Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 106 (2017) 321 – 328

Procedia Computer Science

13th International Conference on Current Research Information Systems, CRIS2016, 9-11 June 2016, Scotland, UK

The important role of CRIS's for registering and archiving research data. The RDS-project at Radboud University (the Netherlands) in cooperation with data-archive DANS

Ed Simons[a], Mijke Jetten[b], Maaike Messelink[c], Marnix van Berchum[d], Hans Schoonbrood[e], Marion Wittenberg[f]

[a b c e] *Radboud University, Comeniuslaan 4, 6525 HP Nijmegen, the Netherlands*
[d f] *Data Archiving and Networked Services (DANS), Anna van Saksenlaan 51, 2593 HW Den Haag, the Netherlands*

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 realization 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 data sets, and – last but not least – an adequate organizational framework to curate and archive the data sets professionally and provide optimal support and services to the researchers.

The paper presents the *Research Data Services (RDS)* project of Radboud University (the Netherlands) 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 data sets via the CRIS (Current Research Information System) of the university, including both the registration of the metadata as well as the actual upload of the data files to the DANS archive. It is argued that an optimal solution is not only a technical matter, but also requires the definition and organization of appropriate support, management structures and workflows, involving both local and national partners. In this respect, attention is paid to the explanation of the *frontoffice-backoffice model (FoBo)* that is being defined and implemented as part of the project and which forms the organizational 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 it ends with an overview of lessons learned and a peek into the future of the RDS-project.

Peer-review under responsibility of the Organizing Committee of CRIS2016

Keywords: Current Research Information System (CRIS); Radboud University; Data Archiving and Networked Services (DANS); research data management (RDM); data archiving; data registration; data curation; frontoffice-backoffice model; research information services (RIS)

Peer-review under responsibility of the Organizing Committee of CRIS2016
doi:10.1016/j.procs.2017.03.031

1. Setting the stage; the benefits of using a CRIS for registering and archiving data sets

Within present-day science and science policy, proper registration and archiving of data sets resulting from research has become an important issue. This allows data sets to be reused by peers and accessed for monitoring and control purposes. Among other things, optimal storage, curation and sharing of research data are considered keystones for the realization of Open Science. A crucial element is the availability of adequate metadata linked to and registered together with the data sets, in order to make them *FAIR* (findable, accessible, interoperable and reusable)[1]. Without a sufficiently complete and detailed set of metadata that provides information about the location, nature, subject, context and conditions of the use of the data set, the latter would be untraceable and it would be difficult if not impossible for a user to interpret and determine whether the data set may be of use for him or her.

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 be the obvious systems of choice for the registration and handling of metadata on data sets. Surprisingly, practice is a little different, as the registration of data sets and its accompanying metadata is often treated as a separate reality and not integrated with the broader context of research information aspects and objects to which they are - nevertheless - indissolubly connected. This has resulted into the coming of new and separated systems and applications uniquely targeting the registration and management of data sets, 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 at least two clear disadvantages:

- Valuable contextual information that could act as extra, additional entry or source for finding and interpreting the data sets (or related data sets) is left out and remains invisible, thus creating serious limitations and a suboptimal situation from the point of view of the *FAIR*-ness of the data (see above).
- Researchers, management and policy makers are confronted with 'yet another administrative system'. This situation may lead to irritation and it is inefficient as it hampers a comprehensive, integrated view on (all aspects of) the research they are involved with.[1]

As indicated, these drawbacks can be avoided by using a CRIS for registering and archiving data sets. As CRIS's nowadays more and more function as the primary and leading source for an institution's Open Access repository, integrating the registration of data sets 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, including directly linking publications and data sets to one another. Additionally, for 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 data set information in the CRIS's, the latter become optimal information nodes for (international) Research e-Infrastructures and Open Science Clouds.

2. Operationalizing the concept: the RDS-project at Radboud University (the Netherlands)

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 registering and archiving data sets resulting from the institution's research. To implement this policy, a project called *Research Data Services (RDS)* was defined to expand the existing CRIS functionality with the possibility for a researcher to upload her of his data sets to a data-hosting provider and register the accompanying metadata in the CRIS itself. This is in line with the broader view and policy in the institution that the CRIS should be the *one-stop-resource* for both input and export of the institution's research information.

---

[1] 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 data sets resulting from the institution's research are stored, as (the information on) the registration of data sets may be spread over various applications or systems, not linked to one another. This is usually in contrast to the information on publications.

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

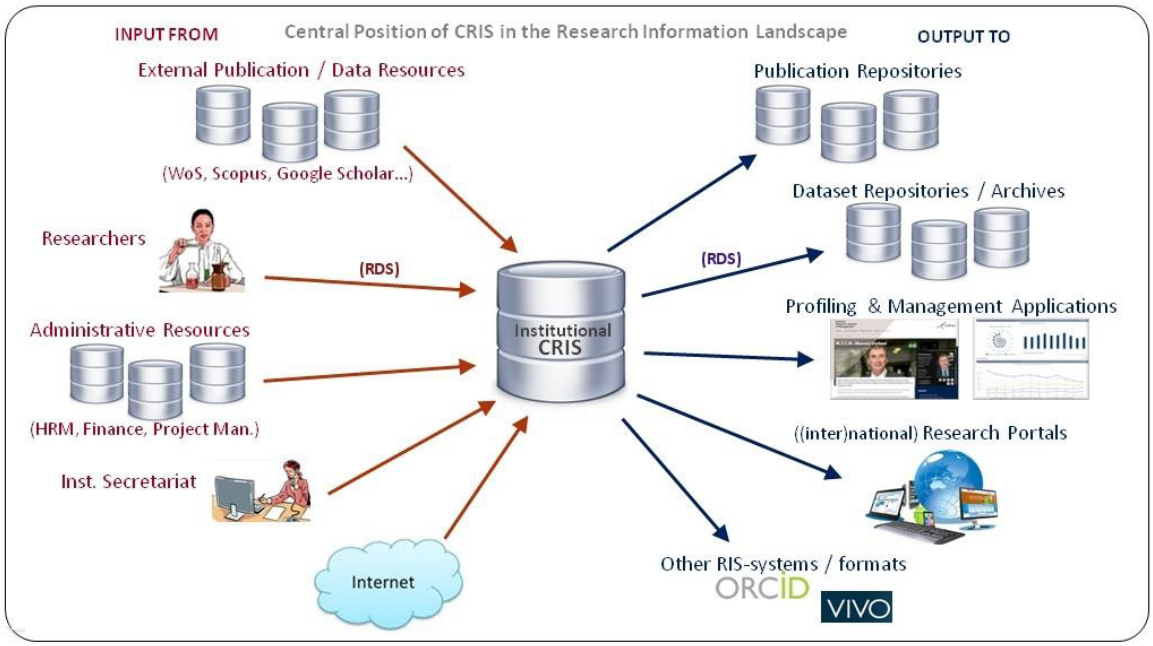

Figure 1. Central position of CRIS in the research information landscape

As Radboud University itself does not primarily offer data storage services for archiving data sets, the cooperation of a data-hosting provider – to accept upload of data sets and metadata from the CRIS – was necessary in order for the CRIS-oriented solution to work. Luckily, one of the Dutch national data archives, DANS (Data Archiving and Networked Services), was willing to work out a solution. This was ideal, because they were CRIS-oriented themselves already, as they host the national CRIS, called NARCIS.

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

- An *online interface* for the researcher, containing functions for uploading the data files and registering the metadata.
- A *technical solution*, handling the actual upload of files and metadata to the data-hosting provider and the automated communication workflows involved (e.g. the necessary e-mail alerts to involved parties/persons).
- *Support services*, the so-called *frontoffice-backoffice* model, distinguishing between services at the institution's level (Radboud University, the frontoffice) on the one hand and the national data-hosting provider (DANS, the backoffice) 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 that allowed a researcher to register his or her publications. The first and most visible goal of the RDS-project was to extend this interface with functions for registering and uploading data sets. Basically, this involved the following activities:

- The choice of a *standard (generic) metadata schema* for describing the various aspects of the data set. As point of departure, the international *DataCite* metadata schema was chosen.

- The implementation of the chosen metadata schema in an *online registration form,* i.e. fields on the screen to be completed by the researcher.
- Functions allowing the researcher to select and upload the *data files*.

An important function that was built into the interface was the possibility to directly link data sets to publications. The researcher can do this him- or herself, either from the already existing publication part of the interface or from the newly developed data set part.

Concerning the metadata, it should be noted that, apart from the generic descriptive metadata mentioned above, a data set may also pertain discipline-specific metadata, generally consisting of extensive, specific thesauri/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 part of the total package of data files uploaded to the data-hosting provider.

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 science, computer science, language studies and management science.

4. Making things work: the technical solution

4.1. The source: the CRIS (Metis) at Radboud University

This year is the 23rd anniversary of Radboud University's CRIS called Metis, developed by the university itself. At present, the Metis database holds approximately 200.000 metadata records on research results: publications, annotations, dissertations, lectures, patents and recently, as a result of the RDS-project, data sets. In Metis, there are different modules for different users. For the researcher, there is a separate (CRIS) interface, giving the researcher direct access to the registration and management of his or her personal research results. A researcher can add or modify results, add unique identifiers, define personal profiles, upload full text to the Radboud repository and, since 2015, archive data sets at the national DANS archive. In addition, it is possible to define relationships between data sets or between data sets 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,* a widely used and popular HTML-, CSS-, and JS-framework for developing responsive web applications and mobile first projects.
- *iBATIS,* a persistent framework which automates the mapping between SQL databases and objects in Java, .NET, and Ruby on Rails.
- *jQuery*, 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 web services. 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 data sets, with the data archiving system at DANS.

4.2. At the national data-hosting provider: DANS EASY

The data sets in the RDS-project are uploaded from the local CRIS to DANS through the so-called SWORD protocol and stored in *EASY* (Electronic Archiving SYstem), the certified long-term preservation archive of DANS (http://www.easy.knaw.nl; EASY is Data Seal of Approval (DSA), World Data System (WDS) and NESTOR-DIN certified). EASY offers functionality for manually uploading data sets, 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 data sets as in the RDS-project, other agreements apply, involving the frontoffice of the archiving institution (see below). EASY is built upon the Fedora platform, and maintained and further developed by DANS' technical staff.

Currently, EASY contains a rough 31 thousand data sets, mainly originating from the humanities and social sciences.

4.3. The SWORD-protocol

The SWORD-protocol (Simple Web-service Offering Repository Deposit) is a lightweight protocol for depositing content from one location to another. It is profile of the Atom Publishing Protocol (known as APP or AtomPub; see http://swordapp.org/about). Since SWORD is using an existing protocol, AtomPub clients and API's can be re-used. SWORD 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.

DANS has experimented in recent years with the SWORD-protocol, e.g. in the Enhanced Journals Made Easy!-project.[2] From 2014 onwards, work focused 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 setup with version 2.

4.4. Technical workflows

A researcher registers the metadata of the data set in the CRIS and uploads the data 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 University. The frontoffice at Radboud University Library checks the metadata and data files and, if necessary, communicates with the researcher for corrections/additions. Subsequently, the frontoffice initiates the upload to DANS through the upload function in the CRIS.

The Radboud University's CRIS makes an HTTP POST request to EASY, with the required header(s) and body. The zip file of the body contains (1) a file containing metadata with the name: DansDatasetMetadata.xml, and (2) a folder named 'data' containing the data files. The metadata file DansDatasetMetadata.xml is in the DDM (DANS Dataset Metadata) format, as defined by DANS.[3] The folder containing the data files may include subfolders, which will be included in EASY. DANS makes use of so-called preferred or accepted formats, to guarantee preservation for the long term.[4]

When the data set is accepted, a 202 HTTP response is sent. A confirmation e-mail is sent to the depositor (Radboud University), including the basic information of the data set and the persistent identifier (PID), more specifically a DOI, which uniquely identifies a data set. If the received data set is validated by DANS, the data set is saved in EASY. From this point on, the data set travels the same path as 'regular' data sets deposited through the EASY web interface. Since data curation is done by the frontofffice (see below), all data sets deposited from the Radboud University's CRIS will be published, with only incidental checks by the DANS data managers.

5. Support services as the key to success: the frontoffice-backoffice model

Radboud University's research data management policy states that 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

---

[2] 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).

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

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

data that researchers may have, an organization with a robust IT infrastructure and adequate support service is essential. Hence, Radboud University Library has established this support service as a frontoffice-backoffice model (FoBo), in close cooperation with the university's policy department, IT services and DANS. The library serves as frontoffice, while national data archive DANS functions as backoffice. The frontoffice-backoffice model is the organizational component of a federated data infrastructure under development in the Netherlands. The frontoffice supports, advises and trains the researchers and students in responsible data management. The backoffice ensures that the research data delivered is permanently and sustainable archived and made optimally available for discovery and reuse.

We describe the support services at Radboud University Library. Three aspects can be distinguished:

- The role of the library in research data management (RDM).
- Supplemental data policies by research institutes, combined with CRIS training sessions.
- Daily practices of data archiving by the frontoffice.

5.1. The role of the library in RDM: developing Research Information Services (RIS)

Characteristic of Radboud University's RDM policy is the steady, continuing role of the library. The library carries out the support, training and curation tasks. For several years already, offering services on *publication* management to researchers has been a shared undertaking by the library and the research institutes. New is the establishment of a *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 frontoffice-backoffice shifted from DANS to the institution's library. These integrated services of the library are summarized 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.
- 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 students and researchers. The RDS-project profits from the historically comprehensive network and integration of the library in the various faculties and research institutes.
- Most RDM-projects are temporal, delivering specific infrastructures. The role of the library is stable, irrespective of the duration of the RDM-project. RDM-support has become part of the library's daily business.
- Where policy departments focus on the development and implementation of policies, and IT departments emphasize the technical aspects of system development, 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?

5.2. Data policy services and training

Central in the RDS-project stands the development of the CRIS-interface (see above) and its deployment and implementation within (the institutes of) the university. Two aspects are fundamental for a successful deployment: (1) the existence of a data management policy in the institute, and (2) CRIS training sessions. First, 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 (such as: what to archive, who archives, what documentation etc.). To aid research directors, the university library and the policy department developed a checklist and offer support in developing and formulating these policies.

Second, in each research institute training sessions to groups of researchers are organized to implement disciplinary data protocols and explain the CRIS-interface. Practice learns that it takes time to develop and implement policies, while training sessions are easy to organize. At Radboud University, training sessions prooved to be an efficient way to introduce the CRIS-interface. They make data and publication management practical and approachable. 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).

5.3. The frontoffice-backoffice model in the daily process of data archiving

The collaboration between the Radboud University Library as frontoffice and data archive DANS as backoffice 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 data sets (DANS Easy interface) and register the accompanying metadata (CRIS and/or repository). DANS did communication with the researcher and the curation of the data set, while for publications the communication was taken care of by either the CRIS-staff or the repository manager. With the introduction of the new CRIS-interface and the frontoffice-backoffice model, Radboud University Library is handling the curation task and the communication with the researcher in an integrated way. The library acts as a *one-stop-service*: all information the researcher needs – website, support services, training sessions – can be found in one place. The library acts as an intermediary between the researcher and available infrastructures for publication and data management.

5.4. The frontoffice in practice

Radboud University Library checks data sets that are deposited via the CRIS-interface, before they are sent to DANS. First, the library checks the metadata, since adequate and rich metadata are vital to the *FAIR*-aspect of data sets (see above). Therefore, the metadata must be understandable for potential re-users, including a meaningful title and solid description.

Furthermore, the data files are checked, particularly on privacy-sensitive information. The files that are sent to DANS may no longer contain information that enables the identification of an individual. This information has to be removed or adjusted by the researcher. Moreover, the data files must be readable and understandable to a potential re-user. Therefore, the library checks if for instance 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 data set or metadata, the library contacts the researcher. Once the metadata and data files are correct, the library sends the files to DANS using the SWORD-protocol.

5.5. The backoffice in practice

Via the SWORD-protocol, metadata and data files are automatically deposited from Radboud University's CRIS to DANS EASY. Every data set is automatically assigned a DOI upon deposition, to enable sustainable and unique reference of the data set. In this model, Radboud University's data librarians do the data curation. The DANS data manager assist the data librarian if necessary, checks the data sets randomly and publishes the data sets. DANS ensures long-term storage and sustainable accessibility to the research data.

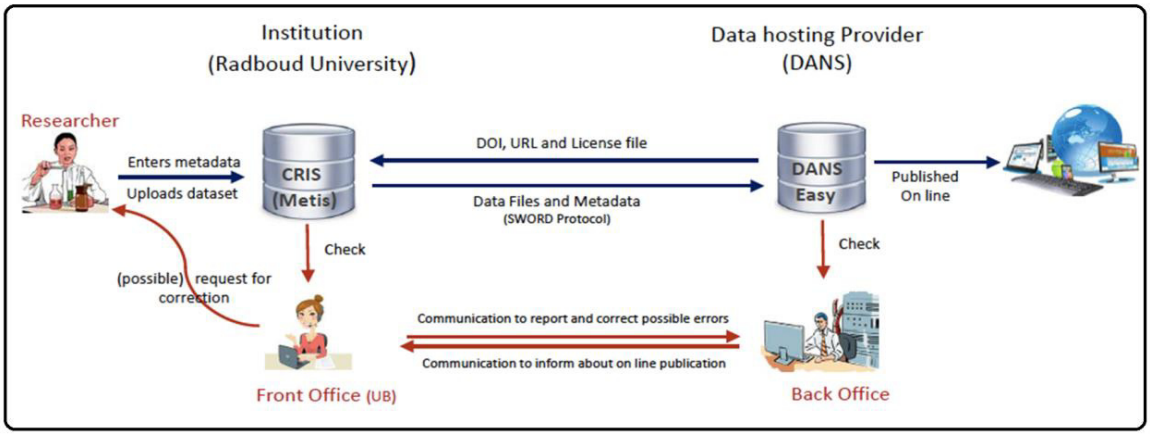

Figure 2. Frontoffice-backoffice model

6. Conclusions: lessons learned and future plans

The RDS-project has been a new and innovative development to the involved partners. To make sure that the resulting products are in fact useful for the researchers, the project involved pilot groups from the various research institutes. The groups reacted very positive to the concept of integrating research data management and archiving functions into the CRIS. Working together with researchers as well as our first experiences, revealed the following lessons learned and future plans:

- The best approach to implement the products resulting from the RDS-project is the benefit-for-researcher approach instead of the compliance approach. Thus, we made sure that researchers were actually involved when building infrastructures.
- From the researchers' perspective, a CRIS-oriented *one-stop-service* is essential to manage research information. From the institution's perspective, the RDS-project has strengthened the idea to use the CRIS as the primary resource for research information.
- By implementing one distinct frontoffice, for example at the university library, researchers have an easy accessible point of contact concerning their questions on the management of publications and data sets.
- Building a proper infrastructure is an iterative process of discovering flaws and finding solutions. Various aspects of the interface and workflows were adjusted during the project, in Radboud University's CRIS as well as in the DANS archive.
- The availability of optimal metadata, registered in an integrated system, is vital to guarantee the discoverability of research information. This is an important conclusion and hopefully one that will be taken to heart by the policy makers and project planners of research e-infrastructures or Open Cloud solutions.
- The concept of a *one-stop-resource* is taken a step further in a recently started project at Radboud Faculty of Medicine and the University Medical Centre, aimed at developing a *Digital Research Environment* (DRE). Whereas initially the DRE was meant to provide IT tooling to support research activities, under the influence of the RDS-project a research information functionality (CRIS-functionality) will be integrated.
- Further on in the project, CERIF will be introduced as XML format to send metadata from the CRIS to the DANS EASY archive. In this respect, a new project has just started, aimed at supplying project, person and organization metadata from institutional CRIS's in the Netherlands to the national NARCIS research information database by means of CERIF.

References

1. 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.