories. The results showed that metadata harvesting from external sources currently comes with a lot of challenges. First, research institutions struggle to find their own datasets from affiliated researchers because many external repositories use metadata standards with a low degree of mandatory attributes (e.g. DataCite, DublinCore) resulting in incomplete information. Metadata fields that are important for discovery - such as ORCID, authors' affiliations or institutional ROR- are often not mandatory and standardised in many repositories. For example, some repositories still allow author registration with name instead of a PID such as ORCID. Hence, researchers may be wrongly affiliated leading to extra or missing datasets (e.g., Vrije Universiteit Brussel wrongly referred to as Free University Brussels).
This challenge has been described in detail in the paper 'DataCite Dataset Metadata: a Flemish Case Study'[36]. Other attributes such as subject, format, description, access rights and links between datasets and publications or projects are also not mandatory according to the DataCite metadata standard. This hinders the discoverability of dataset metadata and requires time and effort to manually validate and enrich the metadata in order to align with the FOSB application profile for datasets.

Furthermore, the quality of the metadata is typically uncurated by the repositories resulting in missing, inaccurate or unclear information that inhibits complete automatic harvesting. Since many dataset repositories cannot guarantee high levels of quality control, there is a considerable amount of pollution in the current data ecosystem. It is common for files, such as images, supplementary materials and collections, to be registered with a DOI as if they were datasets. For these reasons, harvesting proactively from external sources using APIs (e.g., WoS Data Citation Index API, OpenAire API, DataCite API) thus results in either too many results to weed through, or too little to have any use. Therefore, most research institutions mainly rely on importing metadata manually either using specific DOIs or search queries via SoP (Standard operating Procedure).

Moreover, the different metadata schemes that are currently in place within the data repositories are not always used consistently. For example, the *DataCite "Rights" field* is not mandatory and may contain either access level information or licence information. Licensing information is not harmonised across repositories either and does not necessarily follow the OpenAire guidelines and SPDX licence identifiers as requested by FOSB. Hence, metadata harvested through various repositories might differ slightly (e.g. licence naming conventions) and some information might be missing (e.g. formats). Also, regarding linking datasets to projects and other research output (publications, patents etc.), the researchers are the main source of enrichment since they currently hold this information.

An additional issue is that references to specific funder programs or funder/project ID numbers are often non-existent in dataset metadata. The funder and project metadata fields are currently open and not globally standardised. Similarly, when datasets with other PIDs than DOIs will be registered, it will be even more challenging to automatically prefill the dataset record based on harvested metadata, since other types of dataset PIDs (Handle, accession number etc.) do not have

[36] Van Wettere, N., 2021. Affiliation Information in DataCite Dataset Metadata: a Flemish Case Study. *Data Science Journal*, 20(1), p.13.DOI: https://doi.org/10.5334/dsj-2021-013

an obligation to register certain metadata.

4.2 Analysis and results.

Overview of available technologies.

In order to get an insight into possible solutions for the registration and finding of externally published datasets, we decided to first take a look at the available technologies/metadata sources that are available in the current RDM ecosystem. In order to properly analyse our options, we first made an overview of the available technologies/sources, then we compared the metadata models of these selected technologies/sources against the minimal requirements set-out by a previous analysis of this project group in relation to the FOSB metadata model for datasets. Based on these two steps, we could then move on to explore how the investigated technologies/sources could be used and to do a feasibility analysis of our initial ideas and brainstorm possible solutions.

List available relevant technologies and make a shortlist.

To get an overview of the available and relevant technologies/sources, we did a brainstorm with the project group members and some research into what was out there when it comes to metadata aggregators, repositories, and PID services that register metadata to name a few examples. We listed sources with and without APIs to ensure we did not limit our scope to just the API accessible sources. An overview of the reviewed sources can be found in appendix Chart A1.

For each technology/source we researched the domains covered by the technology/source, the estimated number of datasets in the source at that time (Summer 2023) and, where available, estimated the number of datasets affiliated to a selection of eight Flemish research institutions found in the source at that time (Summer 2023).

It is worthwhile to note that we based our selection of services to free and open-source services. For example, we excluded from our analyses such services as Elsevier's Data Monitor[37], as the results from the interviews with Flemish research support staff indicated that there was a preference toward avoiding paid services which would increase reliance on large publishing companies. Instead, in the spirit of open science, we focused on existing technologies, sources, and services which had free and open functionality.

[37] https://www.elsevier.com/products/data-monitor

Review and test selected metadata source (technologies) against the minimal requirement of 2.1.2. Assess and make an overview of strengths and weaknesses.

To assess each selected technology/source against the previously established minimal requirements, we listed for each of the metadata elements in the FOSB model the likelihood of availability in each technology/source. The possible values were; 0 for elements that are never available in the technology/source for a dataset, 1 for elements that are always available in the technology/source for a dataset (typically these are the mandatory fields of the technology/source), and 0.5 for elements that have a likelihood of being available, but no certainty (typically these are the mandatory if available or optional fields of the technology/source). The scoring of the different technologies/sources based on this method and the estimates from 3.a. were compared to each other to discover overlap and possible unique fields only available in certain sources.

We found that there is considerable overlap, though there are also a lot of unique fields that only certain sources can provide. From this comparison, the suggestion arose to create two categories: "discovery sources" that are used for finding datasets, their IDs and the metadata's initial entry point, and "enrichment sources" that can be used to enrich the discovered metadata.
Two pieces of information that are unlikely to be in any of the discovery sources are: link to project and link to publication.

A short list was created of sources that could serve as discovery sources and sources that were possibly better suited for the enrichment phase. In total, we decided on OpenAire, ORCID, DataCite, and a series of repositories to further investigate as discovery sources. For the enrichment sources, we mainly focused on the repositories themselves as they are the source of all available metadata in the UI. They are also included in the investigation of the discovery sources to see if they can be used effectively in a direct way.

Feasibility analysis of ideas. Possible solutions brainstorm & analysis.

With the selection of sources decided on, the next step is to check the feasibility of actually using them as discovery sources. To do so, we did a series of test queries for the selected discovery sources.

For each of the sources, a predetermined set of queries were compared: the text string "Belgium", the names of five Flemish research institutions, the corresponding RORs of those institutions, ORCIDs of representative researchers from each of the five selected institutions, the Full Names of those selected researchers, and the Full Names of the selected researchers in combination with the names of their respective Institutions. Five people per institution were selected from FRIS to form our sample. For each data source (DataCite, OpenAire, ORCID, Dryad, Zenodo, Figshare, Pangaea), the same set of queries were conducted. The results were exported and aggregated for analysis (aggregation can be found in: Aggregation20230830-V01.2 (repos separated).xlsx). It was immediately apparent that the "Belgium'' query was inappropriate and should be excluded, as it generated very large numbers of poor-quality results.

An overview was made for each identified DOI and with what combination of source and query the DOI was found. We limited our search to DOIs as that was the best way to uniquely identify datasets across sources. Note that the "Belgium" query was largely excluded, as we quickly discovered its scope as too broad (it included both datasets generated by researchers at Belgian universities, but also datasets about Belgium) and the number of results were too high to properly parse during this test.

It was of course possible for any one DOI to be discovered by several different sources or queries (e.g., a record in Zenodo might be found alternatively by querying Zenodo with the name of the institution or by querying OpenAire with that institution's ROR). Therefore, the next step was to identify the degree of overlap between the various sources and queries to determine whether there exists an efficient set of sources and queries that could be utilised to maximise coverage (i.e., find the greatest total number of unique DOIs) while minimising overlap (i.e., reduce the number of times the same DOI is found using separate sources/queries). The overlap of the different sources for the discovered DOIs was investigated using UpSet (https://gehlenborglab.shinyapps.io/upsetr/), which functions as a more intricate Venn Diagram and allowed us to visually represent the sets of DOIs and where they are found.

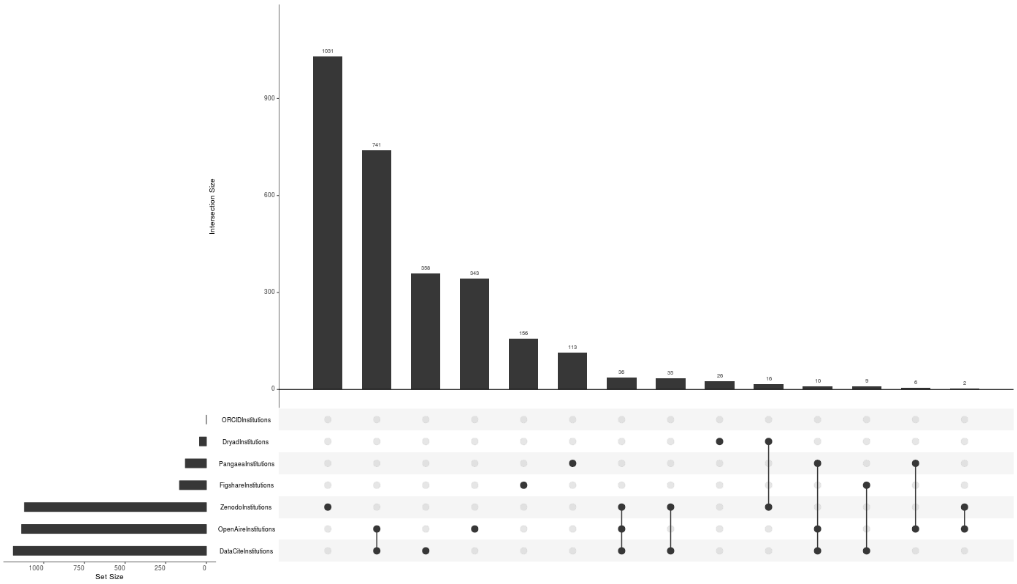


Figure 6. UpSet graph: results querying sources with institutional names.

The above UpSet graph shows the results of querying the various sources with the names of the five chosen institutions. We will explain briefly how to read an UpSet graph. UpSet graphs are composed of three components: a histogram in the lower left, a larger “main” histogram, and a matrix of dots between the two. The histogram in the lower left (next to the list of names) gives the total number of DOIs found by querying each source (e.g., the largest bar next to the DataCite entry indicates that >1000 records were found when querying DataCite with the institution names). The large, main histogram then gives the *sets* of DOIs (in descending order), wherein the definition of sets is described by the matrix of dots below. For example, the leftmost column indicates that 1031 records were *uniquely* in Zenodo, whereas the next column indicates that 741 records were found in both DataCite and OpenAire.

As can be seen from the UpSet graphs, there was no single source (or source + query combination) that found all - or even most - DOIs (UpSetR Graphs folder with the results.) as evidenced by the lack of considerable overlap. Instead, large numbers of records were found in specific single sources (and no other). Although the existing aggregator services (OpenAire and the DataCite API) often found the highest numbers of datasets, they inevitably missed large numbers of datasets that were found by querying the repositories themselves. Often, this is not because these records are absent from the OpenAire or DataCite databases, but rather because the metadata availability and search functionality of these aggregators often do not link these records to their associated institutions or authors in easily transversable ways. For example, records found “exclusively” by querying Zenodo could then subsequently be found in DataCite by querying DataCite with the DOI found in Zenodo. Therefore, these services remain useful for metadata validation and augmentation, although their use for record *findability* is limited. Nevertheless, this highlights the importance of determining the best queries for each source. A frequency table of the found DOI’s can be found in Chart A2 in the appendix. From querying these sources and comparing their results and coverage, some additional noteworthy discoveries were made:

1) We found that the ORCID API is not as straight-forward to use for the discovery of datasets as expected, as you can only query for outputs based on ORCID. If you query on

e.g., institution or name, you get the ORCIDs of researchers affiliated with those institutions as the only available response, not the related outputs of those researchers. Furthermore, in order for an ORCID-based approach to work, researchers must first link their dataset records to their ORCID account. Such an intermediate step does not align with the "Enter Once, Reuse Often" principle, and therefore it would be better to find a method which does not require additional work on behalf of the researcher.

2) The overlap between Zenodo and Datacite on the institution searchers was minimal. This was unexpected as all Zenodo DOI's have a DOI and their registration and metadata should therefore be findable in DataCite. We suspect this inconsistency might be due to the aforementioned different technical search set-ups/index technologies in Zenodo and DataCite. (E.g., an affiliation search requires the ROR in DataCite.)

3) Zenodo provided many results for the institution searches without overlap with other sources. This implies that there are datasets which are findable in Zenodo that aren't findable anywhere else based on the same search.

4) The general overlap between the different sources was less than expected.

5) Even the smaller sources have data that the bigger sources/aggregators don't have. Which indicates their possible added value to include in the discovery process.

6) In DataCite, the overlap between results of the ROR searches and institutions searches was minimal. This was unexpected. It points to a technical search set-up that needs to be kept in mind when using DataCite to query on institutions.

8) The ROR search results from the repositories were completely captured in the DataCite ROR search results.

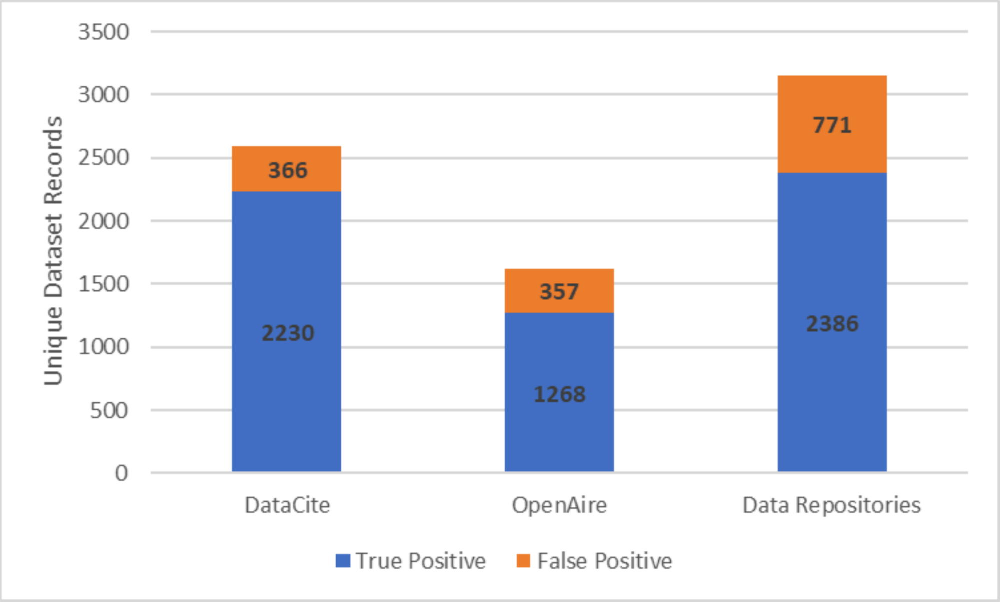


Figure 7. Accuracy analysis of various data sources based on query experiments.

The next step in our feasibility analysis was to perform a preliminary analysis to evaluate the *accuracy* of our various source+query combinations. The goal was to determine what proportion of the found DOIs represented false positives (i.e., were not actually from the authors/institutions being examined). We queried DataCite with the DOI of each found DOI and searched the metadata for mention of a Flemish RPO. Any record without mention of a Flemish RPO was considered a false positive, making this a strongly conservative test of accuracy. As can be seen in the above graph, the number of false positives remained relatively low (from 14% in DataCite up to 25% in the data repositories). Moreover, the process of eliminating false positives was neither difficult nor time-consuming. Therefore, the presence of false positives should not be considered an impediment to implement such automatic dataset identification strategies - as long as a similar process is also implemented to remove false positives.

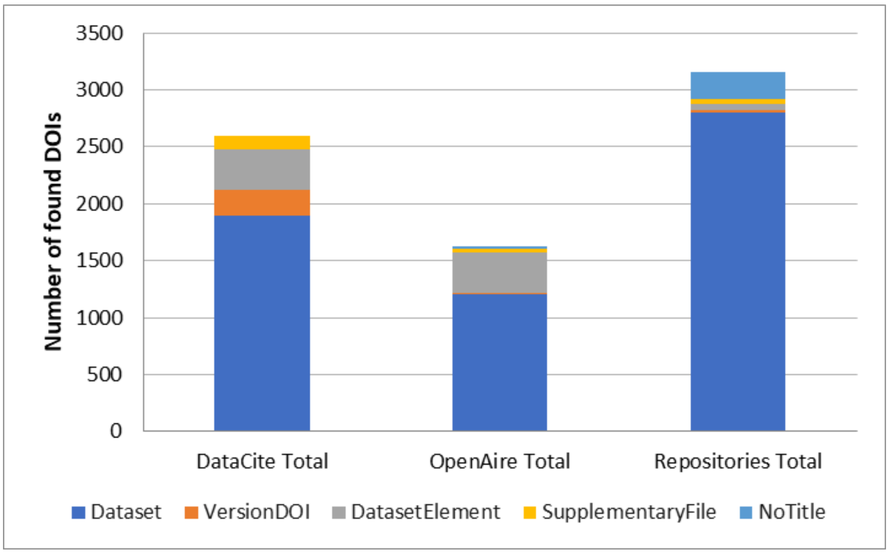


Figure 8. Comparison of number of actual dataset DOIs found (dark blue) and reasons for exclusion from sample set.

Another perspective from which to gauge the overall *accuracy* of the search results for each source and query was to make an assessment of how likely it was that each found DOI points to a unique dataset. As it stands, a number of the result DOIs were found to point to files within a dataset, instead of the dataset itself, to a version of a dataset for which there was already another DOI pointing to the latest version of the dataset's landing page and DOIs leading to datasets that are named "supplementary files". This analysis was made based on the structure of the DOI itself and on the resulting metadata. It is important to note that this method isn't watertight and just gives an overall picture of the general distribution of what kind of pages/datasets the found DOIs resolve to. From the above graph, it's clear that overall, the majority of the found DOIs lead to unique datasets in the most strict definition of the word for each source queried, with the repositories giving the highest percentage of 'strict' datasets.

The above analyses lead us to the conclusion that there is no single source+query combination that is capable of finding even the majority of relevant datasets. It is possible that as existing aggregation services improve their coverage (or new aggregation services are developed), there may emerge an obvious, single, comprehensive source for dataset metadata. However, until such a service emerges, we advocate for a modular approach when trying to query and harvest metadata, wherein specific API integrations with various sources are built individually on an as-needs basis. Casting as big a net as possible will maximise the number of dataset records that are captured. Although this will necessarily produce some overlap as some DOIs are found via different sources, it is trivially easy to deduplicate using PIDs. Additionally, as mentioned in Part 2.1.2, this modular method allows efforts to be concentrated on the repositories which are thought to contain the most dataset records. The main takeaways from this initial feasibility analysis that explored a series of possible metadata sources are that:

(1) The query selection within a source has a big impact on the results, depending on the source and its querying options, certain query types are more successful than others, with the success rate differing between the different sources.
(2) The indexing systems of the different metadata sources influenced the search results in an unexpected way. Because these indexing systems aren’t transparently documented or explained, it’s difficult to predict what you’ll get from the same search query in a different source/system.
(3) There is no source that is able to find all datasets, there are always unique datasets found in each source.
(4) The overlap between sources and the respective queries was less than expected, therefore ruling out the option to choose just one source to find them all.

These conclusions are important to take into account when brainstorming possible solutions to propose. An expected solution at the start of this analysis was that there would be a source that stood out as being the master-source, the one source to find them all, both finding the majority of the datasets and being able to supply most if not all of the necessary metadata. However, the analysis has shown that there is no such source at this moment. This leads to the conclusion that a more modular approach would be better suited for discovering datasets and enriching found datasets. With this in mind, we continued on with a brainstorm of possible solutions, of which three were chosen to further explore with the above established knowledge of the landscape in mind.

Key deliverable 2: Proof of Concept: proposed solutions.

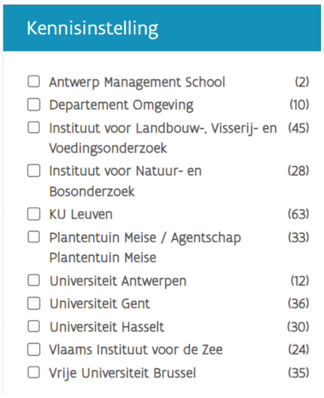
From the above analysis, we came to the conclusion that there were three feasible alternatives for how we could proceed, with each representing a different degree of commitment by the involved institutions (and a concomitant number of datasets found in response). We refer to these three options as:

- The Modular Approach - As described above, the modular approach involves iteratively querying various sources (aggregators, repositories, etc.)
- The Registration Approach - Already employed by several RPOs (e.g., UGent, KU Leuven), the registration approach involves researchers entering DOIs (or other PIDs) into a system which then automatically harvests the relevant metadata from the appropriate repository.
- The Null Approach - The situation as it is currently, without modifications. This serves as a floor value for the amount of resources invested by the RPOs and the number of datasets that make it through to FRIS against which we can compare the above two models.

We compared the quantity and breadth of coverage of the above three methods. To provide a suitable basis of comparison, we chose to compare datasets found for 2022, as at the time of writing (December 2023) the number of records already in FRIS (i.e., those acquired through the null approach) would have likely reached their ceiling and would be unlikely to change.

The Null Approach

At the time of writing, there are 320 datasets that were published in 2022 and subsequently registered in FRIS, and these datasets are distributed across 11 different RPOs. Of these 320 datasets, 180 have DOIs, whereas the remainder have only some alternate form of DOI. Therefore, the modular approach (which at present searches only in repositories which use DOIs) would not be able to find the 140 non-DOI datasets.



Kennisinstelling

□ Antwerp Management School	(2)
□ Departement Omgeving	(10)
□ Instituut voor Landbouw-, Visserij- en Voedingsonderzoek	(45)
□ Instituut voor Natuur- en Bosonderzoek	(28)
□ KU Leuven	(63)
□ Plantentuin Meise / Agentschap Plantentuin Meise	(33)
□ Universiteit Antwerpen	(12)
□ Universiteit Gent	(36)
□ Universiteit Hasselt	(30)
□ Vlaams Instituut voor de Zee	(24)
□ Vrije Universiteit Brussel	(35)

Figure 9. Datasets from Flemish RPOs registered in FRIS for 2022.

The null approach should not only be evaluated by the amount of datasets successfully registered, but also by the amount of work necessary to register those datasets. As identified in the surveys of various Flemish RPOs, the process of dataset registration is a laborious process which often involves a high degree of manual data entry. Thus, the null approach should not be considered as a "do nothing" approach, which would disregard the significant amount of work necessary to maintain the current situation.

The Modular Approach

To compare with the null approach, we queried five major data repositories (Dryad, Figshare, GBIF, PANGAEA, and Zenodo) and one major data aggregator (DataCite Commons) for datasets that were published in 2022 by the RPOs that already have datasets registered in FRIS for that time period. In total, this approach found 1033 DOIs across the various sources and queries, which after deduplication was reduced to 834 unique DOIs. Figure 10 below shows the number of datasets found per institution (stacked bars) relative to the amount currently registered in FRIS for the same period (black line).

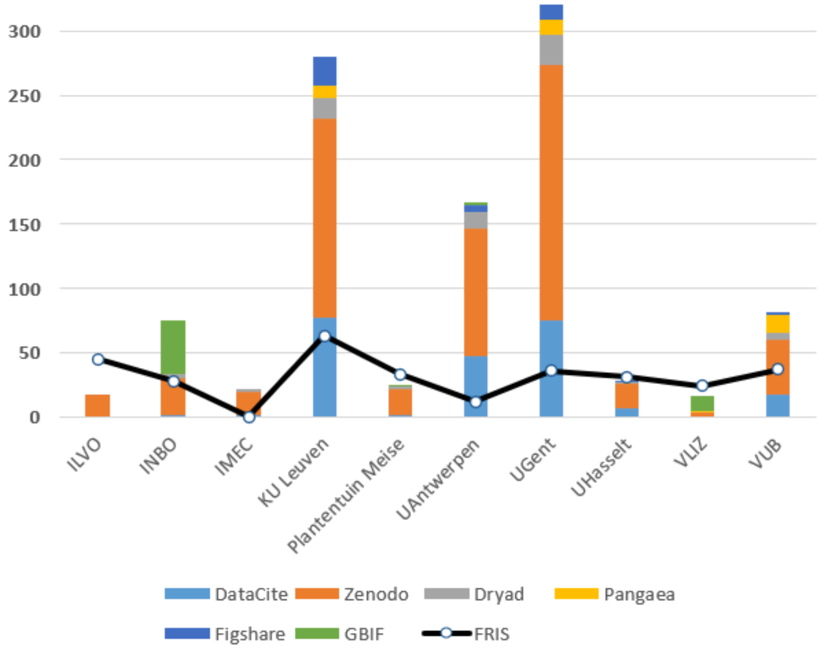


Figure 10. Datasets from Flemish RPOs (2022) found through repository APIs (stacked bars) and registered in FRIS (black line).

As is evident from Figure 10, for the majority of institutions there were significantly more datasets found through searching the various sources than were found in FRIS (and therefore can be assumed to be present in the institutions' respective CRISs). Even accounting for duplicates, there were over 800 more datasets found through this approach than are presently registered in FRIS. Thus, this modular approach can quickly produce significant improvements to the different RPOs' oversight of their respective research data outputs, especially when used in a retroactive sense to find already published datasets (which the authors might be less likely to register with their institutions retroactively than they are of new datasets published now).

It is worthwhile to address the areas of non-overlap between what exists in FRIS and what was found via the modular approach and the datasets registered in FRIS. Why did the modular method not find all of the datasets already registered in FRIS? To explain, we can divide these missed datasets into two categories: datasets missed from repositories that the modular approach did *not* search and datasets missed from repositories that the modular approach *did* search. The former is simple to account for, as the modular approach obviously cannot find records that are not in the repositories in which it searches. We initiated this analysis with API integrations for the repositories that we considered the most popular amongst Flemish researchers, but the intention of the modular approach is that it is easily extendable, such that new integrations can be easily created in "plugged in" to accommodate searches for different repositories. Therefore, as the modular approach expands to incorporate integrations with the APIs of more repositories, we expect this number of missed records to decrease.

In contrast, the datasets that are missed within repositories that were actually searched could have been missed for several reasons. This could either be due to gaps in the queries used - which could potentially be refined over time - or due to limitations in the metadata provided in the record on the repository page. For example, if a researcher creates a dataset record in a particular repository but does not provide any affiliation in the metadata, then querying the repository with the name or ROR of their affiliation will not find that record. One method of counteracting this would be to strengthen efforts in educating researchers on the benefits of thorough metadata documentation and encouraging them to include affiliation information when publishing dataset records online.

The Registration Approach

As seen from the above description, the modular approach finds a significantly greater number of datasets than currently sit in FRIS, but it of course does not find *all* possible datasets. Therefore, there remains a need for researchers to identify specific datasets and notify their institutions that they have published these datasets. For this, we envision a service that is as simple for the researcher as possible. Ideally, the researcher need only supply the PID for their published dataset and the system automatically harvests all of the relevant metadata from the repository page (preferred) or, secondarily, through an aggregation service such as DataCite or OpenAire.
The researcher can review the harvested metadata and make edits if necessary. This system could retain much of the same infrastructure as is necessary to implement the modular approach, instead querying the various sources for individual PIDs rather than more general searches.

Such systems already exist at various Flemish institutions as integrations into their respective CRISs. However, the implementation of this approach at a broader level would require more development than the strictly modular approach because there would also need to be an interface wherein researchers could enter their PID and review the metadata of their published datasets.

To minimize the development necessary to integrate such an interface into the CRISs of the various institutions (or to build a standalone application for institutions with no CRIS), this registration approach can be implemented at different levels.

In conclusion, this project group advises to use a combination of the above described modular approach and the registration approach. We further recommend establishing a formal group or community within the FRDN including research data staff (RDM support staff and technical staff) to develop and maintain a platform to exchange knowledge regarding metadata harvesting procedures (codes, queries, best practices etc.) based on the POC that resulted from this project group. This will benefit both dataset metadata registration and harvesting.

Conclusions & Future Directions

Based on the analysis conducted by this Project Group, we have delivered two Key Deliverables. The first (Key Deliverable 1), describes the current status of research metadata sharing and enhancement amongst Flemish RPOs and provides a list of candidate improvements, ranked according to ease-of-implementation and strength of impact. The second (Key Deliverable 2), outlines a proposed method of finding a greater number of datasets affiliated with Flemish RPOs by harnessing the APIs of various popular data repositories and dataset aggregator services. We provide a set of basis code that can be built upon to extend the reach and accuracy of the data harvesting software.

Through Key Deliverable 1, we have provided a list of improvements to the Flemish data and metadata ecosystem, however their actual implementation remains outside the scope of this project. Given that many of these recommendations (e.g., making FRIS business rules less restrictive, or using FRIS for harvesting) involve multiple stakeholders, careful consideration is necessary to determine how best to implement them. What we have presented here is a survey collecting the points of view of the participating stakeholders on their current and on their ideal technical metadata flow. In brief, we have presented *what* the identified most important steps are, but not "*how*" they can be practically instantiated.

Similarly, we provide the basis for harvesting dataset metadata in Key Deliverable 2, there is still more to be done regarding its implementation that was outside the scope of this project group. Ideally, the code should be hosted in a publicly accessible location (e.g., GitHub, Zenodo) where it can be accessed by the various RPOs. Furthermore, precise documentation is necessary to explain the download, set-up, and use of the software, along with documentation explaining - in detail - strategies for generating accurate search queries from each repository in question.
Furthermore, efforts should be taken to build a community around this harvesting software, such that as staff from different RPOs identify new repositories not serviced by any of the available modules there exists a community wherein they can find tools and assistance for creating a module for the new repository in question, and then subsequently a place for them to share that new module once created. We note that GitHub seems suited to this task, but leave it for future groups to decide.
This project group advises to establish a formal group or community within the FRDN existing of research data staff (RDM support staff and technical staff) to develop and maintain a platform to exchange knowledge regarding metadata harvesting procedures (codes, queries, best practices etc.) based on the POC that resulted from this project group. This will benefit both dataset metadata registration and harvesting.

Appendix

Chart A1: Technologies/sources explored in 3.a for Key Deliverable 2

OpenAire Explore	https://explore.openaire.eu/
ORCID	https://orcid.org/
DataCite Commons	https://commons.datacite.org/
Data Citation index	https://clarivate.com/products/scientific-and-academic-research/research-discovery-and-workflow-solutions/webofscience-platform/data-citation-index/
Google Dataset Search	https://datasetsearch.research.google.com/
Crossref API	https://www.crossref.org/
BASE	https://www.base-search.net/
ARK	https://arks.org/
ARXIV	https://arxiv.org/
PURL	
URN	
URL	
Handle	https://www.handle.net/
EOSC Portal	https://eosc-portal.eu/
CORDIS	https://cordis.europa.eu/
e-CORDA	
Open Research Europe (ORE)	https://open-research-europe.ec.europa.eu/
CoARA	https://coara.eu/
euroCris	https://eurocris.org/
Zenodo	https://zenodo.org/
Dryad	https://datadryad.org/stash
Pangaea	https://www.pangaea.de/
OSF	https://osf.io/
Figshare	https://figshare.com/

EGA	https://ega-archive.org/

Chart A2: Frequency table of number of results per query type and source combination

	DataCite	OpenAire	ORCID	Repositories (aggregate)
Institution names	1231	1184	0	2480
Institution RORs	917	0	0	509
ORCIDs	146	131	45	83
Author names	225	230	0	78
Author names + affiliation	77	80	0	7