Neuroimaging Research Data Life-cycle Management

Hurng-Chun Lee
Donders Institute,
Radboud University
h.lee@donders.ru.nl

Robert Oostenveld
Donders Institute,
Radboud University
Karolinska Institute
r.oostenveld@donders.ru.nl

Erik van den Boogert
Donders Institute,
Radboud University
e.vandenboogert@donders.ru.nl

Eric Maris
Donders Institute,
Radboud University
e.maris@donders.ru.nl

ABSTRACT

Research Data Management (RDM) aims to improve the efficiency and transparency in the scientific process and to fullfil the requirements of the funding agencies and (local) regulations. Failures in reproducing some key empirical phenomena have resulted in the research process being questioned. In response to this, many research institutions have prioritized the development of RDM. In the neuroscience and neuroimaging domains, RDM is confronted with challenges in managing large volume and diverse data, which furthermore may contain sensitive personal information. At the Donders Institute (DI), we have developed an iRODS-based research data repository, an essential component for realising a RDM workflow that spans the whole research lifecycle. The objectives of this workflow are: (1) long-term data preservation for internal reuse, (2) documenting the analysis pipeline, allowing for collaboration, and reproduction of the published results, and (3) easy sharing of data and analysis pipelines with researchers around the world.

Keywords

Research data management, research life-cycle, neuroimaging data.

INTRODUCTION

Over the last ten years, researchers have experienced an increase in the pressure (by funding agencies, scholarly journals, and universities) to share their research data and to document the process via which they obtained their published results. This pressure asks for a Research Data Management (RDM) protocol and, given that almost all research data are stored digitally, a set of IT services that are needed to implement this protocol. In this paper, we describe such a RDM protocol and the associated IT services.

Research data are all the information that is (1) generated as a part of the research process and (2) on which a scientific report is/will be based. This definition of research data does not only include empirical data, but also simulated data, computer scripts for analysis and simulations, stimuli presented in experiments and the computer scripts for presenting them, etc. These data are generated in different stages of the research process, beginning by acquisition and ending by sharing. The RDM protocol that is described here pertains to the documentation of the data that are generated at *all* stages of the research process. Therefore, it is called *life-cycle RDM,* and it is to be compared with RDM restricted to the time of publication. Life-cycle RDM has the advantage that the act of documenting data becomes an integral part of the research process, with the sharing of documented data being a natural endpoint.

We implemented a particular protocol for life-cycle RDM in a large (600 researchers) and heterogeneous neuroscience institute, the Donders Institute (DI) for Brain, Cognition and Behavior. The DI acquires data of many

iRODS UGM 2017, June 13-15, 2017, Utrecht, Netherlands

different types, such as magnetic resonance imaging (MRI), different electrophysiological signals, whole-genome DNA, proteomics and transcriptomics data, limb movement trajectories, behavioural data, questionnaires, etc. These data are collected using a diverse set of instruments and machines, which often store the data in proprietary file formats. In addition, these data are analyzed using a diverse set of software tools, of which many also store their output in a proprietary file format.

In the following, we first describe our RDM protocol. This is followed by a description of the IT environment that is used to implement this protocol. We conclude this paper by giving an overview of the strengths and weakness of our RDM method, which is partially based on experiences with the method in a production environment.

THE RESEARCH DATA MANAGEMENT PROTOCOL

The objective of the Donders Institute (DI) Research Data Management (RDM) protocol is threefold: (1) data preservation for institute-internal reuse, (2) documenting the analysis pipeline, allowing for collaboration and reproduction of the published results, and (3) easy sharing of data and analysis pipelines with researchers around the world. To realise these objectives, three types of collections[1] have been defined: (1) Data Acquisition Collections (DACs), (2) Research Documentation Collections (RDCs), and (3) Data Sharing Collections (DSCs). These three types also correspond to three different phases of the research data life cycle: acquisition, analysis and reporting, and sharing. We have written a protocol that specifies how these collections are initiated, managed, built, closed and shared. Starting from this protocol, we have designed and built a digital infrastructure, the Donders Research Data Repository (DRDR) [1].

Collections are considered as resources that are provided to researchers upon request, just as lab space, access to MRI scanners, computing resources, research assistant hours, etc. These resources are managed at the level of a so-called Organisational Unit (OU). The DI has four OUs, that all manage their own resources. Collections are initiated by a member of the OU's administrative staff, the so-called Research Administrator (RA). Each collection belongs to a single OU, namely the OU of the RA that has initiated this collection. Upon initiation, disk quota is assigned to the collection, and the requesting researcher is granted the authorization to manage this collection (see further). Every collection includes a required attribute that specifies the research project with which this collection is associated. As a consequence, every OU must have an administrative system in place in which research projects are uniquely identified and in which resources are organised accordingly. Most OUs already had such a system in place, linking research projects to financial budgets.

Upon collection initiation, one or more researchers are assigned to this collection as so-called managers. From the perspective of the OU, these managers are responsible for building and curating the collection. This explicit responsibility implies that only members of that OU can be assigned as a manager. Within the DRDR, managers are authorized to assign other users to their collections. These other users can be assigned as manager, contributor or viewer. These roles map onto the familiar LINUX-authorizations own, write and read. A viewer can only read/download a collection's files, a contributor can also add/delete/modify these files, and a manager can also assign/remove other users to/from the collection (in a particular role). For a user to be assigned as manager to some of the OU's collections, the user has to be flagged as "eligible manager" in that OU[2]. As a contributor or viewer a user does not have to be specifically linked to one of the OUs. This feature accommodates on the one hand the formal responsibility (within the OU) and at the same time the frequent collaborations between researchers that belong to different OU's, or between researchers of which only a subset belongs to a OU that is represented in the DRDR.

[1] The term "collection" used in this paper refers to the DRDR collection. It should be distinguished from the iRODS collection. The DRDR collection is a conceptual container of folders and files. In practise, it is implemented as a iRODS collection in certain namespace hierarchy.

[2] The list of users eligleble to be manager within an OU is maintained by the RA of that OU.

The IT-system that implements the DRDR does not enforce how collections are to be built; this is rather specified in a written protocol. This written protocol specifies the collections primarily in functional terms (What should a collection viewer be able to do with it?), rather than in operational terms (Which files of which type must go in the collection, what must they contain, and how must they be organised?). Specifically, the protocol specifies that a DAC must contain all raw data (with "raw" meaning "without any manipulations that limit future analyses of these data") plus a description that would allow a well-informed colleague to make sense of the data. The written protocol is augmented by online documentation [2] in the form of a Frequently Asked Questions (FAQ) page, which contains concrete suggestions on how this functional requirement can be realised. Further, the protocol specifies that an RDC must document the scientific process, allow for the sharing of preliminary results within the project team (i.e. co-authors of a publication), and document the editorial and peer-review process. Again, this is augmented by a FAQ page that, in this case, contains concrete suggestions on how this documentation can be realised (e.g., by uploading analysis scripts). Finally, the protocol specifies that a DSC must contain all the relevant information (1) to reproduce the published results, and (2) to extend on these published results; this is again augmented by a FAQ page.

When a collection is complete, it can be closed, after which it cannot be changed anymore. Only a manager can close a collection. Closure of both a DAC and a RDC differs from closure of a DSC, this is crucial given that the DRDR can be used both for institute-internal research data management (realised by DACs and RDCs) and for sharing of data with the rest of the world (realised by DSCs). Specifically, only when closing a DSC, a persistent identifier is created via which this DSC can be accessed over the web. When closing a DAC or RDC, no persistent identifier is created.

Accessing data from a closed DSC cannot be done anonymously. To access a collection, users are required to be registered in the DRDR. To facilitate data sharing with a wide audience, the system allows users to register with the credentials of a social ID (e.g. LinkedIn, Google, Facebook, Twitter, Microsoft). As a DSC may only contain de-identified data, we obtained permission from our university's security officer to share data in this weakly-constrained fashion. To access DACs and RDCs (which may contain identifiable data, such as audio, video, MR scans, etc.) the authentication is more constrained, requiring evidence that the user is employed by an institute for scientific research (e.g., another university).

Accessing and re-using the data of a closed DSC is not unconditional. The conditions for access and re-use of the data are specified in a so-called Data Use Agreement (DUA). The minimum condition for access and re-use is that the data may not be used to try to identify the participants that have contributed the data. On top of that minimum condition, additional conditions may be imposed, such as conditions pertaining to the commercial and scientific use, and appropriate credit for acquiring and sharing these data. The DUA is selected by the DSC manager as one of the required collection attributes without which the DSC cannot be closed. Importantly, selecting an appropriate DUA, is the only way in which a DSC manager can control how the data may be reused; there is no individual review of data access requests by the manager. When a user (after registering with appropriate credentials, such as those of a social ID) accepts the DUA of a DSC, the DRDR automatically assigns this user as a viewer to the specific collection.

THE DONDERS RESEARCH DATA REPOSITORY

The objective of the Donders Research Data Repository (DRDR) is to provide a digital data management system that allows research data to be managed according to the protocol discussed above. Based on iRODS, the repository features:

- *A file-based system* that offers flexibility for managing unstructured data stored in various format.
- *A single and uniform namespace* that is understandable and familiar to researchers.
- *Access-controlled metadata management on collection level* providing essential information for documentation and searching.
- *User authentication* via trusted identity providers.
- *Role-based authorization* on collection and OU level.

- *Data replication* for disaster recovery.
- *Workflow automation and policy enforcement* to reduce administrative and operational effort.

Figure 1 shows the architecture of the DRDR system in three layers: storage system, data-management middleware and interfaces.

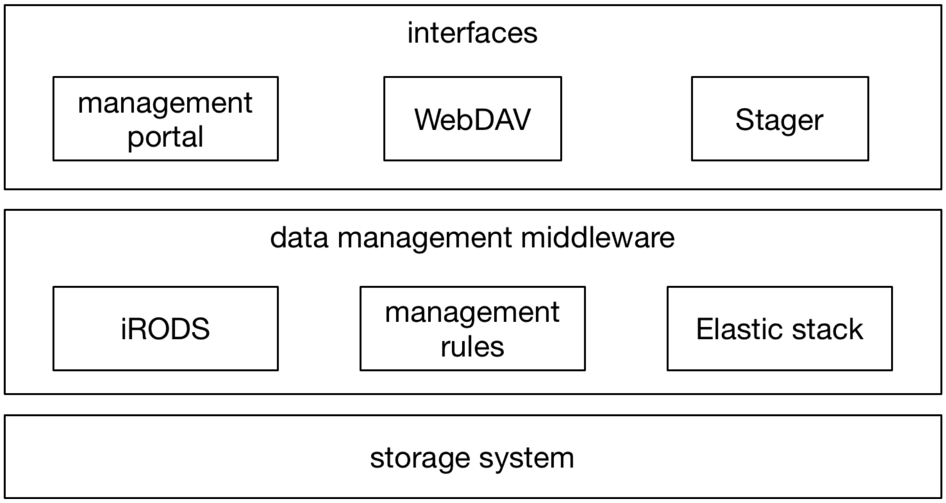

Figure 1: the DRDR architecture in three layers.

Storage system

The storage system is where the data are physically stored. It is mounted as a filesystem on the iRODS resource servers. In principle, the storage system can be of any type and specification. However, we anticipate some level of data duplication in different collections and thus rely on the de-duplication feature of the storage system to reduce the cost.

In the DRDR, two identical storage systems are installed at different locations for data safety and disaster recovery. For better control of dataflow and less dependency on storage functionality, data replication between the two storage systems is managed by the data management middleware.

Data management middleware

Leveraging on iRODS, the data management middleware implements the core functionality of the DRDR. We will discuss below the way iRODS is used in various aspects.

Resource arrangement

Using the tree metaphor and composable resources, the iRODS resources are arranged accordingly to maintain two data replicas in the system, and to reflect the administrative boundaries of different OU's. The arrangement is illustrated in Figure 2.

For the incoming data (i.e. the first replica), multiple coordinating (random) resources are created, each for an OU. Within each coordinating resource, storage (UNIX filesystem) resources are mapped to the filesystem mounting the first storage system. Quota may be configured on the filesystem to restrict the overall storage usage of an OU. With a proper design of the iRODS namespace (see later), the incoming data is controlled to flow accordingly to its OU-specific coordinating resource. Although we are using a single file server for now, this configuration allows OUs to grow with different storage (financial or technical) decisions independently without interference to other OU's and users.

As an “internal backup”, resource for the second replica is made as a single coordinating resource consisting of a storage resource pointing to the second storage system. Data replication is performed by a delay rule triggered upon creation (or update) of the first replica.

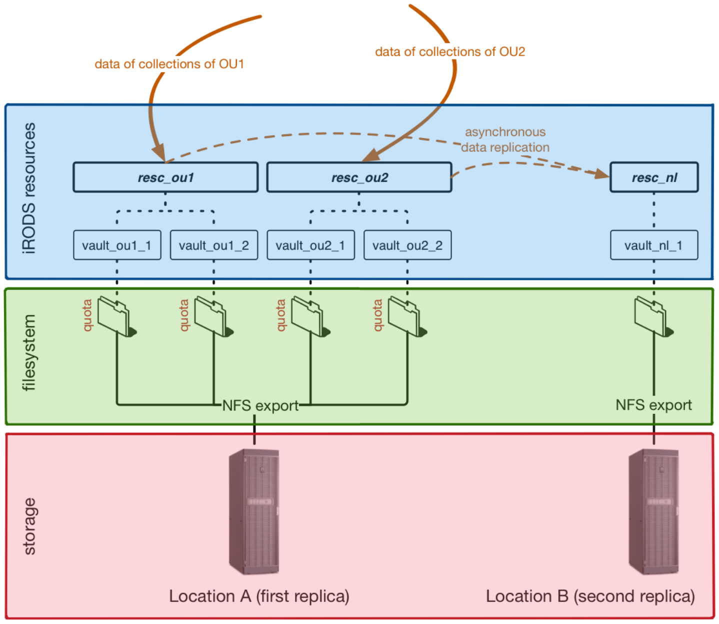

Figure 2: organisation of iRODS resources and dataflow in the DRDR system.

Collection namespace and role-based authorisation

The iRODS namespace of the DRDR collection is structured to reflect the hierarchy of organisation (O), organisational unit (OU) and the DRDR collections. Within a DRDR collection, the researcher has the freedom to organise data according to the specific (research) needs. Figure 3 shows an example namespace of a DAC belonging to the organisational unit DCCN.

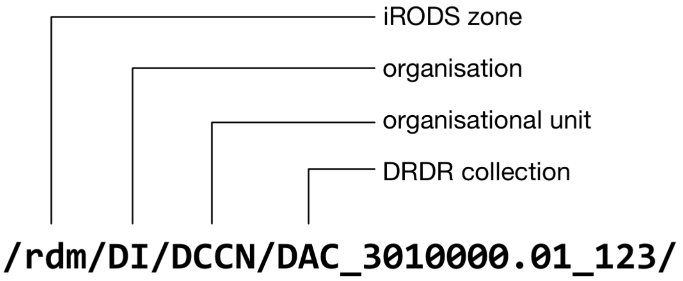

Figure 3: an example of the DRDR collection namespace.

The namespace hierarchy helps to implement the role-based authorisation defined by the protocol using the group-level access control. For instance, OU groups (e.g. ‘ou_admin’) are created to govern the access permission of the OU-level namespace; while three groups are made for each collection to control the permissions of ‘manager’, ‘contributor’ and ‘viewer’ roles. In this approach, granting/revoking access permission is simply achieved by adding/removing the user to/from a corresponding group.

Collection metadata

In the DRDR, metadata is only assigned to collections. On the collection-level namespace, metadata attributes are stored as the key-value-unit (KVU) triplets of iRODS. The attributes are largely derived from DataCite [3] with few control vocabularies specific to neuroscience and medical science (e.g. MeSH [4]). Attribute values containing data structure are represented as a JSON string.

Policy-enforcement points are adjusted to trigger data-management workflow when certain collection attribute is changed. For example, setting the 'state' attribute to 'closed' triggers a chain of actions to 1) set the collection to 'read-only', 2) clone the collection to a versioned snapshot, and 3) register the versioned snapshot with a global identifier (e.g. ePIC [5] or DOI) if the collection is a DSC.

Role-based authorisation is applied to metadata attributes according to the protocol. This is achieved by using the management rules (see later) given that iRODS doesn't support fine-grained authorisation on the KVU triplet.

Management rules for clients

A powerful feature of iRODS is the rule engine, which allows the data management policy and workflow to be customised. In the DRDR, we also use rules to create RPC-like client-server communications for high-level management such as editing collection attributes. The benefit of this approach is that complex logic and workflow managed on the iRODS server side are transparent to the web-application client. The management rules can be easily reused to build different management clients.

Figure 4 illustrates a client-server interaction using a management rule (i.e. rdmUpdateCollectionMetadata) for editing collection attributes. The client simply calls the rule with inputs; while on the server-side complex logics are wrapped together with the actual attribute update for, for instance, authorisation enforcement and verifying whether the values to be set are valid.

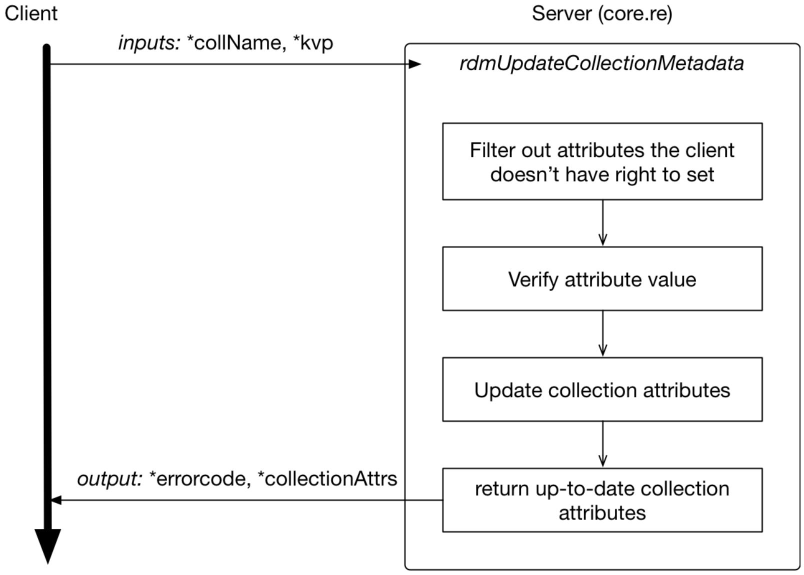
Client
Server (core.re)
inputs: *collName, *kvp
rdmUpdateCollectionMetadata
Filter out attributes the client doesn't have right to set
Verify attribute value
Update collection attributes
return up-to-date collection attributes
output: *errorcode, *collectionAttrs

Figure 4: schematic illustration of the client-server interaction for updating collection attributes.

User provisioning and authentication

Users in the DRDR are provisioned as iRODS accounts when they sign-up the first time to the DRDR management portal via a trusted identity provider (IdP). Using the national IdP federation (i.e. SURFConext [6]), identities issued by Dutch research and educational institutions are supported. Social IDs are also supported with limited permission in DRDR. User attributes retrieved from IdP are stored in iRODS as user profile which is used to communicate with user (via e-mail attribute), and determine the user's eligibility in the system.

For security reason and to allowing auditing, users sign in to the management portal via a trusted IdP. This allows the system to record actions under an account that is traceable to an actual person. To access the data stored in a collection, the user has to retrieve the iRODS username and a short-term password from the management portal (after authentication). The iRODS username is used to interact with the WebDAV and iRODS data access interfaces. The short-term password prevents data from being accessible to a person whose authorisation has been revoked or expired by the trusted IdP.

Logging

User interactions with iRODS are logged as "events" written to the iRODS log file (i.e. rodsLog). The event content contains necessary details for auditing. Events in the log file are processed by Filebeat [6] and transferred into the Elastic stack [7]. Additional tools for accounting, notification and reporting are built on top of the Elastic stack. This dataflow is illustrated in Figure 5.

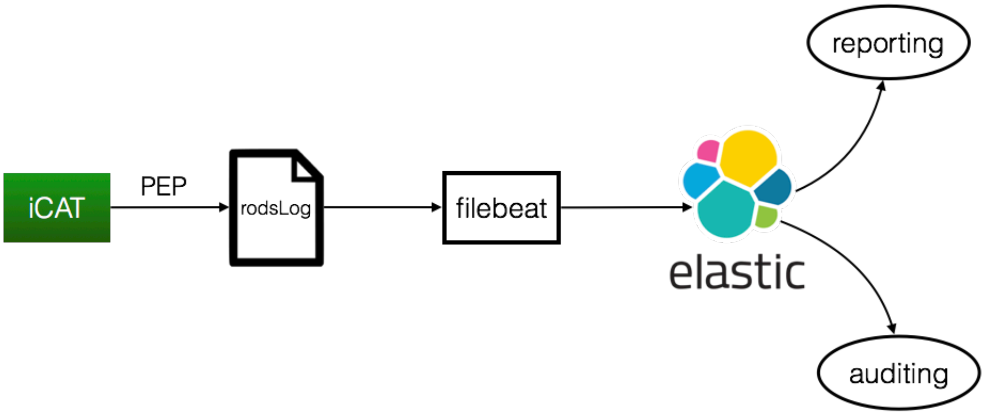

Figure 5: the dataflow of user events from iRODS to the Elastic stack.

Interfaces

One important feature in the DRDR interface is that it separates data access from the collection and user management. While the high-level rules consolidate the way management interfaces interact with iRODS, various data access mechanisms and interfaces are required for different scenario.

Management portal

The main management interface of the DRDR is a user-friendly web portal [1]. It is mainly used for managing user and collections attributes. Integrating with the federated identity provider, it is also used for user authentication and retrieving temporary credential for data access. Most of the functionalities in the manager portal require user authentication. The management portal also hosts the landing pages for published Data Sharing Collections which are publicly accessible. Under the hood, it makes use of the high-level rules to perform actions in iRODS.

WebDAV for easy data access

The main interface for transferring data in/out the DRDR is the WebDAV gateway using the Davrods [9] implementation. It allows researchers to use familiar tools to access data, and offers seamless integration with desktop by "mounting" collections as desktop drives.

The downside of WebDAV is that it is not trivial to maintain a reliable connection for transferring massive amount of data. Any failure along the connection will result in interruptions which then require human intervention. For researchers, these interruptions result in a considerable amount of effort, just to manage the transfers.

Stager for massive data ingestion and retrieval

The stager is a service implemented for data transfers between the DRDR and a local (high-performance) storage system, eliminating the need of researchers' effort in dealing with massive data transfers.

The stager provides a web interface for users to specify data transfer tasks in a graphical way similar to the file manager [10], and submit them to a task queue. Internally, an agent process performs the task of transferring data using the efficient "irsync" command and monitors the transfer progress. In case of failure, the task is rescheduled (up to a certain number of attempts). At completion, the user is notified by email about the result of the transfer.

The stager provides an efficient way for massive data ingestion and retrieval. Using the stager we implemented the automatic streaming of data from acquisition equipment (specifically the MRI and MEG scanners) to the DRDR. Upon the completion of a data-acquisition session in the lab, a data-transfer task is posted to the stager for ingesting the raw data to the DRDR. This automatic raw-data streaming into the DRDR provides not only a convenient automatic data stream for the researchers and an improvement in scientific integrity, but also an opportunity to organise the raw data into a standard structure (e.g. the BIDS standard [11]).

STRENGTHS AND WEAKNESSES

We outline two strengths and four weaknesses. The first important strength is that we have realised a protocol and associated IT environment that fits with the combined scientific-administrative workflow of a large and heterogeneous institute. Importantly, by treating storage on the DRDR as a resource for which researchers must apply just like other research facilities, we obtain adoption of the initial part of our RDM protocol. A second important strength is that DRDR provides the necessary functionality for (1) the sharing of the data of published papers, and (2) the implementation of the Data Management Plan (DMP) of a research grant. In a DMP, it must be described how the data will be handled, stored and shared in the different stages of the research project; most of these functions belong to the core functionality of the DRDR. Publishing papers and obtaining grants are important objectives for researchers, and this contributes to the user adoption.

A first weakness of DRDR is that it provides weak integration with the data analysis workflow using High Performance Computing (HPC) or using desktop computers. For instance, at this moment, we do not have a cross-platform solution allowing data analysis programs to access the DRDR collections as seamless as using a local storage. A second weakness of DRDR is that users can only be authorized for access to DACs and RDCs after they have authenticated via SURFConext as federated IdP service for research institutes and universities. Unfortunately, SURFConext operates on an opt-in basis: the DRDR has to be added by the IdP administrator for each of the research institutes or universities. This will only happen if employees of that organisation explicitly express to the local IdP administrator that they want to authenticate themselves in DRDR using their employer's IdP service. Only 9 of the approximately 100 institutions using SURFConext (which includes only 2 of the 14 Dutch research universities) has so far opted in. A third weakness of DRDR is that it does not implement standards for the collection content, simply because broadly applicable (cross-discipline) standards for collection content do not yet exist; the DI is an institute with a large variety of neuroscience disciplines. The absence of these standards hampers interoperability and reusability of the data. The fourth weakness is that the current DRDR interface doesn't provide enterprise-level searching-filtering-sorting functionality on collection metadata. This is due to limitations of the iRODS querying interface. For the moment, users can only filter collections on a few predefined attributes.

FUTURE DEVELOPMENT

Given the three collection types, DRDR provides a single repository that facilitates data management for both internal collaboration (via DAC and RDC) and data publication (via DSC).

In the perspective of internal collaboration, we are looking forwards developing an efficient integration between the DRDR system and computing facilities so that researchers can access collection content seamlessly for data analysis. Linked to this is an ongoing challenge of structuring the data in standard ways so that machine actions, such as automatic data processing, can be implemented within the collection.

For the data publication, our goal is to enable open data access following the FAIR principles [12]. Assigning persistent identifiers to published DSC's is the first accomplished step. In the future, we will address other aspects of the FAIR principle by, for example, exporting attributes of the published DSC's to FAIR data points with standard schemas (e.g. Dublin Core [13]). Given the sensitivity of the specific neuroimaging and medical data that we are managing, we note that not all FAIR aspects may be fully realised.

Furthermore, we will have to improve the system for better adoption and user friendliness. Speeding up data transfers, integration with a service interconnecting identity federations around the world (via eduGAIN [14]) and improving advanced searching-filtering-sorting functionality are important issues.

Although we believe that the method we have developed for managing neuroimaging data can be adopted by other research domains, we anticipate that some adjustments are needed to fit smoothly in a different domain. It is interesting to know how the current implementation will land in other research institutes. Feedbacks from the adoption process will help improve the system towards a more generic RDM solution.

CONCLUSION

We have structured the data management workflow in which both researcher and administration take part of responsibility. The workflow is specified by protocol and implemented by a digital data repository, and it pertains to the three data-management processes in the research lifecycle, namely the generation, documentation and publication of data.

The workflow defined in the protocol is content independent. It allows the protocol to be easily adopted by different research domains; while still offering flexibility for researchers to manage content towards their need.

The data repository is based on iRODS. The flexibility of iRODS allows us to leverage various iRODS features to implement the functional requirements derived from the protocol, and to provide feasibility for possible changes on the protocol.

Both the protocol and data repository are designed to provide a generic way of the research data lifecycle management. Thus, from our perspective, they can be either directly adopted by similar research institutions, or provided as a reference implementation for other research domain.

ACKNOWLEDGMENTS

We would like to thank the Management Board and the RDM steering group of the Radboud University for their guidance and financial support, the Radboud ICT Service Centre (ISC) for the technical support in developing the CMS portal of the repository, and the researchers at the Donders Institute for their invaluable input on design and usability aspects.

REFERENCES

[1] Donders Research Data Repository, **https://data.donders.ru.nl**

[2] The Frequently Asked Questions page of the Donders Research Data Repository, **http://www.ru.nl/donders/research/data/faq/**

[3] DataCite, **https://www.datacite.org**

[4] NCBI Medical Subject Headings (MeSH), **https://www.nlm.nih.gov/mesh/meshhome.html**

[5] ePIC, **http://www.pidconsortium.eu**

[6] Filebeat, **https://www.elastic.co/products/beats/filebeat**

[7] The Elastic stack, **https://www.elastic.co/products**

[8] SURFConext, **https://www.surf.nl/diensten-en-producten/surfconext/index.html**

[9] Smeele, T., Smeele C., Davrods, an Apache WebDAV interface to iRODS, Proceeding of iRODS User Group Meeting 2016, 41--48 (2016)

[10] Orthobox file managers, **https://en.wikipedia.org/wiki/File_manager#Orthodox_file_managers**

[11] Brain Imaging Data Strucutre (BIDS), **http://bids.neuroimaging.io**

[12] Wilkinson, M. D. et al., Comment: The FAIR Guiding Principles for scientific data management and stewardship, Scientific Data 3:160018 doi:10.1038/sdata.2016.18 (2016)

[13] The Dublin Core, **http://dublincore.org**

[14] eduGAIN, **https://www.geant.org/Services/Trust_identity_and_security/eduGAIN**