IJDC | *Conference Paper*

Development of an Integrated Lifecycle of RDM Tools: Looking Back and Forward at KU Leuven

Dieuwertje Bloemen
KU Leuven

Ingrid Barcena Roig
KU Leuven

Abstract

After the creation of an institutional research data management (RDM) policy some years ago, KU Leuven decided it was crucial to create an ecosystem of RDM tools for the entire RDM life cycle to enable KU Leuven researchers to manage their data well and in a FAIR way. In 2018, KU Leuven started on the road to an institutional data repository, which resulted in the launch of the Dataverse-based Research Data Repository (RDR) in January 2022. In line with this, there was a strong focus on tools to manage data in a structured and metadata-rich way during the research project, when data was still being worked on. Based on this, an Integrated Rule-Oriented Data System (iRODS) instance that would enable easy and integrated metadata management from the start of a research project via a self-developed portal was launched in March 2023 under the name "ManGO," which is a shortened version of MANagement van Gegevens voor Onderzoek. Because one of the four FAIR principles is interoperability, and to fully enable FAIR data management, it was important to facilitate this interoperability of data by ensuring that KU Leuven RDM tools are interoperable, with a first focus being the connection between ManGO and RDR to select and copy data ready for publication. As with ManGO, RDR, the integration work, and other RDM developments made or adapted at KU Leuven, all relevant work will continue to be shared in open source where possible to ensure that investments that the university makes in RDM tool development can be used by and are accessible to other institutions.

Submitted 10 February 2025 ~ *Accepted* 20 February 2025

Correspondence should be addressed to Dieuwertje Bloemen, Email: dieuwertje.bloemen@kuleuven.be

This paper was presented at the International Digital Curation Conference IDCC25, 17-19 February 2025

The *International Journal of Digital Curation* is an international journal committed to scholarly excellence and dedicated to the advancement of digital curation across a wide range of sectors. The IJDC is published by the University of Edinburgh on behalf of the Digital Curation Centre. ISSN: 1746-8256. URL: http://www.ijdc.net/

Introduction

In 2014, KU Leuven established its first set of research data management (RDM) policy guidelines and best practices, and an RDM working group to explore the RDM needs of the university's researchers. Four years later, a Research Infrastructure Self-Evaluation (RISE) analysis[1] was performed, which resulted in a new set-up of working groups at the university focusing on RDM policy, infrastructure, and advice and training, as well as the establishment of the university's RDM Competence Centre (RDM-CC). Alongside this, KU Leuven decided it was crucial to create an ecosystem of RDM tools for the RDM life cycle to enable KU Leuven researchers to manage their data well and in a Findable, Accessible, Interoperable, and Reusable (FAIR) way.

With KU Leuven being the largest university in the Low Countries and dedicated to education and research in nearly all fields, the challenge would be to create and provide infrastructure and tools that could be used by as many researchers as possible at the university. The university, for example, has a wide range of research domains across 15 faculties, which are clustered into three thematic groups: (1) Humanities and Social Sciences; (2) Science, Engineering and Technology; and (3) Biomedical Sciences. At the university, innovative research forms the basis of all academic programmes, with the university being the #1 grant recipient in the Horizon Europe framework programme (FP9).[2] Therefore, KU Leuven made it a priority to support open science by investing in the necessary innovative research infrastructure(s) to enable its researchers to practice research according to the open science principles using the FAIR-enabling institutional RDM tools.

Background

The development of RDM support services at research institutions is, in part, driven by the need for robust and high-quality research, as well as external factors. For KU Leuven, a research-intensive university in Flanders, Belgium, key external drivers were: (1) the adoption by Fonds Wetenschappelijk Onderzoek (FWO)—the main research funder in Flanders—of RDM requirements into their policy in 2018; (2) the European Commission mandating open science and FAIR data[3]; and (3) the Flemish government established the Flemish Open Science Board (FOSB)[4] in 2020 and provided funding to all universities to recruit data stewards and develop data infrastructure.

A first university-wide policy guideline on RDM was adopted at KU Leuven in 2014. In 2018, an RDM steering group was created with three working groups: one on policy, one on advice and training, and one on infrastructure. In 2019, a new RDM policy was adopted, focusing on high-quality research and scientific integrity. This policy was further expanded on in 2022, with guidelines on publishing and sharing research data.

The RDM-CC was set up at the university in 2020 with five RDM experts to coordinate the development and roll-out of RDM guidance, training, services, tools, and infrastructure to support researchers with improving their RDM skills, practices, and competencies. The

[1] RISE analysis: https://www.dcc.ac.uk/sites/default/files/documents/publications/UsingRISE_v1_1.pdf

[2] KU Leuven is the highest-ranked university in Horizon Europe: https://nieuws.kuleuven.be/en/content/2023/ku-leuven-is-highest-ranked-university-in-horizon-europe

[3] The EU's open science policy: https://research-and-innovation.ec.europa.eu/strategy/strategy-research-and-innovation/our-digital-future/open-science_en

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

university's RDM service capabilities and infrastructure were analysed, and further developments were planned using the RISE Framework in 2018 and 2022. This benchmarking tool is designed to facilitate RDM service planning and development at an institutional level. At KU Leuven, it was used to set priorities for future developments, which were then translated into annual RDM action plans.

In the 2018 RISE analysis, it was concluded that there was a need for the development of infrastructure to store and manage metadata of research datasets and to better store the data itself. This infrastructure aimed to provide an ecosystem of tools at KU Leuven that would help researchers manage their data better and make their data more FAIR. In practice, this meant exploring the option of establishing a research data repository and charting what data management tools were already available at the university to identify gaps in the assortment of tools. These initial analyses resulted in the start of two major infrastructure projects: (1) the creation of an institutional research data repository; and (2) the creation of a tool for managing data in a metadata-rich way during the active phase of research.

These two infrastructure projects were further supported and boosted by the creation of the FOSB in 2020, which enabled KU Leuven to invest even more in its RDM tools due to the FOSB providing co-financing to its member organisations to support open science and FAIR research data management and by the board facilitating the Flemish Research Data Network (FRDN) to bring together a community of Flemish data stewards to exchange knowledge and experiences. The exploration of an institutional data repository for the publication of research data resulted in the launch of the Dataverse-based KU Leuven Research Data Repository (RDR) in January 2022. The strong focus on tools to manage data in a structured and metadata-rich way during the active phase of research resulted in the expansion of the range of RDM tools offered at KU Leuven with an iRODS[5] instance that would enable easy and integrated metadata management from the start of a research project. This iRODS installation was launched in March 2023 under the name MANagement van Gegevens voor Onderzoek (ManGO) and was adapted to the needs of KU Leuven research by developing a full customised in-house iRODS portal.

As such, KU Leuven has made significant investments in RDM over the last five years, which has resulted in a growing set of services and infrastructure, such as the institutional RDR to facilitate data publishing, and the ManGO platform to manage active research data, with metadata, automated workflows, and collaboration functionalities. In 2025, KU Leuven supports a wide array of RDM tools, including OSF, GitLab, REDCap, SharePoint, DMP-online, and a DMP Monitoring Tool, in addition to the iRODS and Dataverse instances. This assortment of tools results in a varied landscape of RDM tools that enable and support KU Leuven researchers in good data management. Because one of the four FAIR principles is interoperability, and to fully enable FAIR data management, it was important to facilitate this interoperability of data by ensuring that tools are interoperable where possible. Continued efforts are being made to integrate the institutional data repository and the ManGO environment with each other and with other RDM tools as much as possible. This has resulted in a strong focus on using exchange standards for data and metadata in the development and implementation of tools, in an effort to help researchers use RDM tools to their full capability and to simplify data transfers from active data management tools to a data repository. The strongest connection is currently between ManGO and RDR, due to the extensive control and customizability that KU Leuven has in these tools, with dedicated developer teams supporting their further development and maintenance.

[5] iRODS – Open Source Data Management Software (GitHub): https://github.com/irods/irods

Exploration of Requirements and Solutions

Looking back, for both infrastructure projects, a first step was to gather the requirements of researchers and research support staff. Once the requirements were defined, the exploration phase started, where different software options and tools were tested against these requirements. This includes internal and external tools to ensure that no redundancy is created unnecessarily. For the institutional data repository, LIBIS (a division of KU Leuven Libraries) would take the lead, starting with a requirements analysis in 2018, and for the active data management infrastructure, KU Leuven ICTS would take the lead.

Institutional Data Repository or Registry

To establish an institutional data repository, requirements were based on the results of an RDM tour organised in 2016, where support staff went around different departments and surveyed researchers for their needs in data management. On top of that, the FAIR principles were the foundation for most of the requirements to ensure that any system implemented would adhere to the open science best practices around data sharing. This initial list of requirements was then discussed with different groups of stakeholders to ensure the prioritisation according to the MoSCoW system[6] was agreed on, with the result being a prioritised list of requirements for a post-publication infrastructure set-up at KU Leuven.

With this finalised list of requirements, it became clear that the original idea to expand the existing literature repository (Lirias) would not answer most needs on this list. Therefore, it was deemed necessary to explore all options available in-depth to ensure the best possible choice was made. A list of software options, solutions, and tools was drawn up, and for each option, each requirement was rated on its availability or viability. Some of the top contenders were Figshare, Esploro, Dataverse, using the existing DSpace underlying Lirias, or setting up a separate DSpace for datasets. For each top contender, a SWOT (Strengths, Weaknesses, Opportunities, Threats) analysis was carried out in addition to checking the requirements. Although initially expanding on Lirias' DSpace was expected to be a strong contender, this quickly turned out to cause too many issues due to the rigidity of the established infrastructure. For example, the interface researchers use for depositing institutional copies of publications is limited when it comes to file management and structuring, and because this interface is not open source, adapting it to the needs of data publication would be challenging, if not impossible, on a short-term basis.

After discussing the options and their analyses with the RDM infrastructure working group and with approval of the steering committee, Dataverse[7] was chosen to start the pilot phase with, due to the relatively short time necessary to get a Dataverse instance up and running, but primarily because of its open source nature being in line with the spirit of open science. With Dataverse being open source, it would allow for more integration options and send a good message with regard to the open science context for which the infrastructure is intended. A caveat for all non-Lirias options was that a Lirias connection would need to be established, because this integration was shown to be one of the top priorities for many researchers for ease of use. However, an API exporting metadata would be able to perform the most basic integration, with such an API call existing in all explored options. Generally, a challenge for a Dataverse implementation is the need for more technical profiles within the institution. However, at KU Leuven, the pre-existing

[6] MoSCoW Prioritisation: https://www.agilebusiness.org/dsdm-project-framework/moscow-prioririsation.html
[7] IQSS/Dataverse – Open source research data repository software (GitHub): https://github.com/IQSS/dataverse/

experience of LIBIS with systems such as Lirias and Teneo, the university's preservation system, would be more than sufficient to get an initial pilot up and running. Once the decision was made to continue with Dataverse in early 2020, the pilot phase was prepared and started.

Active Data Management System

User requirements for the active research data management infrastructure (ManGO) were gathered during the RDM tour organised in 2016. The selection of iRODS as the software base for the platform was influenced by synergies with another ongoing platform initiative within the Flemish Supercomputing Centre (VSC),[8] a collaboration between all Flemish universities and the FWO to provide computing and data services for Flemish researchers. In 2019, KU Leuven initiated a pilot project to evaluate iRODS as an active RDM platform for the VSC High Performance Computing (HPC) users. Given the experience that KU Leuven had already gained with this platform, it was decided to use the same technology for a KU Leuven-wide platform.

The main criteria for selecting iRODS as the software included its capacity to store large amounts of data, the ability to add metadata to describe the files and folders and make them findable, the flexibility to define custom data policies within the system, and the availability of a framework to automate data management tasks, reducing repetitive tasks and manual errors. The diversity of clients who could interact with the system (a web portal, command line interface and Python API) was also a crucial feature, allowing users to choose the client that best fits their needs and technical skills. The open source nature of iRODS and its large user community, organised around the iRODS consortium, were also important factors in the decision. The opportunity to learn from and be inspired by the experiences of other institutions using iRODS undoubtedly helped to implement ManGO over a shorter time frame.

The development of ManGO as a service then proceeded in phases. In the first phase, a proof of concept (PoC) for ManGO was set up so that the wide network of RDM support staff (data stewards and local IT staff) could become familiar with the ManGO platform and provide feedback on its design and functionalities.

Establishment of Pilot Instances

Once the exploration phases finished and all requirements were listed, pilot phases were started. In the pilot phase, actual system deployments were set up to test the chosen systems and software to explore their technical feasibility and user friendliness. The latter was achieved by gathering feedback from relevant parties in a recurrent manner, to improve the systems incrementally until they were ready for production.

Road to RDR

In 2020, the pilot phase for the institutional data repository with Dataverse was started. Initially, a basic Dataverse installation was set up locally, and in tandem, the metadata model was decided on. A working group within the FOSB on metadata and standardisation was working on a metadata model for dataset based on the DataCite metadata standard, which would allow the FOSB to calculate key performance indicators (KPIs) for open data publication. In line with these developments, the pilot's metadata model was set up with these adaptations to ensure that the KPI reporting would be facilitated as best as possible. Once the basic pilot set-up was up and running with the metadata adaptations and internal

[8] VSC: https://www.vscentrum.be/

initial testing was performed, documentation was drafted to start a broader pilot testing phase where researchers and research support staff had the opportunity to test the system and provide feedback based on their experiences. This broad testing phase started in June 2021. Alongside the pilot of the system, a poll was launched to decide on the name for the upcoming repository, with the name RDR, pronounced as "radar," being chosen as an acronym for "Research Data Repository," and in pronunciation indicating that it is a platform in which to look for and find KU Leuven data.

In the meantime, the DataCite membership was set up to allow for DOI generation for each published dataset. An organisational decision to be made was whether the Dataverse installation should use collections, and if so, how they should be structured. The options considered were to have one collection for each of the three groups, one for each faculty or even one for each department. However, after deliberation, it was decided to start with no collections whatsoever to ensure that reorganisation of the university could not impact these collections and to prevent complicated decisions for interdisciplinary research as to which collection to choose to publish their data in. Based on how the Dataverse collections work, this decision could always be reviewed at a later date. However, up to this point, it has proven to be the right decision for now.

Once the broad pilot testing phase was over, the feedback was analysed and, where necessary, adaptations were made to the Dataverse set-up and the documentation. At this point, the main components of the set-up were done. However, items such as terms of use, protection of sensitive data, and the integration with Lirias and the university's library search system were not yet in place. This was worked on during the following months. The connection with Lirias was set up using the systems' APIs, which allows for the metadata of each dataset published in RDR to be registered in Lirias. From here, it could be exported to the library's online catalogue. The work on the terms of use, but particularly on how to correctly protect all types of sensitive or protected data in policies and workflows, took a long time. Dataverse allows for data to be restricted; however, this does not cover all the support that is necessary to help researchers when publishing their data as openly as possible, but specifically to do so as closed off as necessary. Once these crucial last items were in place, the repository was ready to go live.

Path to ManGO

After the initial PoC in August 2021, a pilot phase began in collaboration with various research groups from different domains. These groups worked with a Minimal Viable Product (MVP) and provided feedback and input for new features. This pilot phase was crucial when further developing the MVP and ensuring a well-tested service that met the research community's needs, which was ready for the production launch in 2023.

During the pilot phase, the name ManGO was chosen, and the system was promoted using various events and actions to increase the number of testing groups. Interested research groups were invited to intake interviews to confirm ManGO's suitability for their uses, gather needs and use cases for new functionalities, and assist them when getting started with ManGO. From August 2021 to March 2023, 26 research groups participated in the pilot phase: 12 from Science and Technology, 11 from Biomedical Sciences, and 3 from Humanities. Their needs included centralised and scalable data storage, automatic data capture from instruments, metadata management, workflow automation, sharing data with collaborators inside and outside KU Leuven, and long-term data archiving. The ICTS RDM support team assisted these groups, documenting unmet needs for future development. Follow-up meetings were held in May 2022 to check progress and address issues. The pilot phase concluded in March 2023, transitioning ManGO to production, with 80% of the pilot groups (21 out of 26) continuing to use it. In addition to the piloting by researchers, feedback was gathered from RDM support staff, similar to RDR, from departments actively engaged in RDM support, including the Research Coordination Office, KU Leuven Libraries, and local ICT staff through information and feedback sessions.

In addition, LIBIS, the team in charge of RDR, had a pilot project to integrate ManGO with RDR.

In addition to the initial needs expressed during the intake meetings, researcher feedback from the daily use of the platform was particularly valuable and led to significant improvements. Early in the pilot phase, it became clear that a graphical web interface was essential for accessibility, because many users were unfamiliar with command-line interfaces or Python. A user-friendly and portable web portal would facilitate the adoption of the new system by as many types of users as possible. Two community-developed web interfaces for iRODS, Metalnx and Yoda, were tested at the start of the pilot phase; however, both had drawbacks and could not fully meet the needs of KU Leuven researchers. Therefore, a new portal was developed based on the Python-iRODS client to provide full flexibility. The new web portal includes a feature for applying rich and standardised metadata through schemas, which was not possible in the flat metadata system in iRODS, which uses triplets of Attribute-Value-Unit. These metadata schemas, forms with required attributes and conditions, allowed users to add metadata to data objects and collections with immediate validation, ensuring accuracy and reducing errors. In addition, a Python library was developed to automate metadata management. The metadata schema module has since become the primary method for many research groups to work with metadata for their research data.

Moving to Production

Once the pilot phase was completed and all feedback had been processed, the project was ready to go live. For RDR, this happened in 2022; for ManGO, the launch into production happened in 2023. However, it should be noted that ManGO was already actively being used for research during its pilot phase by a selection of research groups. With a production instance going live, it was necessary to build awareness among researchers who could use the systems. Therefore, for ManGO and RDR, outreach was organised and in-person support was provided for all adopters.

RDR – An Institutional Research Data Repository

On 20 January 2022, KU Leuven RDR went live. This also meant the start of a campaign to promote its use to researchers. The promotion campaign consisted of posts in different KU Leuven newsletters, organising training sessions, drop-in sessions, and going to different department meetings to give short introductions on what a data repository is and what RDR has to offer. The biggest challenge was not that most researchers did not know that RDR existed, but rather that many did not even know what a data repository was for at that point. So, the campaign was two-fold, raising general awareness about data publication alongside promoting the new RDM infrastructure established for this purpose. This campaign was spread out over almost a year and a half due to the large number of departments at the university and because each update to the software was an opportunity to put the spotlight on RDR once more.

For example, one of the first updates to RDR was implemented just 3 months after launch, including a new feature that allowed users to easily enter author and related output information in the dataset's metadata using look-up functions developed by the RDR team. The related output look-up specifically was a further integration with Lirias, one of the key requirements early on for an institutional data repository. In addition, the RDR team has always strived to keep the software as up-to-date as possible in line with the most recent Dataverse software releases. These releases sometimes included additions to the Dataverse functionality made by the RDR team, where all KU Leuven-made developments flow back to the Dataverse community for reuse as much as possible. An

example of these kinds of KU Leuven-made features that are now part of the Dataverse software is the functionality added in Dataverse 5.11, where a user can configure their email and notification settings. A second example is the API functionality that allows draft dataset creation without supplying all mandatory metadata elements upon creation. This functionality was created to allow for easier integration with other RDM tools that don't contain all necessary metadata to create a draft dataset in Dataverse. The API call was implemented in the integration dashboard,[9] which was developed by the RDR team from scratch. This dashboard connects to a series of RDM tools to easily pull data into a Dataverse dataset, as will be explained later. Another dashboard developed in-house and made available in open source is the review dashboard.[10] The aim of the review dashboard is to easily keep track of what reviewer reviews what dataset, and to streamline feedback generation by providing template feedback points that are given if certain items are indicated as not being in order by the reviewer. Currently, work is underway to further support the RDR reviewers by automating some of the review points for them. However, it should be noted that the intention is never to fully replace the human touch necessary for a good, in-depth dataset review.

In addition to the technical improvements after going live with the first version, the necessary work was carried out to have the metadata appear in external search platforms (e.g., OpenAire, FRIS, and Google Dataset Search). In parallel, work was started on a CoreTrustSeal application. CoreTrustSeal certification is desirable to prove the trustworthiness of the repository and is a valuable exercise to ensure all necessary elements of a trustworthy repository are available and transparent to the researchers who wish to use the platform.

ManGO – An Active Data Management Solution

In May 2023, ManGO was launched in production. Similar to RDR, a communication campaign was initiated, which included posts in various KU Leuven newsletters, presentations at departmental events, a short introduction video, and the organisation of training sessions. Currently, ManGO hosts about 1 PB of research data from 150 active projects across different scientific domains.

The pilot phase underscored the importance of a user-centric approach for the success of a data management platform like ManGO. The system's design should be based on user needs to ensure alignment with their data management practices and workflows. In addition, the system must continuously evolve to keep pace with changes in research practices. To achieve this goal, the ManGO platform follows a well-established feature request procedure. This involves capturing new requests, analysing their viability, impact, and development effort, planning and implementation, and communicating new developments to users. Several different instruments are used to gather user needs. Similar to the pilot phase, each new group requesting access to the system is invited to a 2-hour discussion of their use case, focusing on their data workflow and data types. This is followed by an introduction to ManGO and a discussion where they receive advice on how to start using the platform, and where any new needs for their particular use case are captured and registered.

The support service is also a crucial source of information for understanding user needs. Through the support desk, users could ask for help using the system, report bugs, and request specialised consulting to implement their data workflows on the platform. Consulting could range from technical advice on simple tasks to implementing fully

[9] Integration dashboard: https://www.kuleuven.be/rdm/en/rdr/integration-dashboard
RDM-Integration – Integration tool for importing data to Dataverse (GitHub): https://github.com/libis/rdm-integration
[10] RDM-Review-Dashboard – A Dataverse review dashboard (GitHub): https://github.com/libis/rdm-review-dashboard

automated complex data workflows. In addition, after a year of ManGO going into production, a user survey was launched to gauge satisfaction with different aspects of the platform. The survey analysis identified several improvements, some of which have already been implemented (e.g., improved documentation) or are in development (e.g., an automation framework in the ManGO portal).

The close collaboration with users resulted in the continuous evolution of the ManGO platform. The iRODS software has been updated from version 4.2.9 to the latest available 4.3.3 version, following developments by the iRODS consortium. Eight new versions of the ManGO portal have been deployed, including new features, enhancements of current features, and bug fixes. Notable new features include an audit trail module for data objects that logs actions performed on a file from its creation to its deletion, including modifications, access, permission changes, or path movements. Significant improvements to the metadata schema module include the creation of composite fields, the ability to import and export schemas in JSON format, and the creation of a library of fields based on metadata standards (e.g., DataCite), which users could use as building blocks for their metadata schemas.

Training and documentation are also key aspects of supporting the ManGO platform. Regular introduction and advanced training sessions (e.g., on iCommands/Python) are organised. Users could also request tailored training that focuses on their specific needs or workshops about implementing metadata for their use case. A user-centric approach requires a highly skilled support team that works closely with researchers to understand their needs and translate these needs into specific new features for the platform and to develop them.

Establishing Interoperability

After the two new RDM tools, KU Leuven RDR and ManGO, were set up and made available to researchers, the connection and interoperability between the two infrastructures were established to facilitate the easy publication of research data from ManGO to KU Leuven RDR to make the data as open as possible.

Connecting RDR to RDM Tools

The connection between ManGO and the institutional data repository was the initial use case for the creation of a Dataverse integration dashboard that allows connections with RDM tools that have a data exchange API endpoint. About 6 months after the launch of RDR, the KU Leuven RDR team decided to develop an easier way to import data into a dataset from existing data management systems. This quickly turned into a dashboard that would allow a Dataverse user to pull data from external tools, such as iRODS, OSF, SharePoint, GitLab, and GitHub, into a draft dataset. During the concept phase, it was established that so many more integrations were possible in the dashboard because of the reversal of the usual workflow. Instead of pushing data from an external tool into Dataverse, the dashboard would be pulling data from the external tools into Dataverse. This allowed for more flexibility, because all the tools needed to connect were good API endpoints, which most RDM tools have, and meant that the interface of the dashboard could be designed more flexibly.

The integration dashboard could now pull in data from external tools in a user-friendly and efficient manner. The integration dashboard allows for data comparison between the chosen data source, ManGO in the initial use case, and the dataset the data is intended to be copied to avoid unnecessary data duplication if the data is already available in the RDR dataset. The transfer happens on the server; therefore, users could launch the transfer and then close the applications and wait for email confirmation that the transfer has finished.

This is a workflow that has proven to be very helpful for the publication of large datasets from ManGO to RDR. Because the connection between ManGO and RDR is particularly fast and efficient, due to the two systems using multiple streams in parallel to transfer data and running in the same data centre. The integration dashboard of KU Leuven RDR has been expanded since its initial release with ManGO to include GitHub, GitLab, OSF, OneDrive, SharePoint, Globus, SFTP, and any external non-KU Leuven iRODS instances. The dashboard has been shared in open source so that others could easily add to the list of available tools to pull data from and implement the dashboard for their Dataverse set-up. The only requirement for a connection is that the connected tool has an API point that allows for data streaming and discovery.

On the roadmap for the near future, the integration dashboard is planned to import metadata from as many RDM tools as possible. This could be challenging due to the different metadata available in RDM tools and how this metadata is often not set up and provided with the intention of publication. Therefore, in-depth mapping and stakeholder consultation are necessary to ensure that these metadata imports are correct. Again, the connection with ManGO is easiest, due to the close collaboration between the teams. In addition to further integration with RDM tools, future plans for RDR include expanding the support for publishing research software and code and continuing to improve the overall ease of use, as well as continuing to stay up-to-date with researcher requirements and updating the core Dataverse software to the latest available versions.

Growing ManGO Connections

As mentioned earlier, ManGO is continuously evolving. In addition to the ManGO portal, new in-house developed tools have been added to the ManGO ecosystem to assist researchers with their data management tasks. For example, two tools have been developed for data ingestion. ManGO Ingest is a Python-based tool that seamlessly ingests data into ManGO from any client (e.g., scientific instruments). It includes functionality to monitor a specific directory for new files, automatically upload them to a predefined path, add metadata, and check the integrity of the transfers. SFTPingestor is an iRODS SFTP server designed for ingesting data from untrusted devices (e.g., smartphones and IP cameras) with limited access permissions to the upload zone to safeguard the ManGO itself while allowing easy upload from external devices. KU Leuven also actively participates in the development of the R client for iRODS (rirods), which is currently in the testing phase and will soon be deployed in the production systems.

As for RDR, the interoperability of ManGO with other systems is a major focus for the ManGO team. Improvements to the integration between ManGO and RDR are underway, focusing on an option to push data from ManGO to RDR. This differs from the RDR integration dashboard, because the transfer would be initiated in ManGO instead of RDR. This allows users more familiar with the ManGO environment to start their data publishing process from there, providing a seamless user experience when moving data from one infrastructure to the other. A feature to retrieve data from RDR, or any other Dataverse, to ManGO will be included to make it easier to reuse already published datasets. This development is a collaboration with SURF, the IT cooperative of Dutch education and research institutions[11], to ensure that the connection works for any iRODS and Dataverse combination. All this work, as with other tools (ManGO portal, ManGO metadata schema, and ManGO Ingest), will be made openly available to the iRODS community in open source.[12] This ensures that KU Leuven's investments in RDM tools development could be used by and are accessible to other institutions in line with the spirit of open science.

[11] About SURF: https://www.surf.nl/en/about
[12] ManGO-Portal – A web based front-end to iRODS (GitHub): https://github.com/kuleuven/mango-portal

The roadmap for the coming years includes further enhancing the capabilities of ManGO and RDR, continuing to share work as open source, and creating a cold storage platform integrated with ManGO and RDR for long-term archiving. This will provide users with a set of integrated tools spanning the research data lifecycle, from data creation to publication, archiving, and preservation. The new cold storage system is in its final development phase and will start the pilot phase in March 2025, with production scheduled to begin in 2026. The main goal of KU Leuven's RDM support teams is to ensure a seamless user experience, making the threshold for good research data management and adherence to RDM policies as low as possible.

Conclusion

In 2018, KU Leuven started on the road to an institutional data repository, which resulted in the launch of the Dataverse-based RDR in January 2022. In line with this, there was a strong focus on tools to manage data in a structured and metadata-rich way during the research project, when data was still being worked on. In the context of this aim, an iRODS instance that would enable easy and integrated metadata management from the start of a research project via a self-developed portal was launched in March 2023 under the name ManGO. KU Leuven further supports a wide range of other RDM tools, including OSF, GitLab, REDCap, SharePoint, and DMP-online. This results in a varied landscape of RDM tools that enable and support KU Leuven researchers in good data management. Because one of the four FAIR principles is interoperability, and to fully enable FAIR data management, it was important to facilitate this interoperability of data by ensuring that tools are interoperable, with a first focus being the connection between ManGO and RDR to select and copy data that is ready for publication. This connection was established through the RDR integration dashboard and the work of the ManGO team on connecting with other tools to import data. This interoperability work continues with more metadata exchange and the expansion of tools connected to ManGO and RDR, an example being the development of a cold storage solution for archiving data. In addition, ManGO and RDR continue to provide robust support with their dedicated teams that work on the development, deployment, improvement, and support of both infrastructures. As shown in this paper, KU Leuven is committed to providing a robust and trustworthy ecosystem of RDM tools to support its researchers and their data, while contributing to the broader research infrastructure community by sharing development in open source as much as possible to ensure that the investments that KU Leuven makes in RDM tool development could be used by and is accessible to other institutions in the spirit of open science.

Acknowledgements

The development of ManGO, RDR, and related support services is co-financed by KU Leuven and the Flemish Open Science Board.