Research Data Management Protocol Kapteyn Astronomical Institute
Version 6.0, January 17t, 2019

INTRODUCTION

The amount of data on which Astronomy research is based is of the order of Petabytes and this amount is expected to rise. One can consider two kinds of astronomical data: observational data, coming from large instruments built by multinational consortia; and large theoretical simulations based on computer calculations. In addition to these, relatively small amounts of data are produced during data analysis. For most observational facilities all raw data are stored at large international data archives that routinely store all data from such facilities. These archives are guaranteed to continue to exist for several years in the future and have well-defined rules about data access. In most cases the data is available to the whole scientific community after a proprietary period of about 1 year. Other facilities (e.g. LOFAR) produce so many data that they are analyzed on the fly and only selected data are stored by the data centers. Scientific publications can involve large numbers of authors from many collaborating institutions.
Safeguarding the integrity of datasets stored at these international data centers is beyond the scope of the Kapteyn Astronomical Institute, as is verification of processing of data on computers that are not directly under supervision of the Institute. Since most of the important data belonging to publications of Kapteyn astronomers are published at international data centers, such as the CDS, or the Virtual Observatory, the institute feels its primary responsibility is to make sure that regular backups are made of the data its researchers are working on, so that no data are lost, and that in exceptional cases published results can be verified.

IMPLEMENTATION PLAN

For storing RDM plans, the Kapteyn Astronomical Institute uses a database that has been made available by the central university and the Faculty of Science and Engineering as of October 2016 for storing the individual RDMP's of staff and students.

Two RDMP questionnaires have been developed, one for staff and postdocs and another for PhD students and master students. In the questionnaire students and staff are required to describe the location of the data that they have used, to describe how the data was processed in such a way that processing of the data can be reproduced, and to describe how simulations have been done in such a way that the simulations can be reproduced. PhD-students have to do this within the first six (6) months from when they start their research project and master students within two (2) months. Staff and postdocs have to finish the RDMP questionnaire within three (3) months after their paper has been published.

A pilot project for implementing the RDMP procedures for (PhD) students started in 2016. Now the working methods are almost established, and all members of the institute are required to follow the general procedure, which is described in the protocol for Research Data Management for the Kapteyn Astronomical Institute, below.

Data storage by the Central University (CIT) has not been made available yet (Jan 2019). Research data can be stored in a folder /rdmp_data/users/ accessible on each desktop and

workstation. The folder will be referred to “project data space”.

RDM protocol

1. Data Collection

Scientific staff members (TT and (associate) professors) will write an RDMP when this is required for a grant, or when they are the responsible author of a peer reviewed publication. In the latter case they will do so within 3 months after a paper has been published.
PhD/master students are obliged to write an RDMP within 2-6 months after starting their research.
Postdocs are required to write an RDMP within 3 months of their end of contract.

All staff being a responsible author, all postdoc’s and all (PhD) students conducting research are required to upload the essential data underlying his or her peer-refereed publication and theses to the “project data space”. There is no need to repeat information that is already included in published material.
In astronomy it is often not possible to backup raw data because the size of the data easily exceeds that what can be stored on standard backup systems. Primary data is usually hosted by other institutes or observatories using public databases.
Therefore it is only required that staff, postdocs and students carefully describe the origin of the data and how it can be retrieved. The description is a text file with metadata and can be stored in the “project data space” unless this information is already included in the scientific publication. The data to be uploaded also consists of the final version of, or a link to, the published document, electronic versions of the published tables, and (if present) reduced images, program code and other datasets on which the publication is based.
When such a description is not part of the publication the responsible author should add a text file with a careful description of how results are obtained, also in case of simulations. It will be the decision of the responsible author whether intermediate results, a copy of code, input parameters and execution details are included in the project data space.

The responsible author can upload additional data or download previously uploaded data related to the publication. Previous versions should not be removed.

PhD/master students are obliged to maintain the “project data space” from the start of their research. Their data archive is made read-only within one month after the PhD/master thesis has been finished, and contains the data underlying the thesis and related publications. The supervisor has access to the data archive, and is ultimately responsible for the data. A student is graded only after approval of his/her data archive by the supervisor.

The responsible author or the supervisor decides who has access to the data in their project data space within the Kapteyn Astronomical Institute. The scientific director of the Kapteyn Institute will always be granted access.

2. Data Format, Storage and Back Up

Data is uploaded to hardware dedicated to store backup material. With a predicted average of 100 GB per year, the backups will be preserved for 10 years. A backup of the research data should be provided on a machine in another location (e.g. CIT). The data- and backup servers should be protected by a firewall.

Research data is archived by the author in a tar file. This single file is stored under a name that corresponds to the user name given in the existing RDMP web hosting data base. Together with a folder name equal to the user name, this guarantees easy retrieval of these files when necessary.

There will be no guidelines for data formats. In astronomy many different software packages are used, and each package can have its own data format. A restriction in data format implies a restriction in the choice of software, which we want to avoid at all times. Therefore, a description should be included in a text file with information about the used formats, unless these formats are currently standard (e.g. FITS, ASCII). There is always a potential danger that data formats are no longer supported in new versions of software packages for data reduction and analysis but it is not feasible to include software in the publications repository. Therefore we require that the file with metadata documents the version number of the used software.

When new program code is written, one should document the compiler or interpreter version. If the responsible author uses code for licensed software (IDL, Matlab), we cannot guarantee any period after which reconstruction of results is possible. Luckily, this applies only to a small number of projects because open source software and formats are the standard in the astronomical community.

3. Data Communication

Each entry in the project data space has amongst others reference to the responsible staff member, and the working title of the publication.

After publication, the final title is added. The uploaded files can be downloaded by authorized authors, but they cannot be removed from the archive.

4. Data Access, sharing and re-use

The datasets are accessible on demand only by the responsible staff member, the scientific director of the Kapteyn Institute, the Dean of the Faculty and the Board of the University. The data in the project data space remains with the Kapteyn Institute, after a staff member or a student leaves.

In the future the database will also be used to share data in and outside the University. The project initiator controls sharing. It is his or her decision with whom the data and codes can

be shared. The codes are subject to Dutch law (auteursrecht).

The scientific director of the institute has the role of a “super user” with access to all projects. He or she will only allow access to the data per project for auditing purposes. After leaving the Kapteyn Astronomical Institute the responsible author can obtain access through the scientific director.

5. Data Preservation and Archiving

The information in the document management system is maintained for 10 years. The Kapteyn Astronomical Institute guarantees that the system remains accessible during that period. The institute cannot guarantee the functionality of the (licensed) software, but access to the source code is always possible.

The Kapteyn Institute will appoint a data manager to keep an overview of entries, to make certain that data is deposited in time, to make things clear to new users, and to assist the director. The data manager is supported by the computer group of the Institute.

6. Responsibilities

Responsibilities of individual researchers and the description of relevant data are specified in the sections data collection and data format, storage and back up.

The scientific director of the Kapteyn Astronomical Institute, with support of the data manager, will ensure that their responsibilities are met fully and timely by regular checks. The scientific director will also ensure that the protocol with regard to data access, sharing and reuse, and data preservation and archiving is maintained.

PhD students are required to set up an individual research data management plan (iRDMP) and a matching data archive within six (6) months of the start of their PhD. A new version of the iRDMP has to be created whenever important changes to the project occur. The iRDMP is discussed during the R&D interviews. The first promotor is responsible to ensure the verifiability on the basis of the relevant data in the data archive by his or her PhD student. The supervisor remains ultimately responsible, and will ensure the completion of uploading of relevant data within three (3) months of publication.

Master students are required to set up an iRDMP within two (2) months after the start their research project. The supervisor is responsible to ensure the verifiability on the basis of relevant data in the database. A student is graded only after approval of his/her data archive by the supervisor.

The computer group of the Kapteyn Astronomical Institute is responsible for the infrastructure (hardware), maintenance and backups of the “project data space” system.