Astronomy discipline specific guidance on data sharing

Research data management / Digital Research / Library

06 March 2019

Astronomy discipline-specific guidance on data sharing

Version 1	06/03/2019	Eva Borger	Finalised document following review by the working group, approved by the Head of School

Contents

Introduction

Funders have introduced open data policies[1] that require researchers to share their data. Specifically, all UK Research Councils, including STFC, subscribe to the UK Research and Innovation (UKRI) Concordat on Open Research Data. However, these policies often lack a clear definition of 'data' in different scientific disciplines.

Therefore, the Research Data Management team at the University Library are working closely with Schools to develop a series of discipline specific guidance documents with the aim of answering the following questions:

- What are 'data' or 'digital objects'?
- What data should be deposited to be compliant with funder policies?
- What is a 'theoretical' paper where data sharing does not apply?
- What 'exceptions' to data sharing are acceptable?
- How and when should we publish analysis code and scripts?
- What is a digital object identifier (DOI) and what it is good for?

The discipline specific guidance aims to offer a reference document for researchers depositing data and its development has been collaborative process between the RDM team and researchers at St Andrews. **This document is not intended as a policy, but rather as guidance agreed by the working group**.

Data types and suggested sharing practices

This section lists the most common data types anticipated to be relevant to Astronomy researchers in St Andrews, as identified by the working group. For each type of data, it describes two possible sharing practices:

- **Ideal sharing practice:** the method of sharing data identified as the best possible practice; this, however, might not always be practical or achievable with the current available resources.
- **Acceptable sharing practice**: the method of data sharing identified as most often practical and easily achievable to satisfy funders' data sharing requirements.

As a minimum, researchers should adopt the acceptable sharing practice and, when possible, every should make every effort to implement the ideal sharing approach.

The working group identified the following types of digital outputs as most common in different fields of Astronomy represented in the School:

1. Large statistical datasets
2. Simulations
3. Observatory datasets
4. Observatory datasets without public archive
5. Proprietary data
6. Derived data
7. Analysis pipelines and code

[1] https://www.st-andrews.ac.uk/library/services/researchsupport/researchdata/researchdatapolicies/fundersresearchdatapolicies/

Table 1 summarises acceptable and ideal sharing practices for the different data set types identified by the working group.

Dataset type	File size range	Ideal sharing practice	Acceptable sharing practice
Large statistical datasets	TB - PB	▫ Cite data release and release paper	Cite data release and release paper
Simulations	TB - PB	▫ Share/cite software (where possible) ▫ Raw data	▫ Share/cite software (where possible) ▫ Raw data on requests
Observatory datasets	TB - PB	▫ Share processed data products (where possible) ▫ Cite raw data, pipeline version, data reduction software version, observing log	Cite raw data, telescope, archive name
Observatory datasets (no public archive)	GB - TB	▫ Share FITS header, README, calibration files ▫ Raw data if practical or on request ▫ Observing log	▫ Share FITS header, README, calibration files ▫ Raw data if practical or on request
Proprietary data	GB - PB	Cite data source with maximal available context information	Cite data source with maximal available context information
Derived data	small	▫ Create + cite software release with DOI ▫ Cite VIZIER tables ▫ Provide tables as supplementary text files	▫ Cite GitHub page ▫ Cite VIZIER tables ▫ Provide tables as supplementary text files
Analysis pipelines/ code	small	▫ Create + cite software release with DOI ▫ Cite VIZIER tables ▫ Provide tables with identifiers/ coordinates as supplementary text files	▫ Cite GitHub page ▫ Cite VIZIER tables ▫ Provide tables with identifiers/ coordinates as supplementary text files

Table 1: Data types and sharing practices in Astronomy

Depositing, sharing and citing - What to do.

Permanent identifiers and citation

Datasets and code should ideally be deposited in a repository that provides a permanent identifier for the record, e.g. a digital object identifier (DOI). These represent a permanent link to the record, therefore ensuring sustainable data citation, and facilitate information sharing between electronic systems through registration of structured metadata.

Where a permanent identifier is available, this should always be used when citing the data.

It is generally recognised that many data sources in the field of Astronomy do not provide a DOI, but are instead characterised by identifiers, names or coordinates specific to the repository or hosting platform. In these cases, data should be cited with as much detail as possible, together with a link to the hosting platform.

Where to share

Preference should always be given to recognised and widely used **discipline-specific repositories or hosting platforms**[2].

In the field of Astronomy, **GitHub** is commonly used to store and share computational code[3]. It is recommended as it offers built-in version control with Git and includes integration with Zenodo to make code releases or snapshots citable using a DOI. Other commonly use resources include the Astrophysical source code library.

> *What to do*[4]: 1) Deposit your code to your GitHub repository. 2) At the end of the project/ milestone, create a release in Zenodo and obtain a Zenodo DOI. 3) Create a reference record in Pure based on the deposit in Zenodo with its DOI. 3) Include a statement about data/ software access in the manuscript, including the DOI from Zenodo.

Where no such observational database or repository exists, the University's research information system, **Pure**, can be used. Pure provides a DOI for each record and enables linking these to other content, such as publications, theses and funding projects.

Sharing large observational data or image cubes

Large datasets produced in-house (e.g. calibrated data) cannot always be shared easily through standard mechanisms such as upload to Pure.

Where derived datasets have been produced through computational methods, it will often be possible to **share or cite the computational code** or software that was used, together with any available documentation. It can be assumed that research and development will make computational power and storage less limiting in the future. As such, sharing should also be considered for computations that might require significant computational resources.

In other cases, large datasets might be a significant valuable resource that should be kept and made available in the long term, but where there is no suitable public or discipline-specific repository. Where this is the case, authors routinely indicate that data can be **made available upon request**. In order to be able to honour such requests, it is imperative to provide means of contact that will be available in the long term.

> *What to do*: 1) Use your ORCID iD when submitting your publication and link it in Pure. 2) Create a dataset record in Pure. 2) The RDM team will provide you with a DOI for the record. 3) Include a statement about data access in the manuscript, including the DOI. The DOI will

[2] See also 'How to publish data' on the research data management web pages.
[3] Simon Portegies Zwart (2018), Computational astrophysics for the future, Science 361(6406) pp.979-980, DOI: https://doi.org/10.1126/science.aau3206
[4] Michael Jackson (ed.) (07 August 2018). SoftwareDeposit: Guidance for Researchers (Version 1.0). Zenodo. doi:10.5281/zenodo.1327310

link to a metadata record where contact information can be given (and maintained) and where data can be made available in the future[5].

Sharing information about observatory data, including astronomical catalogues

Datasets arising from a study that are being deposited or catalogued elsewhere (e.g. VizieR) do not need to be shared again separately. In these cases, data deposits or tables should be cited/ linked in the publication with sufficient detail (see Table 1) and a reference record should be created in Pure.

What to do: 1) Create a dataset record in Pure, including a link to the catalogue entry[6]. 2) Include a statement about data in the manuscript, using a DOI, if available, or other identifier provided by the repository/ catalogue and a link to the catalogue entry. Where appropriate, functionality provided by journals to link external data (e.g. links to catalogues, AAS, SIMBAD) can also be used. However, if an article is distributed separately (e.g. as preprint), an access statement should be included in that version.

Re-using and sharing software and code

Code might contain elements contributed or originally created by others or might depend on proprietary code, where **intellectual property** rules apply.

Researchers should ONLY share data and code which they own or for which they have permission to do so. Permission might arise through collaboration with the originator or by re-using code published under an open license (or equivalent statement) which permits re-use and sharing. Where no such permission exists, a legal exception to sharing applies (see below).

What to do: 1) Consider sharing the subset of the code/ software that you *can* share (as above). 2) Include a statement in the manuscript describing the restriction and, if possible, provide details of where requests for data or materials can be made.

Code might also **depend on specific software packages that could become unavailable** in the future. However, it can still be a valuable resource that demonstrates the approach that was taken and the methods that were used. In addition, future end users might have access to a copy of the necessary packages.

What to do: 1) Share and cite the code that is suitable for sharing (as above).

What is a 'theoretical' paper where data sharing does not apply?

Data sharing does not apply where **no primary data/ digital outputs** have been used or produced (for example review articles, editorials) or where a publication describes the development of a theory without including computations or simulations that required the creation of scripts or input parameters.

What to do: Where possible, include a statement in the manuscript that indicates that no primary data was produced or re-used.

[5] The University is actively working on mechanisms that will enable sharing of large data files through dataset records in Pure. Please contact the research data management team for further information.

[6] For example: https://risweb.st-andrews.ac.uk/portal/en/datasets/vizier-online-data-catalog-stellar-kinematics-in-califa-survey-falconbarroso-2017(a35f654b-6720-4b99-9502-6a6ddffe2f04).html

Likewise, if a publication is based on the **re-use or meta-analysis of secondary data** and there is no separate digital output, data sharing would not apply.

> *What to do*: Include a data citation or a statement in the manuscript indicating the sources of the secondary data (e.g. URL/ DOI, link to repository with accession number).

Which 'exceptions' to data sharing are acceptable?

Data sharing can be restricted where there are **legal or ethical reasons** that can include intellectual property rights (e.g. software/ code containing other researchers' code) or data or material transfer agreements with external companies.

Researchers are expected to ONLY share data which they own or where they have permission from the owner to re-share (e.g. through an open license or direct permission). Researchers must NOT share data on behalf of collaborators without their consent.

> *What to do:* Include a statement in the manuscript describing these reasons and, if possible, provide details of where requests for data or materials can be made.

In addition, datasets can be **embargoed** for a limited period, where funder policies permit this, to allow researchers a period of exclusive access in order to complete further studies.

> *What to do*: 1) Create a dataset record, for example in Pure, including the data files and set an embargo on the files (or request it from the RDM team). 2) The RDM team will provide you with a DOI for the record and manage the release of the files at the end of the embargo period. 3) Include a statement in the manuscript about the embargo and add the DOI.

If you have any questions about sharing/ publishing your research data or the guidance provided above, please don't hesitate to get in touch with the Research Data Management team:

Old Union Diner, Butts Wynd

St Andrews, KY16 9AL, Fife, Scotland

Tel: +44 (0)1334 462322 or 462343

Email: research-data@st-andrews.ac.uk

The working group

Ian Bonnell, Claudia Cyganowski, Martin Dominik, Alexander Scholz, Paula De Viveiros Teixeira, Rita Tojeiro, Bert Vandenbroucke, Anne-Marie Weijmans, Vivienne Wild