
# RESEARCH DATA

Prepared by: **Roziana Shamsuri**
Research and Information Services Division,
Perpustakaan Sultan Abdul Samad, Universiti Putra Malaysia

### *What is research data?*

Data can be defined as factual information (such as measurements or statistics) used as a basis for reasoning, discussion, or calculation. It is an information output by a sensing device or organ that includes both useful and irrelevant or redundant information and must be processed to be meaningful. (Merrriam-webster). From an information science perspective, data can be defined more contextually in the scope of research to mean that it is any information that has been collected, observed, generated or created to validate original research findings. Although usually digital, research data also includes non-digital formats such as laboratory notebooks and sketchbooks. (University of Leeds).

Research data does not include incidental or administrative data generated in the course of personal activities, desktop or mailbox backups, or data produced by non-research activities such as University administration or teaching. Research data can be defined by the purpose for which it is used. For instance, the same information might be research data by one researcher but not for another, depending on whether that information is being used as an integral part of a research activity. (University of Bristol).

All research data are valuable. Without it, other researchers can't learn and build upon a research. The files on desktop or USB stick may contain valuable knowledge that other researchers can learn from. (Springer Nature). This is the big reason why we need to manage research data systematically. Managing research data brings many benefits, not only to the project but to future researchers and wider society.

### *Sources of research data*

Research data can be generated from many sources. Basically, there are five sources of research data;

| | |
|---|---|
| **Observational data** is a systematic way to collect data by observing in natural situations or settings. It captured data in real-time and is usually impossible to re-create if lost. Examples of observational data is measurements collected by weather sensors, species abundance surveys, archaeological samples, brain scan images, experience and opinion surveys in the social sciences. | |
| **Experimental data** is data which can be measured or collected through some standard objectives, based on experimental needs. Usually, experimental data is captured from lab equipment. Examples of experimental data are clinical trial data, chemical analyses of physical samples, DNA sequencing of organic material and field trial results. | **Simulation data** is taking a large amount of data and using it to simulate or mirror real-world conditions to either predict a future instance, determine the best course of action or validate a model. For example, climate models and economic models. |
| **Derived or compiled data** has been transformed from pre-existing data points. It is reproducible if lost, but this would be expensive. Examples are data mining, compiled databases, and 3D models. | **Reference or canonical data** is a static or organic conglomeration or collection of smaller (peer-reviewed) datasets, most probably published and curated. For example, gene sequence databanks, chemical structures, or spatial data portals. |

### *Research Data Format*

Research data comes in many varied formats. For examples image, text, audio, containers, databases, geospatial formats, digital posters, presentation formats, web records and many more. It is important to choose the acceptable format for sharing reuse and preservation of data in future.

The formats that are likely to be accessible in the future must be:

- Non proprietary
- Open, documented standards
- In common usage by the research community
- Using standard character encodings (ASCII, UTF-8)
- Uncompressed (desirable, space permitting)

The table below shows types of data and preferred or acceptable file formats for sharing, reuse and preservation.

| | TYPES OF DATA | PREFERRED FILE FORMATS FOR SHARING, REUSE AND PRESERVATION | OTHER ACCEPTABLE FORMATS |
|---|---|---|---|
| 1 | **Quantitative tabular data with extensive metadata**<br><br>A dataset with variable labels, code labels, and defined missing values, in addition to the matrix of data | • SPSS portable format (.por)<br>• delimited text and command ('setup') file<br>• (SPSS, Stata, SAS, etc.) containing metadata information<br>• structured text or mark-up file containing metadata information, e.g. DDI XML file | • MS Access (.mdb/.accdb) |
| 2 | **Quantitative tabular data with minimal metadata**<br><br>A matrix of data with or without column headings or variable names, but no other metadata or labelling | • comma-separated values (CSV) file (.csv)<br>• tab-delimited file (.tab)<br>• including delimited text of given character set with SQL data definition statements where appropriate | • delimited text of given character set -- only characters not present in the data should be used as delimiters (.txt)<br>• widely-used formats, e.g. MS Excel (.xls/.xlsx), MS Access (.mdb/.accdb), dBase (.dbf) and OpenDocument Spreadsheet (.ods) |
| 3 | **Geospatial data**<br><br>Vector and raster data | • ESRI Shapefile<br>• (essential: .shp, .shx, .dbf ; optional: .prj, .sbx, .sbn)<br>• geo-referenced TIFF (.tif, .tfw)<br>• CAD data (.dwg)<br>• tabular GIS attribute data | • ESRI Geodatabase format (.mdb)<br>• MapInfo Interchange Format (.mif) for vector data |
| 4 | **Qualitative data**<br><br>Textual | • eXtensible Mark-up Language (XML) text according to an appropriate Document Type Definition (DTD) or schema (.xml)<br>• Rich Text Format (.rtf)<br>• plain text data, UTF-8 (Unicode; .txt) | • plain text data, ASCII (.txt)<br>• Hypertext Mark-up Language (HTML) (.html)<br>• widely-used proprietary formats, e.g. MS Word (.doc/.docx)<br>• LaTeX (.tex) |
| **5** | **Digital image data** | • TIFF version 6 uncompressed (.tif) | • JPEG (.jpeg, .jpg)<br>• TIFF (other versions; .tif, .tiff)<br>• JPEG 2000 (.jp2)<br>• Adobe Portable Document Format (PDF/A, PDF) (.pdf) |

| | | | |
|---|---|---|---|
| 6 | **Digital audio data** | • Free Lossless Audio Codec (FLAC) (.flac)<br>• Waveform Audio Format (WAV) (.wav)<br>• MPEG-1 Audio Layer 3 (.mp3) - spoken word audio only | • MPEG-1 Audio Layer 3 (.mp3)<br>• Audio Interchange File Format (AIFF) (.aif) |
| 7 | **Digital video data** | • MPEG-4 High Profile (.mp4)<br>• motion JPEG 2000 (.jp2) | • JPEG 2000 (.mj2) |
| 8 | **Documentation & Scripts** | • Rich Text Format (.rtf)<br>• Open Document Text (.odt)<br>• HTML (.htm, .html) | • plain text (.txt)<br>• widely-used proprietary formats, e.g. MS Word (.doc/.docx) or MS Excel (.xls/ .xlsx)<br>• XML marked-up text (.xml) according to an appropriate DTD or schema, e.g. XHMTL 1.0<br>• PDF/A or PDF (.pdf) |
| 9 | **Chemistry data**<br><br>Spectroscopy data and other plots which require the capability of representing contours as well as peak position and intensity | • Convert NMR, IR, Raman, UV and Mass Spectrometry files to JCAMP format for ease in sharing.<br>• JCAMP file viewers: JSpecView, ChemDoodle | |

Source: University of Edinburgh Information Services

### *Why manage research data?*

Research data is important because it provides an evidence of research findings. It is a valuable resource that would have taken a lot of time and money to create. Research data is a vitally important asset and we all have a responsibility to make sure that it is kept safe and used appropriately. There are some reasons why we need to manage research data properly and in a timely manner.

- **Increases transparency**
  Transparency ensures that studies can be reproduced by other researchers in the field, and it helps facilitate proper interpretation and dissemination of results by other stakeholders. Good data management can result in improved research integrity as well as act as validation for research results. Accurate and complete research data are an essential part of the evidence necessary for evaluating and validating research results and for reconstructing the events and processes leading to them.

- **Makes data accessible**
  Making data available enhances the visibility of research outputs and increases the number of citations. Research data, if correctly formatted, described and attributed, will have significant ongoing value and can continue to have impact long after the completion of a research project.

- **Reduces the risk of data loss**
  The risk of data loss through accidents or neglect can be reduce by keeping research data safe and secure (use of robust and appropriate data storage facilities). The right place for research data is likely to be institution's own data repository or possibly a disciplinary repository.

- **Facilitates future reuse and sharing**
  As research becomes increasingly more complex, researchers can provide opportunities for collaboration with other researchers within discipline, or even with other disciplines, by facilitating the sharing and reuse of research data for future research. Sharing research data and enabling others to use it will also help to prevent duplication of effort.

- **Improves citations**
  Researcher profile can be enhancing by gaining credit for the data produce and increase readership of published work, including the papers and articles which the data builds on.

- **Meet publishers' requirement**
  Journal publishers increasingly require data that form the basis for publications to be shared or deposited in an accessible data center or repository. This requirement applies to both commercially and publicly-funded research.

- **Meet funders and institutions requirement**
  Funding bodies and institutions are taking more interest in what researchers do with the data that is generated in the course of a project. If the research project is supported by industrial and commercial partners it is likely they will have their own data management or sharing policy.

**References:**

Data. Merriam-Webster.com Dictionary, Merriam-Webster. Retrieved June 15, 2022, from https://www.merriam-webster.com/dictionary/data

Datamation. What is data simulation? Retrieved June 15, 2022, from https://www.datamation.com/big-data/data-simulation/

Dewitt Wallace Library. Research guide. Retrieved June 15, 2022, from https://libguides.macalester.edu/c.php?g=527786&p=3608583

How and why you should manage your research data: a guide for researchers. Retrieved June 15, 2022, from https://www.jisc.ac.uk/guides/how-and-why-you-should-manage-your-research-data

Springer Nature. Research data. Retrieved June 15, 2022, from https://www.springernature.com/gp/authors/research-data#:~:text=Research%20data%20refers%20to%20the,and%20build%20upon%20your%20res earch.

University of Bristol. Research data service. Retrieved June 15, 2022, from https://data.blogs.bristol.ac.uk/bootcamp/data/

University of Edinburgh. Information services. Retrieved June 15, 2022, from https://www.ed.ac.uk/information-services/research-support/research-data-service

University of Leeds. Research data management explained. Retrieved June 15, 2022, from https://library.leeds.ac.uk/info/14062/research_data_management/61/research_data_management_explained