School of Chemistry discipline specific guidance on data sharing

Author	Eva Borger
Publication date	March 2019
Review date	March 2021
Document owner	Research Data Management team
School / Unit	Research & Innovation Services
Document type	Guidance
Version	2.1

Version number	Date	Author	Purpose / Changes
1	10/12/2019	Eva Borger	Finalised document following review by the working group
2	26/03/2019	Eva Borger	Include additional sections on metadata creation, code, sharing and IP, in response to feedback.
2.1	29/10/2019	Federica Fina	Added document information table and modified version control table

Contents

Introduction

Funders have introduced open data policies that require researchers to share the data underpinning their funded research. Specifically, all UK Research Councils, including EPSRC[1], subscribe to the UK Research and Innovation (UKRI) Concordat on Open Research Data[2]. However, these policies often lack a clear definition of 'data' in different scientific disciplines.

In order to provide guidance tailored to the School of Chemistry, a working group bringing together representatives of the Research Data Management team at the Library and the School of Chemistry was formed. The working group developed a guidance document aimed at answering the following questions:

- *What are 'data' or 'digital objects'?*
- *What data should be deposited to be compliant with funders' policies?*
- *What is a 'theoretical' paper where data sharing does not apply?*
- *What 'exceptions' to data sharing are acceptable?*

This document is not intended as a policy but rather as guidance agreed by the working group.

The working group

Professor Sharon Ashbrook, Dr Bela Bode, Dr Paul Connor, Dr Gordon Florence, Dr Herbert Fruchtl, Dr Finlay Morrison, Dr Renald Schaub, Professor Paul Wright, Dr Eli Zysman-Colman

Data types and suggested sharing practices

This section provides a list of the most common data types anticipated to be relevant to the School of Chemistry, as identified by the working group. For each type of data, the document 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, should make every effort to implement the ideal sharing approach where possible.

The main type of digital outputs identified by the working group as common in different fields of Chemistry represented in the School are:

- Numerical data, typically describing XY-coordinates, which could be plotted
- image files or imaging datasets,
- PDF printouts of raw numerical or plotted data,
- text files with input and output parameters/ code,
- metadata associated with machine outputs.

[1] https://epsrc.ukri.org/about/standards/researchdata/
[2] https://www.ukri.org/funding/information-for-award-holders/data-policy/

Table 1 lists specific techniques, data types and feasible data sharing practices as identified by the working group.

Table 2 lists input/output file types and sharing practices for different computational analyses.

What to include

As a general principle, where **numerical** data can be created or obtained from the instrument, these should be provided., **metadata**.

In addition, a rich set of **metadata and documentation** should accompany each dataset. This might include data provided by the instrument or data otherwise created by the researcher (e.g. describing variables, experimental settings, instrument/ software used). The creation, quality control and maintenance of experimental (data level) metadata and documentation, as described in Table 1 and Table 2, is the responsibility of individual researchers and should follow best practice in the field. Any personal or other information that is not suitable for sharing before should be removed before deposit.

In addition to documentation, structured machine readable metadata is typically produced during deposit in a data repository such as Pure or in discipline-specific data repositories. For data deposited in Pure, researchers are responsible for the entry of appropriate metadata in Pure records that will subsequently be maintained by the University Library.

Data sharing and intellectual property

A dataset or script might contain elements contributed or originally created by others or might depend on proprietary code, where **intellectual property** rules apply. This would also be the case where a collaborator on a manuscript does not agree on data be shared publicly.

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 (and their explicit consent) or by re-using code published under an open license or together with an 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. 2) Create a dataset record in Pure and add a description, which specifies what the record contains. 3) The RDM team will provide you with a DOI, which you can add to the manuscript.
>
> *Alternatively*: Where no permission for data re-sharing exists and there is no subset of the data or code to share, add a sentence to the manuscript describing the restriction and, if possible, provide details of where requests for data or materials can be made.

Large datasets

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

Where such datasets have been produced through computational methods, it will often be possible to share a representative input and running file together with all its outputs. Generally, it will be possible to **share or cite**

the computational code or software that were used, together with any available documentation. Technological 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 themselves that cannot be reproduced might be a significant valuable resource that should be kept and made available in the long term. 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 link to a metadata record where contact information can be given (and maintained) and where data can be made available in the future[3].

Computational code

In some instances, computational code, scripts or pipelines might be produced or existing software or code might be adapted. The University recommends using GitHub to deposit software and code, as it offers built-in version control with Git and includes integration with Zenodo to make code releases or snapshots citable using a DOI. Alternatively, Pure can be used to deposit scripts that are used for data processing alongside the dataset.

> *What to do*: 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.
>
> *Alternatively*: 1) Create a dataset record, for example in Pure[4], including the data files and scripts. 2) The RDM team will provide you with a DOI for the record. 3) Include a data access statement in the manuscript and add the DOI link.

Where to share

Preference should be given to discipline-specific data repositories or centres, which facilitate discovery by other researchers in the field of study and provide the means to create appropriate rich metadata to accompany the deposit. Where no such repository exists or the data format is not suitable for deposit, the data can alternatively be deposited in Pure[5].

[3] 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.

[4] https://www.st-andrews.ac.uk/staff/research/pure/

[5] See: https://www.st-andrews.ac.uk/library/services/researchsupport/researchdata/publisharchiveandpreserve/depositinpure/

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:* Include a clear statement in the manuscript indicating that no primary data or code were used or produced during preparation of the publication.

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 clear statement in the manuscript indicating the sources of the secondary data (e.g. URL/ DOI, link to repository with accession number) or a data citation[6].

Which exceptions to data sharing are acceptable?

Data sharing can be restricted where there are **legal or ethical** reasons that, in the field of Chemistry, can include data or material transfer agreements with external companies.

> *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 of time if necessary, 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[7], 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. 3)Include a statement in the manuscript about the embargo in the publication and add the DOI link to the metadata record where data will be made available in the future.

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

[6] https://www.st-andrews.ac.uk/library/services/researchsupport/researchdata/publisharchiveandpreserve/datacitationandaccessstatements/

[7] https://www.st-andrews.ac.uk/staff/research/pure/

Table 1: Analytical techniques, data types, formats and sharing practices

Technique	Instrument/technique output files	Ideal sharing practice	Acceptable sharing practice
Auger Electron Spectroscopy (AES)	Instrument output files or ASCII export	Raw data file from instrument and ASCII export	Raw data file from instrument
Atomic Force Microscopy (AFM)	XY-type numerical data from the instrument with metadata	Raw file from the instrument + metadata	Raw file from the instrument + metadata
Battery cyclers	Raw data can be exported from instrument internal database as text-based files (iqx/ ciqx)	Numerical/ text-based data export	Numerical/ text-based data export
Circular dichroism	XY-type numerical data	Numerical/ text-based data as exported from the instrument	Numerical/ text-based data as exported from the instrument
Conductivity	XY-type numerical data from the instrument	XY-type numerical data from the instrument	XY-type numerical data from the instrument
Dilatometry	XY-type numerical data: : .cle, .dle; .csv, ASCII (.txt)	Raw data file from instrument + ASCII (.txt) file if the raw data file is proprietary or not human readable	ASCII (.txt) or .csv
Differential thermal analysis (DTA)	XY-type numerical data: RSD62 or .ssu / .dsu / .bsu / .msu; ASCII (.TXT), .csv, Enhanced Metafile (.emf)	Raw data file from instrument + ASCII (.txt) file if the raw data file is proprietary or not human readable	ASCII (.txt) or .csv

Technique	Instrument/technique output files	Ideal sharing practice	Acceptable sharing practice
Differential scanning calorimetry (DSC)	..ngb-dd3; .csv, ASCII (.txt), .pdf	Raw data file from instrument + ASCII (.txt) or .csv file	ASCII (.txt) or .csv, .pdf
Energy-dispersive X-ray spectroscopy (EDX)	Text-based output	Raw data file from instrument (Word/ text document)	Raw data file from instrument (Word/ text document)
Elemental analysis (EA, external service)	PDF of output displayed by instrument (there is no raw output file)	Scanned copy of document (without signature)	Cite numerical data in paper
Electrochemical impedance spectroscopy (EIS)	.z, .fra, .txt	Human readable files + metadata (if more than one file format is available, priority should be given to the one with the highest level of metadata)	Raw data file from instrument. Preferably human readable, e.g. as txt file or outputs compiled in a spreadsheet + metadata.
Electron paramagnetic resonance spectrsoscopy (EPR)	XY-type numerical data: .dta and .spc files and ASCII descriptor files (.dsc and .par)	Raw data file from instrument + ASCII data export	Raw data file from instrument.
Gas chromatography (GC)	XY-type numerical data	Numerical/ text-based data export	Numerical/ text-based data export
Gas-chromatography-mass-spectrometry (GCMS)	XY-type numerical data + PDF	Proprietary raw data file from instrument + numerical/ text-based data export	Raw data file from instrument or pdf output
Gas-chromatography-mass-spectrometry	PDF, (XY-type numerical data, though difficult to obtain)	Proprietary raw data file from instrument + numerical/ text-based data export	PDF

Technique	Instrument/technique output files	Ideal sharing practice	Acceptable sharing practice
(GCMS, Internal School-level service)			
High-performance liquid chromatography (HPLC)	XY-type numerical data + PDF	Raw data file from instrument + numerical/ text-based data export	Raw data file from instrument or pdf output
High resolution mass-spectrometry (HRMS – external service)	PDF	PDF	PDF
High Resolution Electron Energy Spectroscopy (HREELS(	Instrument output files or ASCII export	Raw data file from instrument and ASCII export	Raw data file from instrument
Immittance spectroscopy	.csv, .z, .dat	Human readable files with metadata (if more than one file format is available, priority should be given to the one with the highest level of metadata)	Raw data file from the instrument. Preferably human readable, e.g. as txt file or outputs compiled in a spreadsheet + metadata.
Infrared Spectroscopy	XY-type numerical data	raw data from the machine + numerical/ text-based data export	PDF
Isothermal titration calorimetry (ITC)	XY-type numerical data	Raw data file from the instrument	Raw data file from the instrument

Technique	Instrument/technique output files	Ideal sharing practice	Acceptable sharing practice
Laser scanning microscopy (LCSM)	Instrument/ vendor-specific file or image file (TIF, png, jpg)	Raw data file from the instrument + text files with metadata/ data	Raw data file from the instrument (these can mostly be imported into open software, e.g. ImageJ); or: image file (TIF format)
Low Energy Electron Diffraction (LEED)	Image file from instrument (bmp,png,...) + ASCII for metadata	Image file from instrument (bmp,png,...) + ascii for metadata	Image file from instrument (bmp,png,...) + ascii for metadata
Mass spectrometry (MS)	Instrument/ vendor-specific file + csv	Proprietary raw data file from the instrument + numerical/ text-based data export	Raw data file from the instrument
Nuclear magnetic resonance spectroscopy (NMR)	BRUKER raw data, dmfit	Raw data file/folder from the instrument (BRUKER file)	Instrument generated file (Folder from Nomad for solution state NMR and BRUKER folder for Solid State NMR). Note that files produced by third party processing packages (e.g., MestreNova or dmfit) lack important metadata and should not be used as the only source of raw data. They can be used (i) if the raw data is not available from an external source or (ii) additionally provided if desired for information about processing or analytical fitting.
Optical Microscopy	Image file (.bmp, .jpg, .png, TIF)	Image file (TIF)	Image file (preferably TIF)
Optoelectronic characterisation	Text based XY-type numerical data (Excel or Origin)	Numerical/ text-based data export (Excel or Origin)	Numerical/ text-based data export (Excel or Origin)

Technique	Instrument/technique output files	Ideal sharing practice	Acceptable sharing practice
Polarisation-field measurement	.thf (exportable as .dat)	.dat	.dat
Porosimetry	XY-type numerical data	Numerical/ text-based data export (spreadsheet, csv)	Numerical/ text-based data export (spreadsheet, csv)
Powder X-ray diffraction (PXRD)	.raw, .udf, .xrdml, .dat	Raw data file from the instrument + numerical/ text-based data export (e.g.: .udf and .xrdml)	Raw data file from the instrument or numerical/ text-based data export
Pyrocurrent measurement	XY-type numerical data (.csv)	Raw data file from the instrument	Raw data file from the instrument
Scanning electron microscopy (SEM)	Image file (.jpg, .png, TIF) + metadata text file	Image file (TIF) + metadata text file	Image file (preferably TIF) + metadata text file
Scanning Tunnelling Microscopy (STM)	Instrument/ vendor-specific files: .dat, .tspec, .lat, .vert, .pllspec	Raw data file from the instrument + text-based export	Raw data file from the instrument containing metadata
Thermally Programmed Desorption (TPD)	Instrument ASCII output	Raw ASCII data file from the instrument	Raw ASCII data file from the instrument
Transmission electron microscopy (TEM)	.jpg, .png, .tiff, .dm3 (Gatan) + internal file	dm3 file + image file (preferably TIF)	Image file (preferably TIF)

Technique	Instrument/technique output files	Ideal sharing practice	Acceptable sharing practice
Thermal gravimetric analysis (TGA)	RSD62 or .ssu / .dsu / .bsu / .msu; ASCII (.TXT), .csv, Enhanced Metafile (.emf)	Raw data file from the instrument + ASCII (.txt) or .csv	ASCII (.txt) or .csv
UV-Visible spectroscopy	.jcw	Raw data file from the instrument + numerical/ text-based data export	Raw data file from the instrument + numerical/ text-based data export
X-ray photoelectron spectroscopy (XPS)	.vms	.vms (casaXPS)	.vms (casaXPS)
X-ray diffraction (Crystal XRD)	.cif	CCDC accession number	CCDC accession number in the paper

Table 2: Computation type, inputs, outputs and sharing practices

Technique	Instrument/technique output files	Ideal sharing practice	Acceptable sharing practice
Castep DFT	Input files: text files; Output files: .castep and .magres	Input files + output files	Output files are enough if it is possible to extract the input from the output files.
Computational files	Text-based input and output files + README	Text-based files	Text-based files

Technique	Instrument/technique output files	Ideal sharing practice	Acceptable sharing practice
DFT calculations	ASCII output + binary checkpoint file (software dependent)	ASCII file + binary checkpoint file	ASCII file
Gaussian computation	.log and .out files	Input + output files (text-based)	.out files, if it is possible to extract the input from output + log file
Matlab	Numerical simulation scripts as .m files + ASCII data export	ASCII data export	ASCII data export
Simpson simulation	Text based input and output files	.fid, .spe (output), .in (input)	.fid (output), .in (input)
VASP calculations	VASP files, e.g. contcar, incar, kpoints, poscar, outcar	Include outcar file format	Numerical/ text-based data export