Describing your dataset in PURE

Please contact the RDMS service if you require help with completing the dataset metadata template, organising and uploading your files or in deciding where your data should be deposited.

These guidelines for describing each field in the PURE dataset template should be used in conjunction with the PURE Deposit Upload Guide. The purpose of the dataset module in PURE is specifically to describe the data itself (file format, software required, availability issues) and its relations, to people, publishers, associated content. It should **not** be used simply to duplicate the details of any related publication or project. The “Description” field in particular should be used to provide detailed information on the software environment employed to create the dataset, the resulting file formats to be preserved and any relevant contextual information such as details of any restrictions to access and (re)use.

These guidelines aim to ensure that dataset records added to PURE are of a suitable quality to enable data discovery, determination of value, access and reuse (with citation). The metadata content entered here will be reviewed by RDMS support staff and assessed to ensure compliance with certain minimum requirements (See Appendix). A Digital Object Identifier (DOI) cannot be minted nor will the dataset record be validated and made public on the university research portal until these minimum requirements are achieved. Once a DOI has been minted the data and metadata for the record should not be altered, in order to preserve the integrity of the data.

To create a new dataset record simply log into PURE and open a new “Datasets” template by using the ‘+’ button or by clicking on the large green ‘add new’ button. You will be presented with a template with a number of fields, those starred (✱) are mandatory required fields. Some basic help on the content of the fields is available on the Pure record by clicking the information ℹ icon in the top right corner of the record.

Dataset template Metadata Fields

Identification

Title ✱

The title of the dataset should be informative and concise, but descriptive enough to convey the dataset content.

- Titles should be as descriptive as possible so that they are meaningful to researchers from other disciplines and to future readers. Keywords included in the title are important in promoting discovery by search engines.
- Titles should be unique and generally should not use acronyms or abbreviations that would be a barrier to understanding outside a discipline. If acronyms are important for discovery, expand them or define them in the description.
- If appropriate you can use the title of the study or project.

- If the title of the dataset duplicates exactly the title of a related research publication add Data for: "…" to the title.
- Titles should be limited to a maximum of 250 characters.
- The nature of the data and other information such as the location and date range are useful in the title, particularly if they are not recorded elsewhere in specific metadata fields for spatial and temporal coverage.

Description

The description allows you to provide a comprehensive contextual description of the data. Provide details on what the data is intended for, what was the source of the data, how was it collected or created, the precise hardware/software used, if there are any significant features that would be useful to a potential re-user of the data. Some areas to cover in the Description:

- Contextual information: background, aims.
- Data collection methodology: process, instruments, data validation, derived variables, weighting, secondary data sources
- Dataset structure: data files and relationships
- Variable level documentation: labels, codes, classifications, derivations
- Data confidentiality: anonymisation, consent, access conditions of use
- If access to the data is restricted include a statement outlining why
- Detail on the ownership of dataset
- Identify any 3rd party data that has been re-used

See supplementary guidance from the UK Data Service on recommended data formats

Date of data production

If recording the timescale involved in creating the data is a significant element to the research then use this field to record a specific date or a period of time in which the deposited data was produced.

People *

Records the significant contributors involved in the creation or management of the data. People internal to University of Strathclyde should automatically be recognised by the system and can be added by clicking on the "Add person" button. Use the "Create external person" option for people out with the university.

There is only one mandatory role option:

Creator *: person with main responsibility for the development of the dataset

A number of roles are listed in the drop down menu which ideally should be agreed during the designation of roles and responsibilities for research collaborators and partners in the Data Management Plan.

Use "Add organisational unit" in the same way to record the organisation and department relevant to the person.

Dataset managed by
Managing organisational unit *

This automatically defaults to the Organisation Unit at Strathclyde of the person creating the dataset entry. This field indicates the organisation or department which has overall administrative/editorial permission for the record not for the data itself. Change if appropriate.

Data Availability

Publisher *

The publisher in this context is the institution or facility where data is deposited, managed and, where authorized, made available from. For data deposited in PURE the default publisher is always "University of Strathclyde". For data deposited externally, the publisher is the external facility and can be added using the "Change publisher" button. All data produced at Strathclyde but deposited externally must also have a corresponding metadata record created in PURE which links to the external facility data record and defines the terms and conditions under which the data is held there.

DOI (Digital Object Identifier)

A DOI uniquely identifies a dataset and are considered 'best practice' for the accurate capture of citation metrics and therefore demonstrating impact for the research. Data that is cited in a consistent and machine readable way is readily identifiable by citation indexing services such as Scopus. There are two sources for obtaining a DOI for a dataset:

- On deposit in PURE, the DOI will be minted at Strathclyde by PURE staff on validation of the dataset record.
- Minted by an external publisher/data repository. These can be included in the Pure dataset record using the "Add existing DOI" option.

There are four fields that **cannot be changed** after a DOI has been minted for a dataset record, these are: ***Title***, ***Creator*** (listed under People), ***Publisher*** (publisher of the dataset, usually University of Strathclyde), and ***Date made available*** (the date the dataset record is made public not the data files, which may be under embargo or restricted). If any of these four fields need to be changed, we would advise creating a new dataset record.

Using external Research Data Repositories

Where data has been deposited in an external data repository (e.g. UK Data Service, NERC Data Centre, Zenodo) there is a compliance requirement to ensure a marker for this data is stored in PURE. To this end a ***data registry record*** should be created for this dataset in PURE which includes the DOI supplied by the external data repository.

If you wish to deposit externally please consult the RDMS team for advice prior to any deposit. The main reasons for choosing to deposit data in a repository external to Strathclyde include:

- As a condition of award or funder mandate to deposit in a designated repository

- A “responsible repository” recognised within your discipline where data will receive the maximum exposure and impact

To be considered a “responsible repository” we would expect an external facility to provide, at a minimum:

- A valid DOI
- Evidence of sustainability (e.g. hosted or underwritten by a reputable academic institution or funder)
- Formal agreements of responsibility (e.g. data return policies, terms & conditions of deposit and use)

If an external repository is unable to fulfil these minimum requirements further evaluation by RDMS staff and additional actions (e.g. an additional dataset backup retained in PURE) may be required to ensure the security of the data. Further information on using external data repositories is available on the RDMS support pages and a comprehensive listing and assessment of external facilities is maintained by the re3data.org registry.

Receiving a Place-holder or “Dummy” DOI

There will be occasions when a DOI is required but the dataset record is not complete or the data is not ready to be deposited. This usually occurs when a DOI is required for inclusion in a forthcoming research publication or thesis data statement. In such cases a place-holder DOI should be requested at the manuscript preparation stage but it can only be issued by Pure administrators once a basic Pure dataset record has been created. The process is as follows:

- Create a basic Pure dataset record. The minimum requirement is to add a Title, a Person/Creator and a Date made available
- Save the record “for validation”
- There is no need to upload any data files at this stage
- A Pure administrator will be in contact, and request that a place-holder DOI be issued

The place-holder DOI is an inactive version of, what will be, the final active DOI and can be added to a publication’s data statement immediately. The time prior to publication acceptance should be used to prepare the supporting data, complete the dataset record metadata and upload the final version of the data. Place-holder DOIs will only be issued with the understanding that the dataset record and the data upload must be completed and the record validated, creating a public landing page on the university research portal, **before** the related publication is made public. The DOI will not function if the DOI is made public but there is no public landing page on the portal for it to resolve to.

Upload the appropriate file(s) by dragging the files into the upload box or by browsing the relevant files on your computer. Multiple individual files can be selected but if large numbers of files in one directory are present zip the folder and upload as one zip file. The integrity of a folder will not be retained if dragged & dropped, always zip a folder before upload. We would define a dataset as one or more files plus documentation. By including appropriate data documentation such as a README file or a data management plan (DMP) as part of the dataset record, the data can be made significantly easier to understand, interpret and reuse. This is important for open access and authentication of research findings. If uploading a zipped file containing multiple files it is recommended that a README file is created and uploaded separately from the zipped file, listing the filenames and file formats of the data contained in the zipped file. A simple file list of folders in Windows can be created using the command dir/s >listmyfolder.txt from the command prompt.

See supplementary guidance from the UK Data Service on naming and organising files and folders

Electronic Data

Note on special data types:

Thesis Data

Data which supports a thesis should be assessed for its suitability for distribution at an early stage, including any intention to deposit in support of a research publication. Restrictions on access can include a moratorium or digital embargo of two years or more. Circumstances will vary depending on department and discipline but access to data should be discussed and agreed with supervisors before a thesis is submitted. See information on Restrciting access to your thesis requesting a Moratorium (internal link)

Research publication "Supplementary Information"

In order that potentially valuable data receives proper curation and its own DOI it is recommended that Supplementary Information relating to a research publication is deposited in Pure as a dataset rather than uploaded to a publishers site.

An upload box will appear displaying the defaults for each, the values of the options can be edited here or after the file(s) have finished uploading:

File Name/title

The system will automatically pick up the file name but users may wish to give a more descriptive title to the file in the 'File title' field once uploaded to help others understand what the file is.

File size

Displays the size of the individual file

Visibility

This defines the visibility (i.e. on the research portal) and access that can be applied to individual files within a dataset and should be set appropriately. Select "Public – No restrcitions" to make your data accessible by anyone and "Backend – Restricted to Pure users" option if you want an individual file to remain non-visible to external users (e.g., a DMP associated with the data but not for public consumption). Do not use the "Campus" option as this is **not** currently implemented.

License

Licensing your data clarifies how "open" data is and states the terms and conditions of use, safeguarding against potential legal or ethical disputes resulting from the re-use of data. Legitimate areas where access might be restricted (ethical, legal, commercial, sensitive) should be identified and addressed at the data management planning stage, particularly with collaborative projects where multiple rights and ownership may be an issue. Legal/ethical/commercial/sensitive constraints can be highlighted in the section described below while further details of any restrictions should be outlined in the "Description" field.

Select an appropriate license from the drop down menu provided. Further details of the selected license are available via the “Show license” option. While not a mandatory field yet, assigning an appropriate licence ensures that data owners receive acknowledgement through citation and is therefore ***strongly recommended*** for all files in the dataset.

Non-sharing of data is regarded as an exception and the default position of both RCUK funding bodies and University of Strathclyde is that research data should be made open wherever possible, but in ways that does not harm the research process. As open data is citable and reusable it brings many benefits to individuals, institutions and the broader research community. The optimal combination of fields to support open data and those encouraged by funders would be Access to Dataset: open, no legal/ethical constraints, visibility– public, no restrictions, appropriate credit given, and license set to **CC-BY** on files.

Options currently available in PURE: Creative Commons, Open Data Commons (ODC), Open Data Commons Public Domain Dedication & License (PDDL). See also supplementary guidance from the Digital Curation Centre on How to License Research Data.

Type

This defines the broad classification type for an individual file. Select from the drop-down menu and apply the most appropriate option for the individual file type.

Embargoed end date

Where an embargo on the dataset is required (e.g. for publication or commercial reasons) set the end date of the embargo period in the “Embargoed Until” section. Setting an embargo date does not prevent a DOI from being minted nor the metadata being published it simply prevents the data from being released for download until the specific embargo end date. Data will automatically be made public on the end embargo date.

The properties of all the above fields can be edited once the data has been uploaded by selecting the “edit” option associated with the file.

Physical data

Where any significant elements of the research exist in a physical/non-digital format, please provide as much information as possible in the fields provided

Links

Add a URL to information or resources related to the deposited data (e.g. Project web page, sotware vendor). If data relates to a conference, details of the event should be added in the “Activities” section of “Relations to other content”

Date made available * (Year is mandatory)

This date specifies the date that the dataset record will be made publicly available on the research portal, even if access to the data is restricted.

Access contact details

All dataset records require a current Strathclyde contact. Add contact details of the person who should be contacted should anyone wish to discuss access to or questions arising from the dataset. This contact is for internal purposes and only the generic contact email researchdataproject@strath.ac.uk will appear when this record is displayed on the research portal.

Temporal coverage

This field should only be used if the content of the data has a relevant temporal component, and relates to a specific date or period of time related to the data and not the time-span of the project.

Geo location

State where the data was collected, or to which location the data relates to. This is particularly relevant for field work or where a particular location is of particular relevance to the data (do not enter "University of Strathclyde" simply because the research may have been carried out here).

If the data is available and relevant, choose between a geospatial Point or Polygon.

Legal/ethical

The four constraint options listed here broadly represent the main categories that funders and institutions foresee as preventing data from being successfully shared. As already noted these constraints should be identified and addressed at the data management planning stage where ever possible. Further details on the nature of such constraints and the justification for restricting data should be given in the "description" field, or if necessary in a separate README file supplied with the dataset, especially if the research comprises of data which falls outside the options listed in this section.

Keywords

In addition to title and description, keywords are an essential component in ensuring that your data is found. Along with the content of "title" and "description" the keywords added here will be indexed and searchable, so it is recommended that meaningful and informative keywords which do not appear in those content fields be used to aid classification and discovery.

Relations to other content

Linking data to related publications can provide rich contextual information to support data reuse. Evidence suggests that publications providing access to related data are more likely to be cited. Relate the dataset to relevant projects, equipment, student theses, publications, activities, impacts, or other datasets (including superseded versions) which are ***already recorded in PURE***. Relating different content in PURE increases the usefulness of the research information, and is particularly important for projects and publications which are subject to funder or publisher requirements to make research data available.

For the purposes of individual and institutional funder compliance visibly linking data to publication and funded projects is crucial and is therefore ***strongly recommended.***

Visibility

Set the appropriate level of access to the dataset record here. As mentioned earlier in this document, users can make the **record** publicly available in this section, but access to individual **files** can be restricted in the 'electronic data' section. If access to the record is restricted, this will override any public access settings on the individual data files.

The following visibility settings are available:

Public – this means that the metadata will be made available to the research portal for online release and searching by search engines such as Google. This also allows any associated files marked 'public' or with expired embargoes to be made available online.

Backend – the metadata and any files associated with it (including those marked public or with expired embargo dates) will only be available to those who can log in to PURE (including staff and researchers).

Confidential – able to be seen only by those named on the template (who also have PURE access) and system administrators only. Any associated files will only be visible to the same group of users regardless of the individual visibility setting on those files.

Campus IP - **do not use this option** as we do not implement IP restrictions at Strathclyde.

File Visibility	Record Visibility	Outcome on Portal
Public	Public	Visible, public access to data
Public	Backend	Not visible, no access to data
Backend	Public	Visible, no access to data
Backend	Backend	Not visible, no access to data
Public or Backend	Confidential	Not visible, no access to data
Campus	Campus	DO NOT USE

Status

As with all PURE content types, at the bottom of the screen, users should select the appropriate status of the dataset record and **remember to click 'Save' or all information will be lost!**

When you have completed the metadata fields to your satisfaction set the status to "For validation" which will alert Pure administrators that the record is ready for checking. If you wish to suspend completing the record keep the status at "Entry in progress". Once submitted for validation, a member of the Research Data Management and Sharing team will check the record and contact the depositor if any further information or action is required.

Appendix

Metadata Minimum requirements expected for datasets

Title *

Description – to include:

- name & version of software used to create data

- file format(s) of data uploaded
- outline of criteria for any restrictions on access to data

People * - to include at least one Creator *

Dataset managed by/Managing organisational unit *

Publisher *

Electronic data/”edit” - to include

- visibility
- license

Date made available * (Year mandatory)

Access contact details/Contact person (for internal use)

Relations to other content – Projects and Publications in particular