IJDC | *General Article*

First-line Research Data Management for Life Sciences: a Case Study

J. Paul van Schayck
DataHub Maastricht
Maastricht University

Maarten Coonen
DataHub Maastricht
Maastricht University

Abstract

Modern life sciences studies depend on the collection, management and analysis of comprehensive datasets in what has become data-intensive research. Life science research is also characterised by having relatively small groups of researchers. This combination of data-intensive research performed by a few people has led to an increasing bottleneck in research data management (RDM). Parallel to this, there has been an urgent call by initiatives like FAIR and Open Science to openly publish research data which has put additional pressure on improving the quality of RDM. Here, we reflect on the lessons learnt by DataHub Maastricht, a RDM support group of the Maastricht University Medical Centre (MUMC+) in Maastricht, the Netherlands, in providing first-line RDM support for life sciences. DataHub Maastricht operates with a small core team, and is complemented with disciplinary data stewards, many of whom have joint positions with DataHub and a research group. This organisational model helps creating shared knowledge between DataHub and the data stewards, including insights how to focus support on the most reusable datasets. This model has shown to be very beneficial given limited time and personnel. We found that co-hosting tailored platforms for specific domains, reducing storage costs by implementing tiered storage and promoting cross-institutional collaboration through federated authentication were all effective features to stimulate researchers to initiate RDM. Overall, utilising the expertise and communication channel of the embedded data stewards was also instrumental in our RDM success. Looking into the future, we foresee the need to further embed the role of data stewards into the lifeblood of the research organisation, along with policies on how to finance long-term storage of research data. The latter, to remain feasible, needs to be combined with a further formalising of appraisal and reappraisal of archived research data.

Submitted 6 April 2021 ~ *Revision received* 1 June 2022 ~ *Accepted* 6 June 2022

Correspondence should be addressed to Maarten Coonen. Postal Address: DataHub Maastricht, PO Box 616, 6200 MD, Maastricht, The Netherlands. Email: m.coonen@maastrichtuniversity.nl

An earlier version of this paper was presented at International Digital Curation Conference IDCC21, Edinburgh, 19 April 2021.

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/

http://dx.doi.org/10.2218/ijdc.v16i1.761
DOI: 10.2218/ijdc.v16i1.761

Introduction

Modern life sciences studies depend on the collection, management and analysis of comprehensive datasets. The focus in life sciences is no longer on the collection of a single sample but on arrays of samples analysed and collected in parallel. Researchers have opted for multimodal approaches in their experiments because multiple techniques are required to reveal better insights into dynamic molecular mechanisms underlying biochemical processes. The limiting factor in research nowadays is the pace at which researchers can analyse their data rather than the amount of samples that can be processed (Mons, 2018).

Life sciences research is often conducted by relatively small research groups. These small teams have sometimes been dubbed "small science," opposed to the "big science" of large-scale astronomy or physics projects (Heidorn, 2008). While this is undoubtedly an oversimplification of reality (Darch and Sands, 2015), the term "small science" can be used to quickly sketch the difficulties faced by these research groups in managing their data. For example, reliable collection and transformation of data into open data formats as well as using community standards for metadata can be a bottleneck when tenured expertise and IT infrastructure are lacking. Thus, overall it has proven difficult for individuals in these small- and medium-sized research labs to manage their ever-growing mountain of research data (Borgman et al., 2016).

Parallel to the increasing difficulty for researchers to manage their data, there has been an increasing urgency to openly publish research data (Borgman, 2012). Initiatives like FAIR and Open Science, among others, have been pushing to share annotated research data openly and in a structured way (Wilkinson et al., 2016). These initiatives are mostly imposed top-down on researchers through, for example, the requirements in data management plans (DMPs). They set very high goals and standards, some of which can feel very far from the current daily practice of researchers (Higmans et al., 2019; McQuilton et al., 2020).

In response to both these initiatives and the challenges faced by researchers, research institutes have formed or strengthened existing research data management (RDM) support groups. These groups are based at academic libraries, IT departments, (bio)-informatics research groups and/or other existing research support structures that institutes may have had in place (Cox et al., 2017). Historically, academic libraries have had a central role in the archiving and curation of research output. For decades, they have been at the forefront of research digitalisation in the form of digital repositories or other digital services. This practice made it logical for academic libraries to also step into the field of RDM, primarily serving an advisory role as opposed to providing technical RDM services (Tenopir et al., 2017; Cox et al., 2019).

One of the key challenges for RDM support groups is translating the aforementioned top-down initiatives into day-to-day practice for researchers, while at the same time not losing touch with the primary goal of moving their research forward. In this case study, we reflect on the lessons learnt by the RDM support group DataHub Maastricht (hereafter DataHub) from Maastricht University and the Maastricht University Medical Centre in the Netherlands. We present a brief history of DataHub and its focus on life science by its support of the data-intensive research institutes in Maastricht. We present several strategies and initiatives that have yielded success, including organising incentives, the use of disciplinary data stewards and how to prioritise RDM support. From a technological perspective we present a method for reducing storage costs, the effect we have had in supporting tailored domain specific platforms and the need for cross-institutional access to services. We also review caveats and future prospects for RDM support in life sciences in general.

Background

Maastricht University is a medium-sized and relatively young (45 years old) university in the south of the Netherlands, with approximately 4,400 employees and 20,000 students. The university has six faculties, with over half its employees located at the Faculty of Health, Medicine and Life Sciences (FHML). This faculty has close links with its next-door neighbour the hospital of Maastricht; together they form the Maastricht University Medical Centre MUMC+.

DataHub (initially dubbed Research IT) was founded in 2015 within the FHML IT support department. Appointed by the Board of MUMC+, DataHub's focus is on supporting and improving RDM for both the hospital and the faculty. DataHub consists of a small core team of data and software engineers. This core team is complemented by disciplinary data stewards, many of whom have joint appointments at DataHub and a research group.

Shortly after being founded, DataHub set out to design and build an infrastructure to support its RDM goals. This infrastructure became the Maastricht Data Repository (MDR). At its core, the MDR was built using the integrated Rule-Orientated Data System (iRODS). Briefly, iRODS provides storage virtualisation in one common directory namespace, authentication and authorisation, combined with flexible metadata on object and collection level all under the control of server-side policies written in either its own rule language or Python (Xu et al., 2017). iRODS can be used as a generic RDM platform or to build highly specialised workflows. It is also known for its capability to handle high volumes of data. For example, the Wellcome Trust Sanger Institute uses iRODS to serve more than 30 PiB of molecular sequencing data to hundreds of internal users (Chiang, 2011; Clapham, 2021).

The use of iRODS has gained much traction in the Dutch RDM landscape over the last five years and a thriving community around it has sprung into being (see for example: Lee et al., 2017, Staiger et al., 2017, Zondergeld et al., 2020). Over 10 institutes are now deploying iRODS, including SURF (the national cooperative association of education and research institutes for digital services) that provides several iRODS-based services. Several Dutch institutes have become iRODS consortium members, and iRODS user conferences have been organised twice in Utrecht (in 2017 and 2019).

The MDR serves as a generic data repository for researchers. Data can be ingested to the MDR by users via so called "drop zones" which are accessible via Windows network shares and WebDAV. Users initiate these drop zones via a web interface where detailed metadata can be added. For certain metadata fields there is an ability to select ontology-controlled terms. Once ingested, a drop zone becomes a collection within a research project and is assigned a persistent Handle identifier. Access to project data is controlled on either a user or group level with three distinct roles (manager, contributor, viewer). Depending on the policy assigned to a project, its data are either stored on premises (university and hospital) or remote on offline tape library. Data ready for publication can be published directly from within the MDR to DataverseNL, a national instance of Dataverse (Crosas, 2011). Several applications make use of the different iRODS APIs for domain specific workflows implemented on top of the MDR.

In the five years after the initial conception of the MDR, DataHub broadened its scope of RDM support. One of the challenges faced by the DataHub team was to position the MDR not only as endpoint for inactive data but also as an active repository in all phases of the research data life cycle. Different research domains have different requirements for the active phases of the research data life cycle, which DataHub needed to consider (detailed in Technological Lessons below) when providing guidance and support.

As of this writing, the MDR hosts 276 TiB of data, across 272 research projects and is used by 339 researchers (Figure 1) from approximately 10 different research departments within MUMC+. Two examples are the Maastricht Multi-Modal Molecular Imaging Institute (M4i) and the MERLN Institute for Technology-Inspired Regenerative Medicine. M4i uses imaging mass spectrometry and cryogenic electron microscopy to study the molecular world, while MERLN uses high-content screening microscopy to study the interaction between biomaterials and tissue. Both institutes require very data-intensive research methods but have different RDM needs. As such, they both approached DataHub for RDM guidance and support.

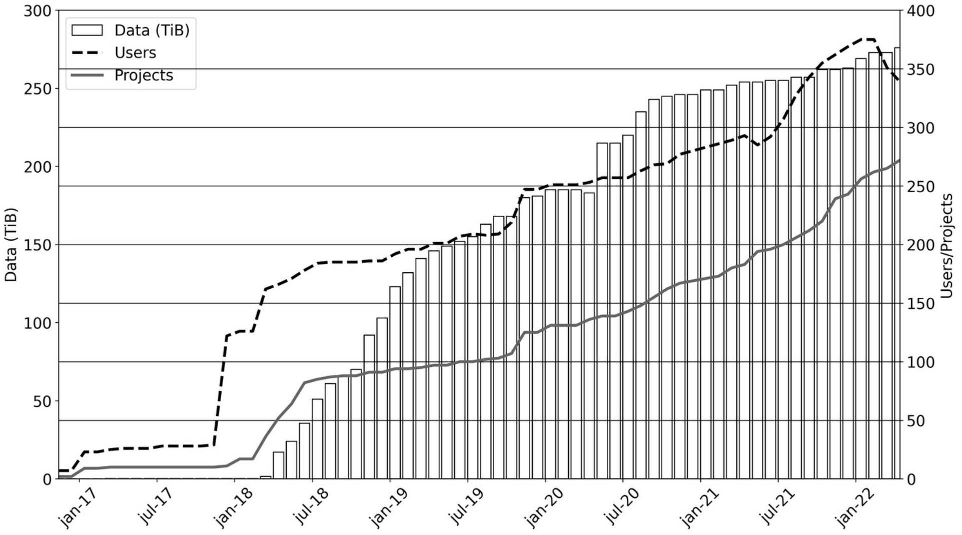

Figure 1. Growth of the Maastricht Data Repository over 5 years by data (bars), research projects (solid curve) and users (dashed curve).

Organisational lessons

RDM and its related support do not take place in isolation. They are part of the broader research support ecosystem of an institute. Typically, the support that can be provided is limited by the available personnel and time, in turn often dictated by financial limits. We identified the following lessons in building and improving the organisational model of DataHub.

Appointing disciplinary data stewards

A data steward can be defined as a person responsible for keeping the quality, integrity, and access arrangements of data and metadata in a manner that is consistent with applicable law, institutional policy, and individual permission (Jetten et al., 2021). In recent years, the role of the data steward has become more and more clearly defined within the RDM and Open Science community and is recognised as a key instrument for moving RDM forward (Versweyveld, 2016). Due to the diversity of requirements between, or even within, research domains in the life sciences, there is also a need for disciplinary data stewards (Teperek et al. 2018). Mons already stated this need for a high granularity of data stewards and recommended to strive towards one data steward for every 20 researchers (Versweyveld, 2016).

The DataHub team has made efforts to encourage the embedding of disciplinary data stewards at different research groups. DataHub has strived to embed data stewards at the level of the research group rather than the faculty; currently, there are about 10 data stewards employed for about 300 researchers. The embedding has taken place in a variety of ways, but preferably in the form of a shared position between the research group and DataHub. Stewards have often been recruited from the ranks of the research groups themselves, but occasionally also newly filled positions have been realised. Due to the mutual benefit for DataHub and the research group, this has always been possible to achieve in a funding neutral way.

The shared placement has had two important effects: (1) the data steward is able to help researchers with their day-to-day practice of data management, data analysis and general IT issues and (2) the data steward possesses or obtains valuable domain-specific knowledge and can translate this knowledge to and from the DataHub infrastructure. Overall, the data stewards have a signalling role by providing early coaching of researchers into RDM. Our experiences show that DMPs can provide good first points of contact to talk about RDM between the data steward and the researcher. Data stewards are encouraged to make use of DMPMaastricht, an instance of DMPOnline (Getler et al., 2014), for this purpose.

In addition to supporting researchers, data stewards act as stakeholder for DataHub's research software engineers, who work according to the Scrum framework (Schwaber and Beedle, 2002). Scrum is a software development methodology where development takes place incrementally and focus on user value is paramount in all phases of the process. It is a way to get business requirements delivered into working code effectively. In Scrum there are multiple roles defined: the product owner is taking care of stakeholder management and transforming business ideas into user stories, the development team is committed to turn user stories into working code or features and the Scrum master is guiding the process and resolving any impediments. Data stewards, in their role as stakeholders, work together with the product owner to define and select the user stories that bring the most user (i.e., researcher) value to improvements of the MDR.

Providing incentives to start RDM

The benefits of organising and sharing data are not always immediately clear, and the gains to be made are sometimes not obvious (Wilms et al., 2020). Regularly, the benefits of RDM only become obvious months or even years after implementation. As famously said by the archivist Jason Scott "metadata is a love note to the future". This "love note" does not only apply to metadata but to research data in general. Short-term members of research teams will have moved on before seeing the benefits of the efforts they put into organising and/or sharing their data (Tenopir et al., 2011; Doucette and Fyfe, 2013). Within small research groups, where the bulk of data is generated by PhD students with fixed short-term contracts, this limited time perspective can especially be dominant.

Over the years, we have learnt that it is instead most effective to incentivise the principal investigators (PI) of research groups to instigate RDM. With the PI of a group on-board and directing this policy, it is much easier to get her/his PhD students to participate. A clear incentive for the PIs can be a financial one, for example, in the reduction of storage costs (see technological lesson "Reducing storage costs through tiered storage").

We have also seen potential pitfalls in the hope to realise new incentives for researcher. We explored the possibility of making a certain technology or feature in the RDM service available as a possible incentive to users to start RDM. Thus far, we have experienced this strategy to be ineffective. For example, we implemented a fully featured semantic search engine, using Ontoforce DISQOVER, on top of the MDR to allow researchers to easily find and retrieve datasets, both their own and of their colleagues. While this feature added value to existing users, it did not propel researchers to initiate RDM and MDR usage. Mainly because finding someone else's data does not encourage you to add your own necessarily. Another example is that the DataHub RDM infrastructure, when used properly, could provide a track-record of all the data's stages in the research data life cycle. Whereas this traceability is important to the Faculty Board, we found that researchers did not use/benefit from it and some reported it just to be a burden.

While single technological features could be a stimulus for RDM initiation, the two we looked at here were generally not effective (as incentive). We will continue to identify others in the future that could be more appealing, and encourage other RDM services to do so.

Reducing bureaucracy: no wrong door policy

For a researcher, RDM can be experienced as yet another (bureaucratic) topic to be covered while doing their 'real work' (Wilms et al., 2020). Similar to any information they need to provide about legal, ethical or (bio)-safety concerns for their research, RDM can just feel like more paperwork. Not knowing where to go for RDM guidance might add to the researcher's feeling about RDM bureaucracy. We have learnt to focus on minimising this feeling by adapting the policy of "no wrong door."

At MUMC+, various groups provide RDM support in one form or another. These include (but are not limited to) the university library, a software engineering department, a data science research group, a clinical trial centre, and ourselves at DataHub. By intergroup collaboration, shared personnel between these groups and active encouragement, we established clear channels and an atmosphere where researcher's requests are quickly directed to the right person at the most appropriate group to handle the solution.

An alternative approach could be to create an overarching layer or support desk to handle all first-line requests. However, this structure does not align with the decentralised disciplinary data stewards who also have flexible roles between the various groups that provide RDM support. It may turn out in the future that an overarching support desk may still be beneficial, but this would require close collaboration with the same data stewards.

Prioritising efforts on the most reusable data

The RDM community is at a stage where, for the foreseeable future, far more research data are being generated than can realistically be supported and assisted in making fully FAIR and Open. Making data truly interoperable on a Linked-Data level, as intended by the FAIR principles, is a daunting task.

In addition, at DataHub more support requests are coming in than can be handled. In effect, it means that choices have to be made as to which research or datasets are to be supported and which are not. DataHub decided to use the criterion of a dataset's potential reuse value once published; that is, additional work put into making such a dataset more FAIR or Open would be greatly amplified in the future by its reuse.

DataHub has taken several approaches to keep the focus on the potential reuse of data: (1) encourage the search for and use of domain-specific repositories. These often offer data type specific functionality and visualisation, provide better indexes and mandate more detailed metadata when compared to generic repositories. Submission to such a repository will automatically make data more FAIR, even if this means only part of the research data of a study is published; (2) Identify, through the embedded data stewards, the research projects that may have the highest potential for future reuse. This early knowledge allows RDM to be designed and implemented as the data are being generated; (3) Use an approach (here: Agile/Scrum) that directs software development efforts to features that have the highest user value. Additionally, one is encouraged to critically think how new features contribute to the effectiveness of their RDM services.

The topic of prioritising efforts on reusable data is related to the question Christine Borgman asked: "If data sharing is the answer, what is the question?". She asked the rightful question whether the effort of making data available is worth the effort put into it. It should remind us that in the triangle of RDM, FAIR and Open Science we should keep the use to which data can be put to at the heart of our activities (Higman et al., 2019).

Technological lessons

We have learnt that establishing and providing appropriate technology for RDM are essential, but very challenging. The requirements for researchers within different life science domains are very diverse and can even differ greatly within the same research group. At the same time, any technology has limited flexibility to support this diversity of use.

Reducing storage costs through tiered storage

A clear incentive for researchers to start RDM is reducing their data storage costs. As is common worldwide, Maastricht researchers often pay directly for storage costs from the research grants they obtain. The consequence of this is that researchers often choose for the lowest costs and store their data on external hard drives or self-managed network storage solutions, which is the opposite of good RDM practice (Tenopir et al., 2011). This decision makes it far more likely for research data to become unavailable after publishing (Abrams, 2016; Ashiq et al., 2020). Therefore, opportunities to reduce storage costs with the additional benefit of RDM functionality will be attractive.

At DataHub, we have formulated this financial incentive in two ways. Firstly, we have prioritised the principle that storage is a public utility, similar to water and electricity, and should be heavily subsidised by the faculties. Up to 100 GiB of storage is free of charge. Above this, storage is offered at cost price or lower, while additional RDM features on top of this are free. Secondly, we have developed a tiered storage system. Most networked storage is expensive because it keeps all data available at high speed at all times. For RDM purposes, data can often be migrated ("tiered") to less expensive and slower performing storage. However, this option must remain transparent in use and easy to execute to be an attractive choice for researchers.

Using iRODS, it was possible to meet these requirements and implement this as functionality of the MDR (Figure 2). Data is stored long-term on the tape library of SURF in Amsterdam, which is offered at a very competitive price. Due to the way tape storage functions, it cannot handle datasets with many small files. Usually, this is overcome by bundling small files together, for example in tarballs. We chose to implement an iRODS policy to migrate only files over 256 MiB to tape, which results in datasets that are stored mixed between remote tape and on on-premise solutions. The choice for 256 MiB was made based on technical grounds of the tape archive. However, for the whole MDR, 70% of the volume of data is contained in files over 256 MiB of size, demonstrating that bulk of the data storage is at the most economical tier. Transparency for end users is maintained: they still see their datasets, presented by the catalogue provider, in the same way as before. Currently, dataset tiering is manually triggered by project managers in the web interface of the MDR. Looking into the future, DataHub would like to extend this functionality to perform tiering automatically based on access time.

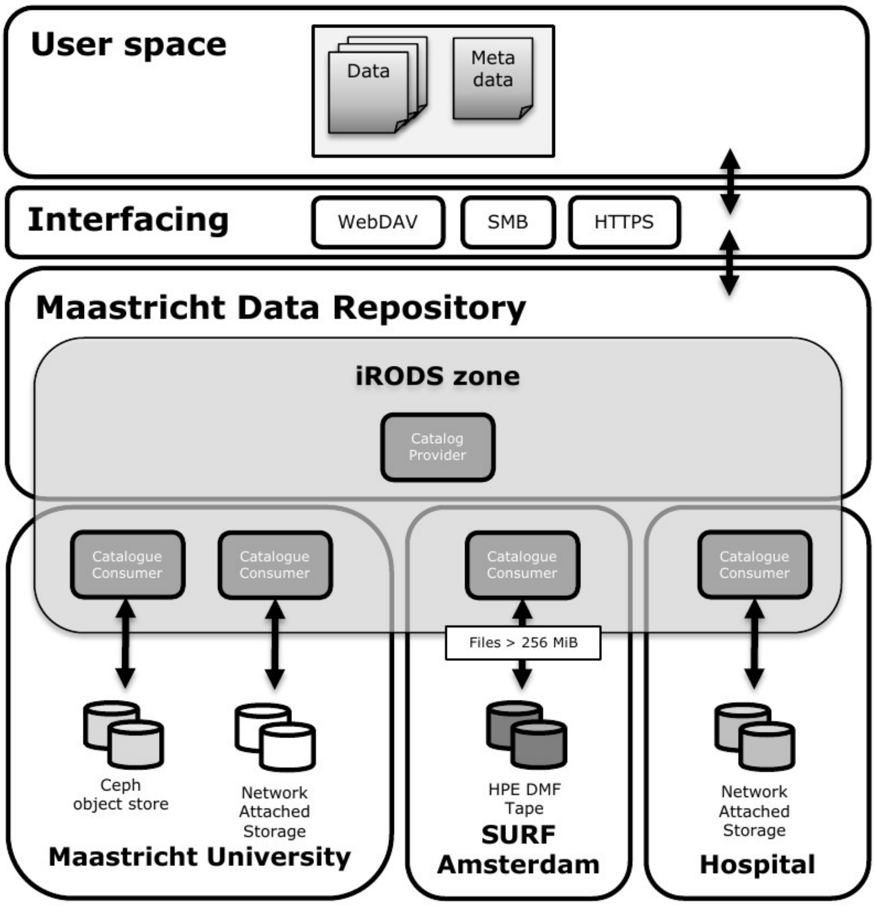

Figure 2. The Maastricht Data Repository uses iRODS for the transparent tiering of data across multiple storage solutions and geographical locations (bottom row). Data and its metadata are organised in projects and collections in iRODS and can be stored and retrieved by different interfaces: SMB, WebDAV or HTTPS. The relational database in iRODS's catalogue provider stores system metadata, logical paths and authorisations, among other things. Data storage is distributed via different catalogue consumers to various on-premise or remote locations based on rules and user requirements. One of these rules (policies) is that only files over 256 MiB in size can be stored on tape.

Supporting data-intensive and diverse life science research

Life science research has become more data intensive in recent years. Dealing with multi-terabyte datasets with millions of files places more strain on all IT infrastructures, especially bandwidth and storage capacity as well as optimised user experience of the tools that are provided. For example, performing uploads and downloads through a web browser is not realistic for terabyte-size datasets.

One of the answers to this challenge was building many of our RDM workflows on top iRODS, which enabled us to support a high volume of data. For example, data transfers during the ingest process go via Windows network shares, thereby bypassing the web browser. However, the generic nature of iRODS limits its functionality, especially for the needs of the diverse subdomains of the life sciences: it does not allow visualisation or specific file format support.

As an additional solution, we implemented support for domain-specific RDM platforms. Various very successful RDM platforms exist for different life science subdomains, for example, XNAT for radiology, OMERO for microscopy and MOLGENIS for molecular genetics (Marcus et al., 2007; Li et al., 2016; van der Velde et al., 2019). These three platforms are successful in their respective domains: i.e., they are providing established community standards and/or methods for researchers to share and manage their research data. While these platforms are regularly setup by research groups on a project basis, long-term support in both funding and expertise is required for continued success.

In implementing these domain-specific platforms, we found that different expertise was required at different levels. DataHub provided support where professional IT skills are required, while the data stewards mainly interacted with the researchers and developed specific workflows on these platforms. We have used this model successfully for XNAT for multiple research groups and are now in the process of implementing it for OMERO.

Obtaining long-term funding is still a challenge for these sorts of local, very domain-specific infrastructural RDM platforms. However, recently, the Dutch Science Foundation (NWO) has started to recognise the need for this as well in their aim to support Open Science. Through their large infrastructure roadmap grant-scheme DataHub has been able to secure funding for supporting OMERO for the next few years.

Offering cross-institutional collaboration

Researchers engaged in international collaborations are functionally limited by software their institute provides (e.g., network shares not accessible by third parties). As a solution, federated access has been developed separately for several RDM platforms. In addition to being very costly, this process also led to suboptimal user experiences because users are burdened with learning new interfaces, and/or remembering/storing accounts for all the services they use (Linden et al., 2018). Researchers thus often resort to consumer (cloud) sharing applications for which governance does not conform to their institute's policies or national privacy policies.

In response, the European e-infrastructure collaboration GÉANT, among others, has put considerable effort in the development of EduTEAMS to offer generic solutions for federated authentication and authorisation. Several efforts to incorporate these technologies are underway in the Netherlands. One of such is SURF Research Access Management (SRAM) (Figure 3), which provides the generic federated authentication proxy through EduTEAMS. In SRAM the collaborative organisation (CO), which can be a research consortium or a research group, is the functional centrepiece of the system. COs are given a digital home in SRAM where CO-members can see the services they collaborate in, while data stewards can manage these members and add services to their CO. DataHub has been collaborating with SURF in the design and implementation of SRAM and has taken SRAM into production as its federated authentication provider and user/group management for the MDR in 2021. Thereby, secure data collaboration between researchers from trusted identity providers has been realised.

Looking broader, there are several parallel developments taking place in the access and authorisation infrastructures space, notably the Elixir-AAI project, the European Open Science Cloud and Internet2's COManage/CILogon (Linden et al., 2018, Basney et al., 2019). Making services quickly available for international access and offering simple self-service group management are key features to effectively support research groups engaged in international collaborations, as typically found in the life sciences.

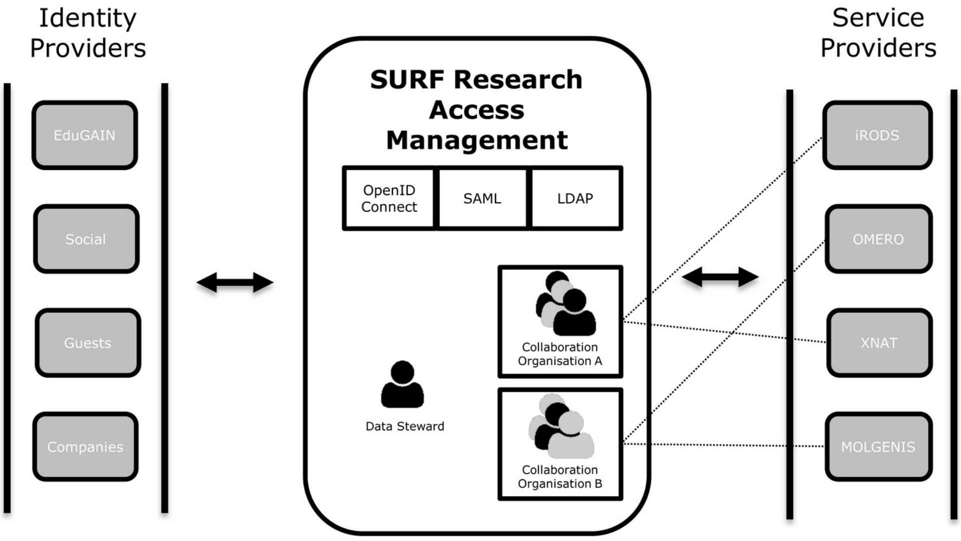

Figure 3. SURF Research Access Management (SRAM) is a federated authentication proxy with self-service group management. It proxies an existing identity from, for example, an institutional account via OpenID Connect, SAML or LDAP to a service provider. Users are organised by data stewards into collaboration organisations. Access to service providers can be configured per collaboration organisation.

Conclusion

Our experience shows a few key lessons for effective institutional first-line RDM support for the life sciences. Firstly, the requirements for effective RDM differ considerably between the different life science domains, affecting the technology and infrastructure to be provided and preventing generalised solutions. This diversity is reflected and addressed in the use of the disciplinary data stewards, but also in the need to provide professional IT support for very tailored RDM platforms. Secondly, prioritising where/how to provide RDM support is an effective way to deal with limited support resources. At DataHub, we chose to focus on research data that are most reusable. This focus can be achieved by disciplinary data stewards knowing the research projects that have the highest potential reuse of their data. Thirdly, finding incentives to motivate researchers to work on their RDM, with the exception of a financial one, remains a challenging task. We found that offering technological features have little effect in motivating RDM uptake. Our limited success with incentives may lead to the conclusion that implementing mandatory requirements and/or punitive measures for RDM practice may still be required. Finally, RDM must allow for cross-institutional collaboration by default by using the appropriate authorisation and authentication infrastructure.

Outlook

Looking ahead, we identified several technological and organisational points that will require attention or will play an increasingly important role in the future for institutional RDM support.

Technological

A topic regularly overlooked in RDM is the archival and maintenance of research software (Howison and Bullard, 2016). Without the software used to analyse the data, it may be impossible to actually reuse the data. This is still a very challenging topic with many possible pitfalls in its solutions, but it needs more attention. Additionally, definitive metadata schemas, even for a particular domain, do not exist and will likely keep changing over time. Currently, only partial technical solutions, in terms of entry, migration, search and presentation, exist to deal with these ever-changing metadata schemas (Philipson, 2020). We are also exploring and developing this further in the use of flexible metadata schemas in iRODS (van Schayck et al., 2019).

Organisational

Our findings underline the important role of data stewards to facilitate RDM embedding in the organisational structure. As such, ongoing efforts to professionalise the role of data stewards at institutions in the Netherlands (Jetten et al., 2021) are critical. In particular, training paths and career perspectives should be defined for data stewards. Secondly, the question of who finances the long-term storage of research data needs to be answered clearly. Currently, in Maastricht, there is an ongoing discussion about the benefits of a clear financial model to pay for the long-term storage of research data. One example model is a discounted, one-time, lump-sum payment for data storage at the end of the project. Finally, data stored must undergo a constant form of curation. At the current rate of data growth, published or unpublished, it is impossible to store everything indefinitely. Therefore, constant appraisal and reappraisal of stored research data, with a continued focus on reusable data, should be a priority for institutional RDM. This practice is only feasible with strong RDM in place and in particular better metadata about the data.

Methods

This paper is based on five years of participant observation by the authors. The authors' findings were further validated by using the results from qualitative semi-structured interviews with key people around DataHub Maastricht. A total of four respondents were interviewed, comprising two long-term DataHub core team members, one Faculty Board member and one data steward. The recordings of the interviews have been deposited in the Maastricht Data Repository and are available upon request (van Schayck, 2021).

Acknowledgements

This work has been supported by the Limburg Invests in its Knowledge Economy (LINK) programme from the Province of Limburg, the Netherlands. We would like to thank the respondents of our interviews for their cooperation. We are very grateful to Hang Nguyen and Raimond Ravelli for critically reading this manuscript.

References

Abrams, S., Kratz, J., Simms, S., Strong, M., & Willett, P. (2016). Dash: Data Sharing Made Easy at the University of California. International Journal of Digital Curation, 11(1), 118--127. https://doi.org/10.2218/ijdc.v11i1.408

Ashiq, M., Usmani, M. H., & Naeem, M. (2020). A systematic literature review on research data management practices and services. *Global Knowledge, Memory and Communication*, *ahead-of-p*(ahead-of-print). https://doi.org/10.1108/GKMC-07-2020-0103

Basney, J., Flanagan, H., Fleury, T., Gaynor, J., Koranda, S., & Oshrin, B. (2019). CILogon: Enabling Federated Identity and Access Management for Scientific Collaborations. *Proceedings of International Symposium on Grids & Clouds 2019 — PoS(ISGC2019)*, *351*, 031. https://doi.org/10.22323/1.351.0031

Borgman, C. L. (2012). The conundrum of sharing research data. *Journal of the American Society for Information Science and Technology*, *63*(6), 1059–1078. https://doi.org/10.1002/asi.22634

Borgman, C. L., Golshan, M. S., Sands, A. E., Wallis, J. C., Cummings, R. L., Darch, P. T., & Randles, B. M. (2016). Data Management in the Long Tail: Science, Software, and Service. *International Journal of Digital Curation*, *11*(1), 128–149. https://doi.org/10.2218/ijdc.v11i1.428

Chiang, G.-T., Clapham, P., Qi, G., Sale, K., & Coates, G. (2011). Implementing a genomic data management system using iRODS in the Wellcome Trust Sanger Institute. *BMC Bioinformatics*, *12*(1), 361. https://doi.org/10.1186/1471-2105-12-361

Clapham, P. (2021). Informatics Support Group – Wellcome Sanger Institute. Retrieved April 3, 2021, from: https://www.sanger.ac.uk/group/informatics-support-group/

Cox, A. M., Kennan, M. A., Lyon, E. J., Pinfield, S., & Sbaffi, L. (2019). Progress in Research Data Services. *International Journal of Digital Curation*, *14*(1), 126–135. https://doi.org/10.2218/ijdc.v14i1.595

Cox, A. M., Kennan, M. A., Lyon, L., & Pinfield, S. (2017). Developments in research data management in academic libraries: Towards an understanding of research data service maturity. *Journal of the Association for Information Science and Technology*, *68*(9), 2182–2200. https://doi.org/10.1002/asi.23781

Crosas, M. (2011). The Dataverse Network: An Open-source Application for Sharing, Discovering and Preserving Data. D-Lib Magazine, Volume 17. Retrieved from http://www.dlib.org/dlib/january11/crosas/01crosas.html

Darch, P. T., & Sands, A. E. (2015). *Beyond Big or Little Science: Understanding Data Lifecycles in Astronomy and the Deep Subseafloor Biosphere*. Retrieved from http://hdl.handle.net/2142/73655

Doucette, L., & Fyfe, B. (2013). Drowning in Research Data: Addressing Data Management Literacy of Graduate Students. *Imagine, Innovate, Inspire: The Proceedings of the ACRL 2013 Conference*. Retrieved from http://www.ala.org/acrl/sites/ala.org.acrl/files/content/conferences/confsandpreconfs/2013/papers/DoucetteFyfe_Drowning.pdf

Getler, M., Sisu, D., Jones, S., & Miller, K. (2014). DMPonline Version 4.0: User-Led Innovation. *International Journal of Digital Curation*, *9*(1), 193–219. https://doi.org/10.2218/ijdc.v9i1.312

Heidorn, P. B. (2008). Shedding Light on the Dark Data in the Long Tail of Science. *Library Trends*, *57*(2), 280–299. https://doi.org/10.1353/lib.0.0036

Higman, R., Bangert, D., & Jones, S. (2019). Three camps, one destination: the intersections of research data management, FAIR and Open. *Insights the UKSG Journal*, *32*(1). https://doi.org/10.1629/uksg.468

Howison, J., & Bullard, J. (2016). Software in the scientific literature: Problems with seeing, finding, and using software mentioned in the biology literature. *Journal of the Association for Information Science and Technology*, *67*(9), 2137–2155. https://doi.org/10.1002/asi.23538

Jetten, M., Grootveld, M., Mordant, A., Jansen, M., Bloemers, M., Miedema, M., & Gelder, C. W. G. van. (2021). *Professionalising data stewardship in the Netherlands. Competences, training and education. Dutch roadmap towards national implementation of FAIR data stewardship*. https://doi.org/10.5281/ZENODO.4623713

Lee, H.-C., Oostenveld, R., Boogert, E. van den, & Maris, E. (2017). Neuroimaging Research Data Life-cycle Management. *IRODS User Group Meeting 2017 Proceedings*, 17–19. Retrieved from https://irods.org/uploads/2017/irods_ugm2017_proceedings.pdf#page=17

Li, S., Besson, S., Blackburn, C., Carroll, M., Ferguson, R. K., Flynn, H., … Swedlow, J. R. (2016). Metadata management for high content screening in OMERO. *Methods*, *96*, 27–32. https://doi.org/10.1016/j.ymeth.2015.10.006

Linden, M., Prochazka, M., Lappalainen, I., Bucik, D., Vyskocil, P., Kuba, M., … Nyrönen, T. (2018). Common ELIXIR Service for Researcher Authentication and Authorisation. *F1000Research*, *7*, 1199. https://doi.org/10.12688/f1000research.15161.1

Marcus, D. S., Olsen, T. R., Ramaratnam, M., & Buckner, R. L. (2007). The extensible neuroimaging archive toolkit. *Neuroinformatics*, *5*(1), 11–33. https://doi.org/10.1385/NI:5:1:11

McQuilton, P., Batista, D., Beyan, O., Granell, R., Coles, S., Izzo, M., … Sansone, S.-A. (2020). Helping the Consumers and Producers of Standards, Repositories and Policies to Enable FAIR Data. *Data Intelligence*, *2*(1–2), 151–157. https://doi.org/10.1162/dint_a_00037

Mons, B. (2018). *Data Stewardship for Open Science*. https://doi.org/10.1201/9781315380711

Philipson, J. (2020). The Red Queen in the Repository. *International Journal of Digital Curation*, *15*(1), 16. https://doi.org/10.2218/ijdc.v15i1.646

Schwaber, K., & Beedle, M. (2002). *Agile software development with Scrum*.

Staiger, C., Smeele, T., & van Schip, R. (2017). *A national approach for storage scale-out scenarios based on iRODS*. 55–63. Retrieved from https://irods.org/uploads/2017/irods_ugm2017_proceedings.pdf#page=59

Tenopir, C., Allard, S., Douglass, K., Aydinoglu, A. U., Wu, L., Read, E., … Frame, M. (2011). Data Sharing by Scientists: Practices and Perceptions. *PLoS ONE*, *6*(6), e21101. https://doi.org/10.1371/journal.pone.0021101

Teperek, M., Cruz, M. J., Verbakel, E., Böhmer, J., & Dunning, A. (2018). Data Stewardship addressing disciplinary data management needs. *International Journal of Digital Curation*, *13*(1), 141–149. https://doi.org/10.2218/ijdc.v13i1.604

Van Der Velde, K. J., Imhann, F., Charbon, B., Pang, C., Van Enckevort, D., Slofstra, M., … Swertz, M. A. (2019). MOLGENIS research: Advanced bioinformatics data software for non-bioinformaticians. *Bioinformatics*, *35*(6), 1076–1078. https://doi.org/10.1093/bioinformatics/bty742

Versweyveld, L. (2016). We need 500.000 respected data stewards to operate the European Open Science Cloud - News blog - e-Infrastructures Reflection Group. Retrieved January 15, 2021, from http://e-irg.eu/news-blog/-/blogs/we-need-500-000-respected-data-stewards-to-operate-the-european-open-science-cloud

van Schayck, J. P. van, Smeele, T., Theunissen, D., & Westerhof, L. (2019). Providing validated, templated and richer metadata using a bidirectional conversion between JSON and iRODS AVUs. IRODS User Group Meeting 2019 Proceedings, 9--17. Retrieved from https://irods.org/uploads/2019/irods_ugm2019_proceedings.pdf

van Schayck, J. P. (2021). Background interviews for "First-line research data management for the life sciences: a case study." *Maastricht Data Repository*. https://hdl.handle.net/21.12109/P000000190C000000001

Wilkinson, M. D., Dumontier, M., Aalbersberg, Ij. J., Appleton, G., Axton, M., Baak, A., … Mons, B. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, *3*(1), 160018. https://doi.org/10.1038/sdata.2016.18

Wilms, K. L., Stieglitz, S., Ross, B., & Meske, C. (2020). A value-based perspective on supporting and hindering factors for research data management. *International Journal of Information Management*, *54*, 102174. https://doi.org/10.1016/j.ijinfomgt.2020.102174

Xu, H., Russell, T., Coposky, J., Rajasekar, A., Moore, R., de Torcy, A., … Chen, S.-Y. (2017). iRODS Primer 2: Integrated Rule-Oriented Data System. *Synthesis Lectures on Information Concepts, Retrieval, and Services*, *9*(3), 1–131. https://doi.org/10.2200/S00760ED1V01Y201702ICR057

Zondergeld, J. J., Scholten, R. H. H., Vreede, B. M. I., Hessels, R. S., Pijl, A. G., Buizer-Voskamp, J. E., … Veldkamp, C. L. S. (2020). FAIR, safe and high-quality data: The data infrastructure and accessibility of the YOUth cohort study. *Developmental Cognitive Neuroscience*, *45*, 100834. https://doi.org/10.1016/j.dcn.2020.100834