Research Data Management (RDM) at the Donders Institute (DI) and the Donders Research Data Repository (DRDR)

June 2017

Eric Maris, Robert Oostenveld, Erik van den Boogert, & Hong Lee
Research Data Management

Research Data Management (RDM) is a part of the research process that aims to improve its efficiency and transparency, and to fulfill the requirements of the university, funders, and legislation.

1. long-term data preservation for internal reuse
2. a documented analysis pipeline, allowing for collaboration, reuse, and reproduction of the published results
3. easy sharing of data and analysis pipelines with colleagues around the world
The Research Data Life-Cycle

1. **Data preservation** for efficient internal reuse
2. **Research documentation** for enhancing collaboration, reusability and reproducibility
3. **Open access data sharing** with the external scientific community
Donders Research Data Repository (DRDR)

A managed institute-level archive plus IT-services that allow for the realization of the three RDM objectives.
Three objectives → three collection types

1. **Data Acquisition Collection** (DAC) → internal
   Contains the original data

2. **Research Documentation Collection** (RDC) → internal
   Contains files that document the process from data to published results & allows for sharing during project-based collaboration.

3. **Data Sharing Collection** (DSC) → external
   Contains the data and analysis scripts on which the published results are based, allowing other researchers to extend the published findings
Challenges

• Large institute with heterogeneous scientific-administrative workflows
  o 600 researchers in PI groups, generating >250 projects per year
  o extensive and intensive international collaborations
  o 4 centers, each with its own administrative domain

• Data complexity
  o imaging, multichannel signals, text, audio/video, etc.
  o privacy-sensitive data
  o sizes ranges from a few large to a huge amount of small files
The RDM protocol

Organisational Unit (OU)

- Collection: a container of data (files/folders)
  - ✓ Metadata
  - Open or closed

- Workflow
- Responsibility
- Eligibility

- Research initiation
- Data acquisition
- Data sharing
- Data analysis & scientific writing

- Research administrator
- Manager
- Contributor
- Viewer
- PID
What data must go in the collections?

• The IT-system does not enforce how collections are to be built. Instead, the collection content is specified in a written protocol.

• This written protocol specifies the collections primarily in functional terms from the perspective of the reader, rather than in operational terms from the perspective of the collection builder.

• The written protocol is augmented by online documentation in the form of a Frequently Asked Questions (FAQ) page, which contains concrete suggestions on how the functional requirements can be realised.
What data must not go in any collection?

Personal data, which includes the informed consent forms and pseudonymization keys.

What data must not be shared?

Identifying data, such as photo, video, recognizable audio, faces in anatomical MRI.
Roles and responsibilities

• Collection-level roles
  o **viewer** (download)
  o **contributor** (download/upload/modify/delete)
  o **manager** (add users, close collection, download/upload/modify/delete)

• A research administrator creates the collections and assigns the initial manager
  o **PI is always initial manager**
  o Researchers can be assigned as contributor or co-manager (requires u-number and OU check-in), and the latter has the same rights as the initial manager.
Two types of interfaces to the DRDR

• A web interface via which
  o Users can be added/removed from collections
  o Metadata can be added, including support for two neuroscience controlled vocabularies

• Two interfaces for data transfer
  o One based on the WebDav protocol, which only requires a WebDav client
  o One based on the native middleware (iRODS) commands, which also requires a staging area
Authentication

• Via the web interface
• Three authentication levels:
  o Via a trusted institute (SurfConext, EduGAIN, …)
  o Via a social ID (Onegini)
  o Anonymous
• After web-based authentication, the user can obtain a one-time password for data transfer
Roles and eligibility

- Manager  Present in the OU administration + trusted authentication
- Contributor  Trusted authentication
- Viewer of open collections  Trusted authentication
- Viewer of closed DSC  Authentication via Social ID
- Viewer of metadata of closed DSC  Anonymous user
Closing a collection

- A **DSC** is closed when a paper is accepted for publication
  - Only a **manager** can close a collection
  - When closing a DSC, a **Data Use Agreement** (DUA) has to be selected
  - As a part of the closing process, a **persistent identifier** (PID) is generated.
  - That PID is inserted in the footnote of the publication (e.g., [http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0154881](http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0154881))
  - After closure, a DSC can no longer be modified
Closing a Collection

• A **RDC** is closed when a paper is published
  o Much of the content of a DSC can be reused in a RDC.
  o If the associated DAC is still open, it also contains the raw data
  o An RDC is linked to a paper via the paper’s PID
• A **DAC** is closed when no more data will be collected
  o When closing a DAC, storage is used efficiently
  o ... but may not be possible in large data collection projects
How FAIR is this all?

• Findable

• Accessible

• Interoperable

• Reusable