The important role of CRIS's for registration and archiving of research data. The RDS-Project at Radboud University (NL) in cooperation with DANS.

Ed Simons, Mijke Jetten, Marnix van Berchum, Maaike Messelink, Hans Schoonbrood, Marion Wittenberg.

Abstract

Optimal research data management and archiving is a key condition for progress in modern science and of vital importance from both the point of view of research as such as well as research policy and management. More specifically it is a *conditio sine qua non* for the realisation of Open Science and at the same time it is indispensable for the monitoring and assessment of the quality and integrity of research. Various aspects play a role here: optimal infrastructures and tools for the actual handling of data during the research lifecycle, appropriate metadata to describe the datasets, and - last but not least - an adequate organisational framework to curate and archive the datasets professionally and provide optimal support and services to the researchers.

The paper presents the "*Research Data Services (RDS)*" project of Radboud University (NL) in cooperation with one of the Dutch national research data archives: DANS (Data Archiving and Networked Services). In this project a model is worked out for the archiving of research datasets via the CRIS[1] of the university, including both the registration of the metadata as well as the actual upload of the data files towards DANS. It is argued that an optimal solution is not only a technical matter, but also requires the definition and organisation of appropriate support and management structures and workflows, involving both the local and national partners. In this respect attention is paid to the explanation of the "FoBo-model" (Front Office – Back Office) that is being defined and implemented as part of the project and which forms the organisational backbone of the solution worked out. The paper starts by arguing that a CRIS-oriented approach in research data archiving holds substantial added value, and ends with an overview of lessons learned and a peek into the future of the RDS project.

1. Setting the Stage: the benefits of using a CRIS for registration and archiving of datasets.

Proper registration and archiving of datasets resulting from research, allowing them to be reused by peers and accessed for monitoring and control purposes, has become an important issue within present-day science and science policy. Among other things, optimal storage, curation and sharing of research data is seen as a key stone for the realisation of Open Science. A crucial element in this respect is the availability of adequate metadata linked to and registered together with the datasets, in order to make

[1] CRIS: Current Research Information System.

them “FAIR”: *Findable, Accessible, Interoperable and Reusable*[2]. Without a sufficiently complete and detailed set of metadata that provide information about the location, nature, subject, context and conditions of use of the dataset, the latter would be untraceable and it would be difficult if not impossible for a user to interpret and determine whether the dataset could be of use for her or him. Optimal availability of metadata supposes adequate underlying models and systems for registration and management of the metadata and this is where CRIS’s come in. Being resources that traditionally hold rich and detailed metadata on all aspects of research, including research results, one should expect CRIS’s to also be the obvious systems of choice for the registration and handling of metadata on datasets. Surprisingly practice is a little different, as the registration of datasets and its accompanying metadata is often treated as a separate reality and looked upon in a silo-ed way, not integrated with the broader context of research information aspects and objects to which they are - nevertheless - indissolubly connected. This has resulted in the coming into being of new, separate, systems and applications uniquely targeting the registration and management of datasets, leaving out the link to the publications based on them, the institutes and projects within which they were created, and so on. This situation has two clear disadvantages:

1. Valuable contextual pieces of information that could act as extra, additional entry or source for finding and interpreting the datasets (or related datasets) are left out and remain invisible, thus creating serious limitations and a suboptimal situation from the point of view of the “FAIR-ness” of the data (see above).
2. Researchers, management and policy makers are confronted with “yet another administrative system” with yet another interface. This situation may not only lead to irritation with these stakeholders, but it is certainly inefficient as it hampers a comprehensive, integrated view on (all aspects of) the research they are involved with[3].

As already indicated, these drawbacks can be avoided by using a CRIS for the registration and archiving of datasets. As CRIS’s nowadays more and more function as the primary and leading source for an institution’s OA Repository, integrating the registration of datasets in the CRIS has the clear benefit for the researchers of being confronted with only one interface for handling all aspects of their research information, and one environment in which they can directly link all these aspects and objects, e.g. publications and datasets, to one another. Also for the research management and policy makers there is substantial added value involved, as they have all the necessary elements for research management information available in one resource. Last but not least, by integrating dataset information in the CRIS’s ,the latter become optimal information nodes for (international) Research e-Infrastructures and Open Science Clouds.

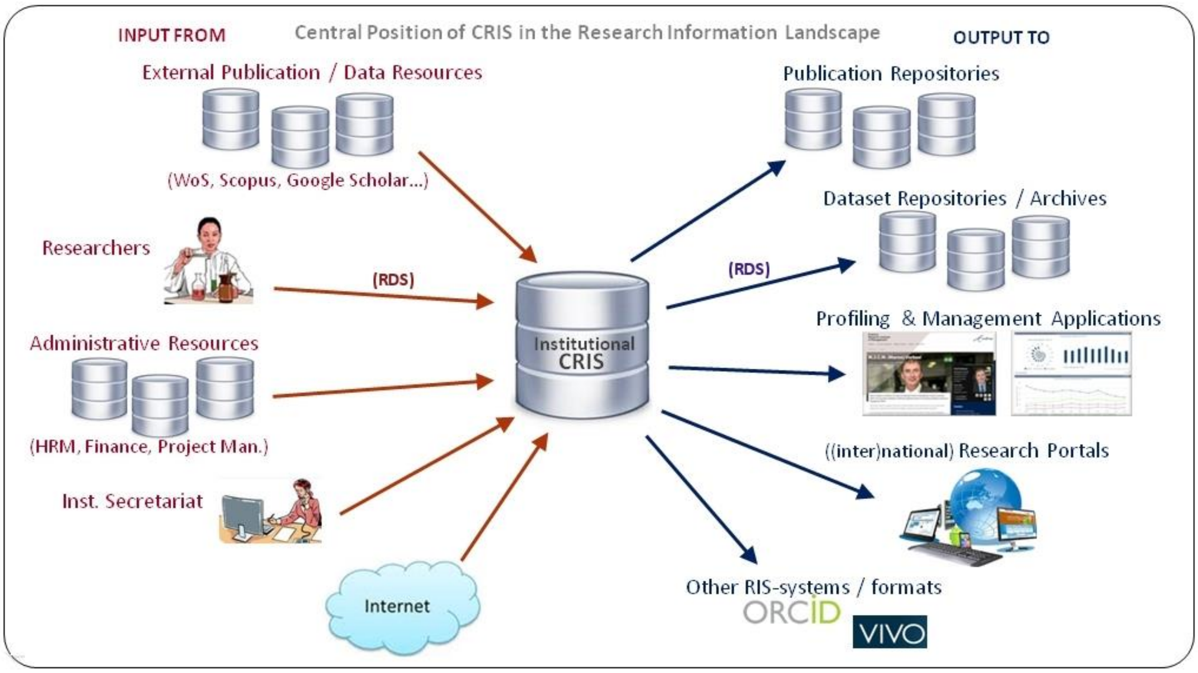
[2] Mark D. Wilkinson, Michel Dumontier, et. al., *The Fair Guiding Principles for scientific data management and stewardship*, Nature - Scientific Data, 3, 15 March 2016. http://www.nature.com/articles/sdata201618

[3] Just one example in this respect: the management of a university or even a single research institute may not have a clue of how many and where datasets resulting from the institution’s research are stored, as (the information on) the registration of datasets may be spread over various applications or systems, not linked to one another. This usually in contrast to the information on publications.

2. Operationalising the concept: the RDS-Project at Radboud University (NL)

Some two years ago, the management of Radboud University, as part of its Research Data Management Policy for the institution, decided to use the institutional CRIS (Metis) as the basic instrument and resource for the registration and archiving of datasets resulting from the institution's research..
To implement this policy, a project called *Research Data Services (RDS)* was defined to extend the existing CRIS functionality with the possibility for a researcher to upload her datasets to a data hosting provider and register the accompanying metadata in the CRIS itself.
This is in line with the broader view and policy within the institution that the CRIS should be the *one stop resource* for both input and export of the institution's research information.

This view on the central position of the CRIS is summarised in the image below, with the aspect dealt with in the RDS-project separately indicated:


INPUT FROM
Central Position of CRIS in the Research Information Landscape
OUTPUT TO
External Publication / Data Resources
(WoS, Scopus, Google Scholar...)
Researchers
(RDS)
Administrative Resources
(HRM, Finance, Project Man.)
Inst. Secretariat
Internet
Institutional CRIS
Publication Repositories
Dataset Repositories / Archives
(RDS)
Profiling & Management Applications
((inter)national) Research Portals
Other RIS-systems / formats
ORCID
VIVO

As Radboud University itself does not primarily offer data storage services for archiving datasets, the cooperation of a data hosting organisation - to accept upload of datasets and metadata from the CRIS - was necessary in order for the CRIS-oriented solution envisaged in the RDS-project to work. Luckily one of the Dutch national research data archives: DANS (Data Archiving and Networked Services) was found willing to enter into a project to work out the solution, being already CRIS-oriented themselves as they also host the national CRIS called NARCIS.

More concretely the RDS-project consists of the following parts:

1. *An online interface* for the researcher containing functions for uploading the dataset files and registering the metadata.
2. *The technical solution "in the background"* handling the actual upload of files and metadata to the data hosting organisation and the automated communication workflows involved (e.g. the necessary e-mail alerts to involved parties/persons).

3. *A (user) support and services organisation, the so called "Front Office - Back Office Model"*, distinguishing between services at the institution's level (Radboud University - Front Office) on the one hand and the national data hosting provider (DANS - Back Office) on the other.

In the following paragraphs these three aspects are looked upon in more detail

3 . The (CRIS) Interface

Before the start of the RDS-project, the CRIS (Metis) of Radboud University had an interface which allowed a researcher to register his publications. The first, and most visible, goal of the RDS-project was to extend this interface with functions for registering and uploading datasets.
Basically this involved the following activities:

- *The choice of a standard (generic) metadata set* for describing the various aspects of the dataset (identifier, title, creators, global keywords, collection period, geographical location, etc…). For this the international *DataCite* metadata set was chosen as point of departure.
- The actual implementation of the metadata mentioned above in *an online registration form,* i.e. fields on the screen to be filled in by the researcher.
- *Functions allowing the researcher to select and actually upload the dataset files themselves*.

An important function that was built into the interface was the possibility to directly link the datasets to publications. This can be done, starting from either the already existing publication part of the interface or the dataset part newly developed in the RDS-project.

Concerning the metadata it should be noted that, apart from the generic descriptive metadata mentioned above, to a dataset could also pertain subject or discipline specific metadata, generally consisting of extensive, specific thesauri / or keyword lists. Given the specific and extensive nature of these metadata, it is almost impossible to implement them as form fields in an interface. Therefore, in the RDS-project they are included as a separate file together with the actual data files, so as part of the total dataset package uploaded to the data hosting organisation.

The interface was developed stepwise, in an iterative approach and in close cooperation with five pilot groups consisting of researchers from various disciplines: Biology, Communication Sciences, Informatics, Language Studies and Management Sciences.

4. Making things work: the technical solution

The source: the CRIS ("Metis") at Radboud Univeristy

This year is the 23rd anniversary of the CRIS called "Metis" developed by Radboud University. At present, the Metis database holds approximately 200,000 metadata records with data on research results; publications, annotations, dissertations, lectures, patents and recently, as a result of the RDS-project, datasets.

There are different modules for different users in Metis. For the researcher there is a separate (CRIS) interface, giving a researcher direct access to the registration and management of her personal research results. She can add or modify results, add unique identifiers, define personal profiles, upload full text to the Radboud repository and since 2015 archive datasets at the DANS national data repository. In addition, it is also possible to define relationships between datasets or between datasets and other results, for example refereed articles.

The CRIS interface is a Java-based application. The underlying relational database management system (RDBMS) is Oracle. In the development environment the following tools are being used:

- *Bootstrap*; Bootstrap is a widely used and popular HTML, CSS, and JS framework for developing responsive web applications and mobile first projects.
- *iBATIS; iBATIS* is a persistence framework which automates the mapping between SQL databases and objects in Java, .NET, and Ruby on Rails.
- *jQuery; jQuery* is a cross-platform JavaScript library designed to simplify the client-side scripting of HTML.

The database consists of more than 300 tables and a large number of views, procedures, functions and database packages. The content is (partly) freely available through webservices. In addition, there are interfaces with local systems, including the personnel system, the repository, personal profile portals, OCLC, Scopus, Web of Science and - for the datasets - with the data archiving system at DANS.

At the national hosting organisation: DANS - EASY

The datasets in the RDS-project are uploaded from the local CRIS to DANS through the so-called SWORD protocol and stored in a system called “EASY” (Electronic Archiving SYstem), the certified long term preservation archive of DANS.[4] EASY offers functionality for manually uploading datasets, as well as opportunities for bulk import of data. All data in EASY is in principle being curated by the DANS Data Managers, although for the SWORD ingested datasets, as in the RDS-project, other agreements apply, involving the Front Office of the archiving institution (see below). EASY is built upon the Fedora platform, and maintained and further developed by DANS’ own technical staff. Currently EASY contains a rough 31 thousand datasets, mainly originating from the Humanities and Social Sciences.

The SWORD Protocol

The Simple Web-service Offering Repository Deposit (SWORD) protocol is a lightweight protocol for depositing content from one location to another. SWORD is a profile of the Atom Publishing Protocol (known as APP or AtomPub).[5] Since SWORD is using an existing protocol, AtomPub clients and API’s can be re-used. The SWORD protocol is used to communicate between two clients: one depositing the content, one receiving the content. The protocol makes uses of existing HTTP methods; the basic functionality is an HTTP POST request for depositing data. This request contains some (optional) SWORD headers and a body, which contains a zip file. The server receiving the request, responds with an HTTP code (e.g. 200 OK). All services of the protocol are described in the so-called Service Document. First step of the communication between two clients is the exposing of this service document. With the description of the service available the preparation of a deposit is made.

[4] http://www.easy.knaw.nl; EASY is Data Seal of Approval (DSA), World Data System (WDS) and NESTOR-DIN certified.

[5] See http://swordapp.org/about/.

DANS has experimented in recent years with the SWORD protocol, e.g. in the *Enhanced Journals Made Easy!*-project.[6] From 2014 onwards work focussed around the connection of institutional data repositories to the long term preservation archive EASY. Currently, DANS has version 1.3 of the SWORD specification implemented in EASY, and has an experimental set up with version 2.
In the next paragraph the specific workflow applied in the RDS-project (Metis-CRIS to DANS) is described.

Workflow description

A researcher registers the metadata of the dataset in the CRIS and uploads the dataset files through the CRIS interface. The files are not directly uploaded to DANS, but in a first stage stored on a local server at Radboud. The Front Office at Radboud controls the datasets and, if necessary, communicates with the researcher for corrections/additions and if everything is found in order, initiates the upload to DANS through the upload function in the CRIS.

The CRIS at Radboud then makes an HTTP POSt request to EASY, with the required header(s) and body. The zip file of the body contains the following:

1. a file containing metadata with the name: DansDatasetMetadata.xml
2. a folder named "data" containing the data files.

The metadatafile DansDatasetMetadata.xml is in the DDM (DANS Dataset Metadata) format, the preferred metadata format as defined by DANS.[7] The folder containing the data files may included subfolders, which will be included in EASY. For preservation purposes DANS makes use of so-called *Preferred Formats*.[8]

When the dataset is accepted a 202 HTTP response is sent. A confirmation e-mail is sent to the depositor, including the basic information of the dataset and the persistent identifier (PID). Every dataset submitted in EASY receives a PID, which uniquely identifies a dataset.[9] This PID (in practice: the DOI) is automatically transmitted to and imported, in the CRIS at Radboud University, integrated with the other already existing metadata of the dataset.

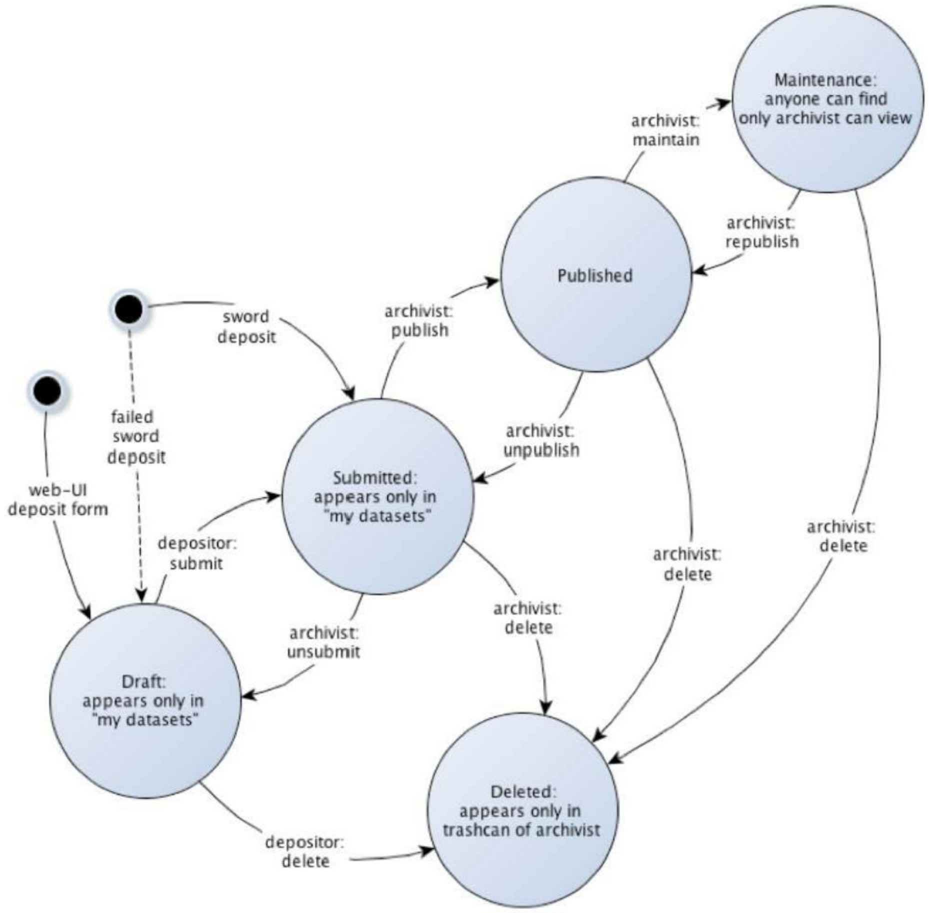
If the received dataset passes validation, the dataset will be saved in EASY and it's state will be set to SUBMITTED. From this point on, the dataset travels the same path as "regular" datasets deposited through the EASY web interface. Since data curation is done by the Front Office (see further), all datasets deposited from the Radboud CRIS will be published, without further checks by the DANS Datamanagers.

[6] This project was part of the SURF funded program on Enhanced Publications (https://www.surf.nl/en/themes/research/research-data-management/enhanced-publications/index.html); it's goal was to develop tools for depositing data, embedded within the workflow of journal publication, in this case with the Open Journals System (OJS).

[7] The full specifications are available at https://easy.dans.knaw.nl/schemas/md/2012/11/ddm.xsd

[8] See http://www.dans.knaw.nl/en/deposit/information-about-depositing-data?set_language=en.

[9] See http://www.dans.knaw.nl/en/deposit/information-about-depositing-data/persistent-identifiers/persistent-identifiers?set_language=en.



5. Support services and organisation as the key to success: the “FoBo model”

In accordance with Radboud University’s Research Data Management (RDM) policy, researchers must store and manage their research data, and make them accessible to others, ultimately at the moment of publication of the corresponding article, book or dissertation. Since general policies often overlook the practical questions on storing, sharing and documenting data that researchers may have, a vital organisation with a robust IT infrastructure and adequate support service is essential. Hence, Radboud University Library has established a Front Office - Back Office model (FOBO model), in close cooperation with the university’s policy department, IT services and DANS. The library serves as the front office, while national research data archive DANS functions as the back office.

The Front Office - Back Office model is the meant to be the organisational component of a federated data infrastructure under development in The Netherlands. The so-called front office (in this case Radboud university library) deals directly with researchers and research supporting staff; it supports, advises and trains the researchers and students in responsible data management and ensure that research data is accommodated in one of the sustainable national research data archives (eg. in this case: DANS). The Dutch national research data archives fulfill the back-office function in the process of data curation: they ensure that the research data delivered to them is permanently and sustainable archived and made optimally available for discovery and reuse.

In the following section we describe the support services at Radboud University that constitute the Front Office. Three aspects can be distinghuished in this respect:

1. The role of the university library in research data management (RMD);
2. the support of supplemental data policies by research institutes in combination with training sessions to researchers in the deployment of the CRIS interface; and
3. the support of the daily practices of data archiving via the CRIS interface at Radboud University.

The role of the library in Research Data Management: developing Research Information Services

Characteristic of Radboud University's RDM policy is the steady, continuing role of the library. The library carries out the support, training and curation task (service desk). Offering services on publication management to researchers has been a shared undertaking by the library and the research institutes of Radboud for several years already. New to the library is the establishment of a full one-stop-service for researchers in archiving and registering research data, registering publications and uploading full text, and for registering the relationships between these products of research. Even more innovative is the data curation role that, as part of the developed *FoBo-model* shifted from the national research data archive DANS to the institution's library. These integrated services of the library are summarised under the label *RIS: Research Information Services.* There are various benefits of this particular role of Radboud University Library.

- The library acts a linking pin. It is involved in most research data management projects at Radboud University and thus guarantees that the involved partners profit optimally from shared knowledge and expertise building at Radboud University.
- The RIS Services are part of the development of the library of the future: nowadays, libraries are so much more than buildings with books. A university will profit from a strong and broad library, which is accessible to a broad variety of both students and researchers. And the RDS project specifically will profit from the historically developed comprehensive network and integrated function of the library in the various research institutes.
- Most research data management (RDM) projects are temporal, delivering specific infrastructures. However, the role of the library is stable, irrespective of the duration of any RDM project. At Radboud University Library, RDM support has become part of the library's daily business. The library thus emerges as a steady factor in the RDM world of temporarily projects and pilots.
- Where policy departments focus on the development and implementation of policies, and IT departments emphasize the technical aspects of system development and management, the library offers a fresh perspective: what are the researcher's questions and needs, and in what way can they be supported to do proper data management?

Data policy services and trainings

Central in the RDS-project stands the development of the CRIS interface (see above) and its deployment and implementation within (the institutes of) the unversity. Two aspects are fundamental for a successful deployment of the CRIS interface in the institutes: the existence of a data management policy within the institute on the one hand and CRIS training sessions on the other.
The first aspect, data management policies, regards supplemental policies within the institutes in addition to Radboud University's general RDM policy. Research institutes are encouraged to develop their own data protocols that cover the more practical aspects of data archiving (What to archive? Who archives? What documentation? etc). To aid research directors, the university library and the policy department together developed a checklist and offers research institutes support in developing and formulating these policies.

Secondly, in each research institute training sessions to groups of researchers are organised to implement disciplinary data protocols and deploy the CRIS interface. Practice learned us that it will take some time for policies to being developed and implemented, while training and demo sessions are easy to organise and make data and publications management (via the CRIS interface, in the case of Radboud University) practical and approachable. At Radboud University, training sessions turned out to be a very efficient way to introduce the CRIS interface. Vital to registering and uploading datasets and publications via the university's CRIS are best practices, easy-to-approach support and workshops, to get researchers acquainted with the existing infrastructure. Instead of the compliance-approach (to funders, journal or university's requirements) at Radboud University we opted for the benefit-for-researchers-approach (why is data and publication management useful for you).

Front Office - Back Office Model in the daily process of data archiving

The collaboration between the Radboud University Library as Front Office and data archive DANS as Back Office is most visible in the daily process of archiving research data.. Before the development of the CRIS interface, a researcher had to use different interfaces to upload his publications (repository interface) and datasets (DANS Easy interface) and register the accompanying metadata (CRIS and/or repository). Regarding datasets communication with the researcher as well as the curation of the dataset was done by DANS, while for publications the communication was taking care of by either the CRIS or the repository managers. With the introduction of the new CRIS interface and the Front Office - Back Office model, the curation task and the communication with the researcher are being handled in an integrated way by Radboud University Library. That way, the library acts as an intermediary between the researcher and the DANS archive.
A big advantage of having the library as a front office is the one-stop-shop idea: all information – website, support services, training sessions – a researcher needs can be found in one place. The following section describes the daily practices of data archiving via the CRIS interface at Radboud University.

The Front Office in practice

Datasets that are deposited via the CRIS interface are checked by Radboud University Library before they are sent to DANS. First, the library checks the information in the metadata fields, since adequate and rich metadata are of vital importance for the FAIR-aspect of datasets (see above). Therefore, the metadata must be understandable for potential re-users, including a meaningful title and solid description. Furthermore, the data files have to be checked, particularly on privacy sensitive information. The files that are sent to DANS can't contain this kind of information such as names, addresses, phone numbers, birth dates and other revealing information about the research subjects. This information has to be removed or adjusted. Moreover, combinations of variables can reveal identities of subjects, for example a data file containing information on deaf people, including their municipality. In case of a small municipality, it is better to aggregate to a higher level, like province or even country.
The data files must be readable and understandable as well. Therefore the library checks whether for instance the variable names and values are explained in a codebook, and -
if applicable - additional syntax files, original questionnaires and measuring instruments are included. If something is missing or needs to be adjusted in the dataset or metadata, the library contacts the researcher. Once the files and metadata are correct, the library sends the files to DANS using the SWORD-protocol.

DANS as Back Office

Via the SWORD-protocol, data and metadata are automatically deposited from the CRIS system at Radboud University into the online archiving system EASY of DANS. Every dataset is automatically assigned a DOI upon deposition to enable sustainable reference of the data. Data curation in this model is in principle done by a Radboud University Library data librarian. The DANS data manager assist the data librarian if necessary, checks the datasets randomly and publishes the datasets. DANS ensures long-term storage and sustainable accessibility of the research data.

6. Conclusions: Lessons learned and future plans

- The researcher pilot groups involved in the RDS-project all, without exception, reacted very positive to the concept of integrating research data management and archiving fuctions into the CRIS. This has strengthen the polciy makers at Radboud University to explicitly put the CRIS as the primary resource for research information and the CRIS interface as the one-stop-shop for researchers to manage their research information (including e.g. the information needed by the OA Repository of the institution).
- The concept of a one-stop-shop interface is taken a step further within a recently started project at the Radboud Faculty of Medicine and the University Medical Centre aimed at developing a *DRE: Digital Research Environment* for the researchers. Whereas initially the DRE was meant to provide IT tooling to support the research activities as such (analysis, data capture and life cycle management, etc…), under the influence of the RDS-project it was decided to integrate also research information functionality (CRIS-functionality) into the future DRE. This is quite an innovative development.

- From January 2016 on the FoBo-workflow is operational. After the first deposits a few flaws were encountered, both on the side of the CRIS at Radboud as well as on the side of the EASY system. Some metadata fields were not available in the CRIS system and some metadata elements were not processed properly by EASY. On both sides updates of the systems were planned, to solve these issues. Another issue was the fact that in the CRIS-system a researcher can vary the access rights per file, whereas in the EASY system this is only possible per dataset. Unfortunately this discrepancy can't be solved in the software of EASY on the short-term. For the time being, differences in access rights on file level will be adjusted manually by the DANS data manager upon publishing the dataset.
- The intention is to, in a later stage, introduce CERIF as the (XML) format to send the metadata from the CRIS to the DANS Easy Archive. In this respect, another project, aimed at supplying project, person and organisation metadata from institutional CRIS's in The Netherlands to the national NARCIS research information database (also hosted by DANS) by means of CERIF, has just started.
- Last but certainly not least and perhaps the most important lesson from a policy point of view: the availability of optimal metadata and metadata registration and management is an aspect that should not be neglected or treated in a stepmotherly way in the discussions on the realisation of research and open science infrastructures and cloud solutions, at the risk of these infrastructures becoming networks or clouds of *closed vaults* of data and as such suboptimal or even useless to researchers. This is not just an abstract or theoretical proposition, but something that was concretely encountered in the practice of the RDS project as it was explicitly mentioned by a member of one of the pilot groups that the current solution for their data archiving was indeed in practice a closed vault because of the lack of adequate metadata. *This is an important conclusion and hopefully one that will be taken to heart and not overlooked by the policy makers and project planners of research e-infrastructures or Open Cloud solutions.*