Association of a federal data repository and the I14Y interoperability platform – Proof of concept

on behalf of the Federal Office of Public Health (FOPH)

contract ID:142006407

Authors

Racine Céline, *Unisanté, University Center for Primary Care and Public Health, Lausanne, Switzerland*

Pochon-Levit Floriane, *Federal Statistical Office, Switzerland*

Kühn Thorsten, *Federal Office of Public Health, Switzerland*

Reviewers

Bochud Murielle, *Unisanté, University Center for Primary Care and Public Health, Lausanne, Switzerland*

Warpelin-Decrausaz Loane, *Unisanté, University Center for Primary Care and Public Health, Lausanne, Switzerland*

Gouzi Fabrice, *Federal Statistical Office, Switzerland*

Beer-Borst Sigrid, *Federal Office of Public Health, Switzerland*

DOI : https://doi.org/10.16908/pub.2025.003

Date of report: 20.05.2025

1. Summary

The dissemination of data is only possible if the party who wishes to (re-)use data is aware of its existence. Consequently, the findability of data sets is a prerequisite for data reuse. One of the methods employed to ensure the findability of data sets is to register them with a data catalogue. The proof-of-concept (POC) documented in this report was initiated to provide and test an automated solution (a so-called harvester service) to register and update metadata for a data set in the I14Y data catalogue.

This POC was produced by the Federal Statistical Office (FSO) and Unisanté in the context of a mandate (Proof-of-concept-DigiSanté) including 3 case studies as part of the DigiSanté program. This specific case study aims at designing a practice-oriented solution to improve the visibility of Swiss health-related research projects through the I14Y Switzerland's national data catalogue.

This POC allowed the development of a harvesting module on I14Y to retrieve public metadata from data repositories. This report details the technical development of the POC, including testing steps, feedbacks and improvements. The section "procedure for future association" is of particular interest for agencies willing to enhance findability of their metadata by means of the national data catalogue.

2. Context

2.1. Program DigiSanté and the project Data space for health-related research

DigiSanté, a program of the Federal Department of Home Affairs, is carried out jointly by the Federal Office of Public Health (FOPH) and the Federal Statistical Office (FSO). It aims at promoting the digital transformation in the healthcare system and the implementation of the Swiss Health Data Space (SHDS) in cooperation with stakeholders in the healthcare systems. DigiSanté establishes digital and standardised health-related services of the federal government and in doing so facilitates efficient day-to-day work. The broad use of a secure health data space by all stakeholders lends support to the high quality of treatment provided within our modern healthcare system and improves both public services and research conducted in the interests of all people.

The FOPH-project "Data space for health-related research" (DSHR) is implemented as part of the DigiSanté program package 4 on secondary use of data for planning, management and research. The project is based on the Federal Council's report of May 4, 2022, in fulfillment of Po 15.4225 Humbel "Better use of health data for high-quality and efficient healthcare". The Federal Council decided to further refine the suggested "national system for the reuse and linking of health data for research purposes" for implementation in the context of the SHDS.

Therefore, the project DSHR aims, among others, to support establishment of conditions for enabling health data controllers in the public and private sectors to manage their data in accordance with the FAIR principles. Focusing on the principle of data findability, this case study should illustrate how data controllers can easily make their data collections visible in the national data catalog I14Y, an interoperability platform. That is, under responsibility of FSO, the POC served to clarify how I14Y can interact with local data repositories (see chap 5.3) using a harvester service to support data presentation on I14Y.

2.2. I14Y and its role in the Swiss administration, laws and obligations

The I14Y interoperability platform is Switzerland's national data catalogue. It helps to ensure an efficient exchange of data between authorities, companies and citizens. In the platform, an index of the data collections and interfaces of the Confederation, cantons and communes is continuously expanded and their metadata are made available centrally.

Public authorities, government-related and private companies can use the platform to create an inventory of their structured datasets. At the same time, I14Y serves to harmonize the data so that they can be used multiple times (once only principle).

The platform is developed and operated by the Interoperability Office (IOS) in the Interoperability and Registers Division (IOR) of FSO. The operation of the platform is regulated by law in the Federal Act on the Use of Electronic Means to Conduct the Tasks of the Authorities (EMBAG). The Ordinance on Digitalisation for the Confederation (DigiV) and the explanatory report provide further information. The platform forms the technical core of the national data management program (NaDB). The features of the platform are continuously expanded and documented in the I14Y public roadmap and the I14Y changelog.

2.3. Data repositories at Unisanté

2.3.1. Presentation of the data repositories

In the healthcare sector, research data are shared under the banner of "as open as possible, as closed as necessary", following the Open Research Data (ORD) principles. Indeed, health-related data might be subject to the Human Research Act (HRA) or the Data Protection Act (FADP) as long as it is not anonymized. Anonymization of health data is extremely difficult to achieve. Its application considerably reduces the usefulness of a data set. It is therefore the approach of relative anonymization, or de-identification, that is generally used in the medical field, when possible. De-identification implies that data sharing is carried out on a restricted-access basis, i.e. that data is accessible free of charge, but only on request, to anyone wishing to carry out a research project. This type of sharing is in line with the ORD and FAIR principles and guarantees the protection of participants. To valorize[1] and share dataset on a restricted-access basis, the research teams use institutional data repositories.

Unisanté is a university center for general medicine and public health based in Lausanne, Switzerland. The institution covers the entire healthcare chain, from primary care, care for vulnerable populations and occupational medicine, to health promotion and prevention, organization of the healthcare system, research and university teaching. Its aim is to maintain and improve the health of the regional population. As an academic institution, Unisanté leads research projects on all its areas of expertise. Moreover, Unisanté conducts health-related studies mandated by the federal government, in particular with the Federal Food Safety and Veterinary Office (FSVO) for nutrition-related studies. As a partner, Unisanté assumes the role of data steward, providing an infrastructure and service for data sharing from commissioned studies.

Unisanté, through its IT and Documentation and Data Unit (UDD) teams manages two institutional data repositories : Unisanté data repository and FSVO data repository.

The role of the first one is to share research datasets produced by Unisanté and to offer a secured alternative for sharing coded and deidentified health related datasets. Due to its expertise in data repositories management, Unisanté provides a data repository to the Federal Food Safety and Veterinary Office (FSVO). Its goal is to share the research datasets (raw data) of two studies conducted by Unisanté on behalf of the FSVO. In the future, it will be used by the FSVO to share all their research datasets. For aggregated or anonymous data, the FSVO uses the opendata.swiss platform.

The data repositories have been respectively created in 2014 and 2016, to answer the need for restricted data sharing, as no other data repository accepted coded health related datasets or provide a secured way to share restricted datasets.

Both systems are based on NADA, an open-source software for data repositories. The system is based on XML metadata in the Data Documentation Initiative (DDI) schema. This metadata schema allows a deep description of the research project, the methodology used, the datasets and their variables.

[1] Data valorization aims to enable the FAIR principles by ensuring the good quality, interoperability and reusability of metadata and datasets. This term is also used when a dataset can't be shared, but its metadata and associated documentation are shared through a data repository.

Moreover, this software provides a secured way to share restricted data, with a traceability on data requests and downloading.

2.3.2. Structure of the research project on the repository

Each research project is presented on a single web page. In the back-end, each project has a secure storage space for data files to be shared. Each web page is composed as follows:

- Study description (DDI-XML metadata)
- Datasets (files can be made available in Open Access or on request, depending on the sensitivity of the data)
- Data dictionary (DDI-XML metadata describing each variable from each dataset)
- Open Access documentation (i.e.: codebook, interview grid, script, surveys…)
- Related publications (list of publications using the dataset)

Metadata is created in XML according to the DDI schema and is made available in Open Access on the repository. They can also be exported in JSON (automatic conversion from the repository).

2.3.3. Findability and access of a dataset

To find a dataset on the Unisanté or FSVO data repositories, a user can either access directly to the data repository or use the data browser Google Dataset Search, as the XML metadata are parsed by this system. The Digital Object Identifier (DOI) pointing on a dataset must also be cited in all scientific articles that utilize the data, to facilitate data discovery.

To get access to a dataset, the user fills in a form through a specific functionality on the data repository. Once all given information is verified (affiliation, ethics approval, agreement sign off…), the access is given through the data repository.

2.4. Opportunities / interest of integration in I14Y

2.4.1. Role of the I14Y platform in the valorization of healthcare data

I14Y is positioning itself as a meta-catalog for data produced in Switzerland by the federal offices and research institutions. This platform only valorizes metadata, as hosting datasets would raise technical and legal issues. To fulfill this role, the platform must be able to automatically retrieve metadata published on data repositories. A metadata-based meta-catalogue would make it possible to localize all types of data, whether from hospitals, research, administrative bodies or the private sector.

Bringing all this metadata together on a single platform is also an opportunity for research teams. As many research teams use statistical data from federal offices, being able to search datasets from other sources on the same system is a way to facilitate data reuse.

Metadata can be used to describe both a research project and the variables making up a dataset. I14Y could therefore help improve dataset interoperability by increasing the utility of metadata at the variable level.

2.4.2. Opportunities for data repositories regarding the FAIR principles

FAIR principles are used to assess the reusability of datasets, but also the quality of data repositories. An evaluation of the levels of FAIR principles was performed in 2024 for the data repositories

managed by Unisanté (Figure 1). It demonstrates that all the aspects of FAIR are addressed but that some are stronger than the others.

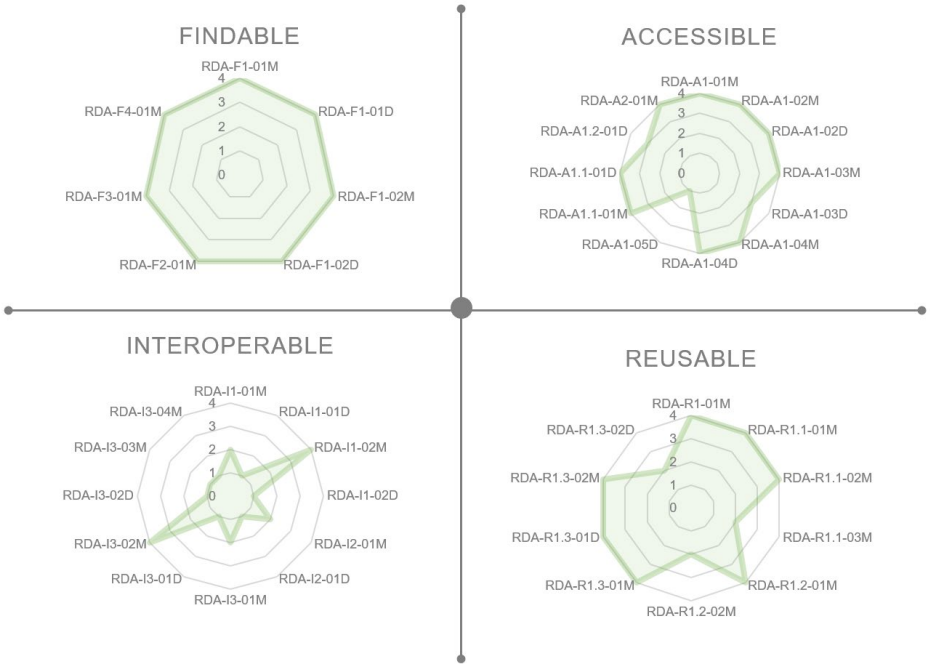

Figure 1 : FAIR maturity model for the FSVO data repository, March 2024

However, according to the statistics of the two data repositories, the number of data requests are lower than expected, even if the indicators are at their maximum for the Findability criteria.

Therefore, if a data repository meets all the FAIR criteria, this does not guarantee the data reuse. Valorizing datasets on the I14Y platform would enable data repositories to gain in visibility, as it offers users a new access point for their search for secondary data.

3. Scope and objectives

3.1.1. Scope

The context, presented in the previous section, shows an interest to valorize datasets produced in Switzerland on I14Y in an automated way. It represents a gain in visibility for data repositories and reenforces the **F**indability principle of the FAIR principles. Thus, Unisanté and FSO associated in the context of the Digisanté initiative to develop a proof of concept (POC) under the lead of FOPH as part of a mandate (add link) including 2 other case studies covering complementary aspects.

As Unisanté manages two data repositories, the scope of this proof-of-concept (POC) is limited to only one repository. The FSVO data repository is selected for this POC since valorization of its datasets will improve its visibility and support FSVO to be compliant with the EMBAG and DigiV. This POC is performed with the agreement of FSVO.

As most of the data repositories do not include variable-level metadata in their systems, this POC is limited to project-level metadata. If variable-level metadata were to be included in the POC, the mapping between the two metadata schemas would need to be extended. As a first step, the variable-level metadata could be imported and displayed on I14Y without being associated to underlying variables (so-called I14Y concepts). For harmonization purposes, the variables could be then linked to I14Y concepts (for example, variables coming from FSO). In this case, the original metadata on the data repository should be modified by the data manager and the harvesting completed with this information.

3.1.2. Objectives

The goal of this POC is to design a practice-oriented solution to improve the visibility of Swiss health-related research projects through the meta-catalog I14Y. To realize this goal, the developed solution should be automated. More precisely, the POC contains:

1. Building a harvester for I14Y to retrieve public metadata from data repositories
2. Testing the findability of metadata of FSVO data repository through I14Y
3. Providing a report including a procedure for future implementation for other data repositories

This automated solution not only meets the visibility needs of research projects, but also the need for a lightweight, easy-to-implement solution for data repository managers.

4. Procedure

4.1. Prerequisites

4.1.1. Prerequisites on I14Y

To set up the harvesting of a metadata catalogue on I14Y, an organization must contact I14Y to create the organization's "Agency" (according to DCAT standard) on the system, which will appear as the publishing entity of the organization's metadata. The organization will also be added as an eIAM (the central login infrastructure of the Confederation) unit to allow access to the I14Y application.

The organization provides the contact of at least one person that will act as its contact person (or local data steward according to NaDB roles). This person receives highest permission for managing the organization's metadata and can further delegate roles with lesser permission within the organization. The Contact person can log in on the platform with their existing FedLogin, CH/AGOV-Login or Switch edu-ID login.

To fully automize management processes such as a harvester, a technical user must be created by the I14Y team on the FSO user management tool, Keycloak, to generate access tokens that can be used to manage metadata through the I14Y Partner API. A technical user specifically dedicated to the organization's harvester is created upon request.

4.1.2. Standard metadata and API

To allow I14Y to harvest metadata, it is important to ensure that the two systems can communicate. The prerequisites to enforce are 1) the availability of an API, 2) JSON exports and 3) standard metadata. As the FSVO data repository is created on the NADA software, metadata exports and API are existing functionalities. The Unisanté IT team and the data repository manager at UDD collaborated to check and secure these two functions.

Regarding the metadata, the data repository manager checked manually if all XML metadata were correctly translated to JSON during the export. All mistranslation between the two formats were documented to be handled in the next step of the project.

The IT team was responsible for activating the API and carrying out a security assessment using the OWASP WSTG methodology, to identify any weaknesses, risks and potential vulnerabilities. Following this assessment, the necessary corrections were made to secure the transfer of information via the API and to guarantee protection against intrusions. The activation has been limited to the API roads allowing the retrieval of public metadata (GET). After the activation, a security control was performed by an Unisanté cybersecurity expert.

4.2. Proof-of-concept development

4.2.1. Mapping

To enable a harvester to retrieve metadata from one system and publish the information in another one, a mapping between the two metadata schemas was necessary. Not all repositories use the same metadata schemas. Dublin Core is widespread, but its tags and level of description are very restricted. The FSVO data repository uses XML-DDI metadata, as it is more detailed.

The data repository manager and the I14Y manager collaborated to create a mapping between the DDI schema and DCAT-CH, used on the I14Y platform.

The first step was performed by the data repository manager. First, a list of the minimum elements needed to describe the data was drawn up, since the level of description varied between datasets. Then, the metadata were exported to check if the minimal information were correctly represented, first in the XML export, second in the JSON export generated by API.

Both sets of information (minimum metadata and JSON export) were sent to the I14Y manager for comparison with the DCAT. Moreover, some mapping between DDI and DCAT already existed and were sent to the I14Y manager as supplementary information.

Based on this documentation, the I14Y manager proposed a first version of the mapping. The I14Y manager highlighted the information missing from the DDI but required in the DCAT. Another highlighted problem was the field format: some of the I14Y fields were lists of choices, but the equivalent XML tag was open text. Therefore, the two teams met to find the appropriate XML tags or repository-specific fields that would be exported in XML.

Once the XML tags were defined, the data repository manager exported again the JSON and sent it to the I14Y manager for testing. This final testing concluded the mapping process.

4.2.2. Harvester

I14Y has been developing a harvesting module, first for the purpose of this POC, then generalized to all I14Y agencies wishing to implement an automatized import and update of their metadata catalogues on the platform.

The harvester uses the I14Y Partner API endpoints to create and edit dataset metadata on I14Y. The standard used to describe dataset metadata on I14Y is the Swiss application profile of DCAT. A DCAT dataset corresponds to one entry in the source data repository and describes one research dataset. The information on data access is managed together with the DCAT dataset itself and displayed on I14Y as a DCAT distribution. The DCAT datasets are created in the organization's private catalogue and can also be directly published to the public I14Y catalogue.

The harvesting module was developed in Python on GitHub with Github Actions to run the harvesting job at regular intervals, or to trigger it manually. A private repository serving as template for all harvesting projects and ready to be customized for each client was later added to the I14Y GitHub organization page.

The template repository contains:

- The client credentials saved as repository secrets
- The workflow that contains the configurations for the frequency of runs and the functions to retrieve the access token and save logs
- A harvesting script with functions to create new datasets from a file or from an API
- A mapping script between the two catalogues

The runs are accessible together with a log file containing the change made since the last run.

4.2.3. Go live and settings

The FSVO data repository harvester was configured on the I14Y GitHub organization to publish to the test environment (ABN) of I14Y.

For the FSVO catalogue harvester, additional functions were added to the base harvester to:

- Update existing datasets based on the timestamp of the last changes made to the metadata in the DDI API response
- Publish the datasets on the public I14Y catalogue (instead of keeping them in the private organization catalogue)

A technical user for the FSVO Agency was also configured on Keycloak.

The harvesting was activated in February 2025 and runs since daily at night. From the I14Y point of view, the results are good and meet the quality standard of the catalogue. In agreement with all parties, the I14Y manager will copy the harvesting to the production environment in July 2025.

4.2.4. Management/administration of the interface

Once the harvester is set up and activated, no other actions are needed from I14Y. The Local Data Steward can check the rendering of the harvested data on the public interface or the private I14Y catalogue. He can also modify the harvested data if needed. Further, they can also be authorized on the Github harvesting repository should they want to make changes to their configuration.

Should errors happen in the harvesting pipeline, a notification is sent to the I14Y mailbox, and the fix is undertaken in the I14Y team or forwarded to the organization if the error happens due to some API misconfiguration.

5. Results

This POC had the first objective to produce an automated solution for data valorization on I14Y. It resulted in two deliverables: a mapping between XML-DDI and DCAT-CH and a harvester for I14Y. The second objective was to test the findability of the datasets on I14Y. Then, the third objective was to produce a procedure for future implementation of the developed solution. A pros and cons analysis based on feedback and a checklist for implementation are available at the end of this chapter.

5.1. Feedback

5.1.1. Unisanté

5.1.1.1 Challenges

This POC revealed few difficulties for the agency / the data repository manager.

The challenges encountered were technical. First, to allow the harvesting, the data repository must have an API. Fortunately, the FSVO data repository already had this functionality, which needed to be activated. Second, the data repository must have a standard metadata schema. The one used on the FSVO data repository is the DDI. However, no existing mapping between DDI and DCAT-CH was available. The mapping was therefore an important step for this POC. The difficulty here was the granularity of the metadata. As the two schemas have different levels of description, some choices had to be made during the mapping.

5.1.1.2 Lessons learned

The project highlighted the need of having a standard data repository, with exports functionalities and API. Most data repository softwares (NADA, CKAN, DSpace, etc.) have these features. Only platforms developed outside these software packages will need to commit development time to ensure system compatibility.

Moreover, the challenges encountered revealed the importance of metadata mapping. The difference of granularity between the XML schemas must be taken into account to allow a clear representation of a research project. It is important to rely on existing mappings, if available, and extend them to get the most complete description possible. This extension should be done as a collaboration between the data repository owner and the I14Y team. With a more precise mapping, the integration of other data repositories to I14Y would be facilitated.

5.1.1.3 What know-how must be available?

To realize the development, the technical competencies needed are the same as the one needed for the management and maintenance of any data repository.

One must know XML and JSON metadata used in the data repository and understand them. It is important to help the I14Y team for the mapping between the metadata standard and the DCAT standard.

Skills in software management are also needed to activate the API. To secure it, competencies in cybersecurity are required.

5.1.2. FSO

5.1.2.1 Challenges

The FSVO data repository harvester was customized in relation to the response of the Unisanté NADA/DDI JSON API, specifically the timestamp for tracking changes. In future, I14Y wants to provide several standard templates for the main harvesting protocols (eg OAI-PMH) so that more regular solutions can be offered.

Also, a user interface allowing the partner organization to define the mapped fields would be more engaging than the customization of a Python script and should be developed.

Several improvements are still required for the user interface to better display the data on the platform, for example in the description field to allow HTML tags in the source data. This is a work in progress.

5.1.2.2 Lessons learned

This POC has allowed us to develop a solution that can be proposed across all agencies using I14Y and develop further scripts for agencies to manage their metadata. This is a very valuable use case for increasing the visibility and usage of the platform across organizations managing data catalogues.

5.1.2.3 What know-how must be available?

The two options for an organization to have either:

- The harvester repository managed by I14Y or
- To fully manage their harvester on their own

allows for more flexibility depending on the know-how available at the organization.

The only requirement for an organization would be to provide an endpoint where the data can be retrieved and let the I14Y team configure the module.

5.2. Pros and Cons

Associating a data repository with I14Y has both advantages and disadvantages.

The first advantage is the increased visibility of research projects, by facilitating access to the data. Firstly, if a dataset is referenced on several repositories, it will be easier to find via a web search. Secondly, I14Y has gained recognition over the course of the project, particularly among research teams reusing data. Its popularity is an asset for datasets valorization. The third advantage is the ease of implementation for data repositories. As the harvester has been developed, only the mapping needs to be done if it doesn't already exist.

The negative points to be raised are not obstacles to platform association, but rather potential improvements.

The first is the time required for metadata mapping, if this does not already exist. However, as many repositories share the same standards, only a small amount of mapping will be required to suit most users. The quality of the mapping could also be improved by extending the descriptions. Currently, only minimal metadata is described, but this POC has highlighted the possibility of extending this mapping.

Another aspect to improve is the description of the research project on the platform. As an example, a research project can have co-primary investigators, which means multiples people responsible for the research. The I14Y metadata cannot represent this information as it currently stands, as the dedicated field is non-repeatable. At present, only a selection of metadata is used for the mapping between DDI and DCAT-CH, and these do not necessarily correspond to the full description of a scientific project. This might represent a challenge for institutions wishing to join I14Y.

5.3. Procedure for future association of external data repositories to I14Y

This POC provided an automated way to connect an external data repository to I14Y. As a summary of this report, here are the steps to follow to associate a data repository to I14Y.

1. The institution contacts I14Y at i14y@bfs.admin.ch to check if the requirements under 4.1.2 Standard metadata and API are met, in particular the existence of an API to retrieve the structured data.
2. If no API exists, an excel template can be provided for a single import of the data. The institution will be responsible to maintain the data up to date.
3. I14Y creates the institution's "Agency" (according to DCAT standard), an eIAM unit (for access management on the Confederation's system) and a technical client (for access token retrieval).
4. To create a user, a contact person must be defined.
5. The organization and I14Y must communicate to define if a mapping between the metadata of the data repository and DCAT is already available or needs to be created.
6. The organization either:
    a. Forks/clones the template GitHub harvester repository to configure it on their own, or
    b. Delegates the management of their harvester to the I14Y team, which will create a specific agency repository under the I14Y organization and set up the necessary configurations.
7. Finally, once the metadata are harvested, the data repository manager must log in I14Y and check if all information are correctly shown. If necessary, manual modifications can be done

This process is currently being documented and will shortly be available on the I14Y handbook

6. Conclusion

This POC reached its goal by providing an automated tool to improve the accessibility of data from, and valorize, research projects in the I14Y meta-catalog. The FSVO data repository has been successfully associated with I14Y.

The deliverables are a harvester for I14Y, a metadata mapping between DDI and DCAT-CH and a procedure on how to associate a data repository to I14Y.

This POC revealed the opportunities for Swiss data repositories to be represented in a meta-catalog.

This experience has also highlighted potential improvements to the I14Y platform. First, the mapping between the metadata schemas could be upgraded, to allow for a deeper description of research projects and their related datasets. Second, I14Y was developed according to the needs of federal offices. Therefore, improvement would be needed to better represent research projects and datasets, as their structure might be different. The association of I14Y and data repositories is the perfect opportunity to collaborate and improve this aspect of the platform.