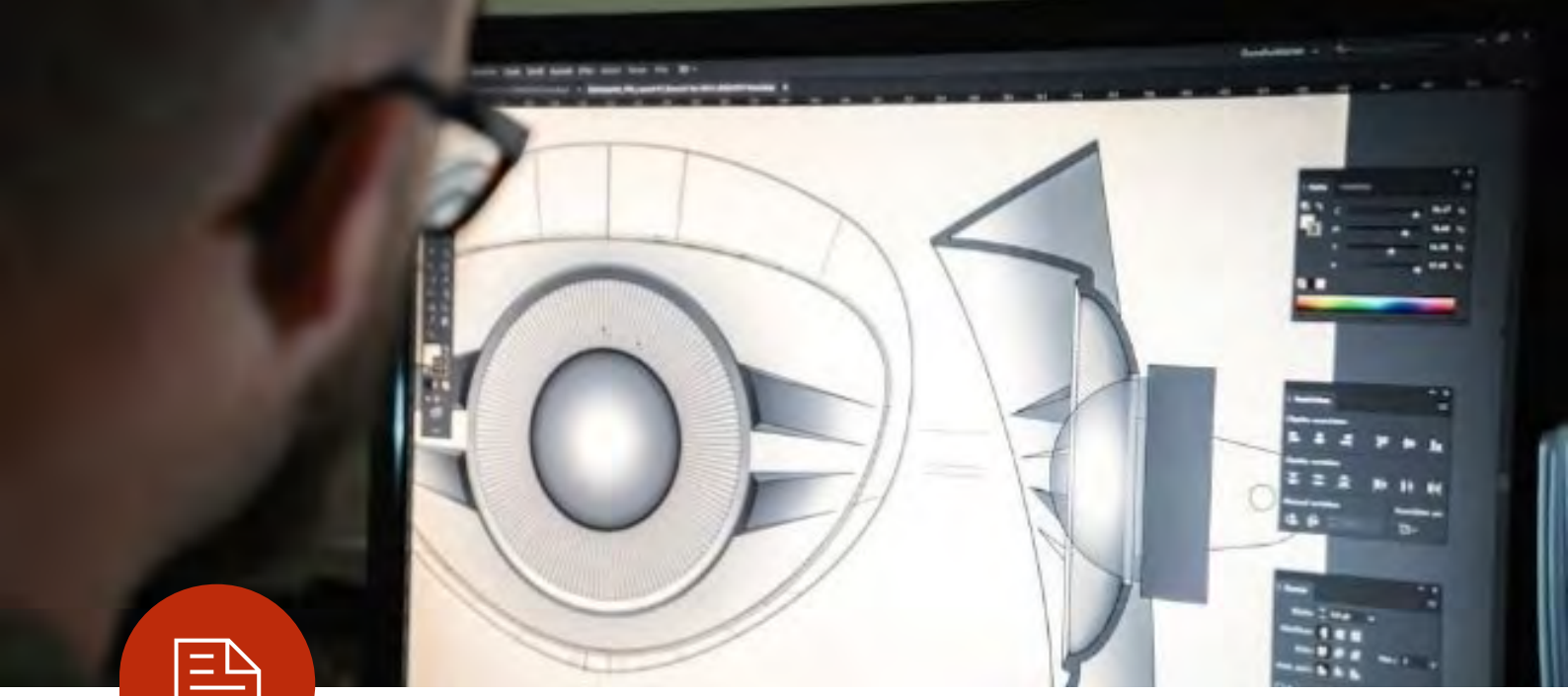
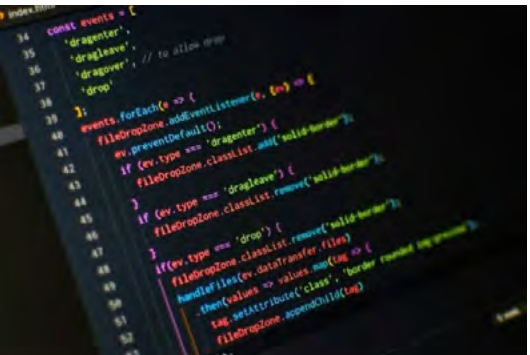


Research data: types, formats, methods

Research data are factual records collected, generated or reused as a basis for analysis, reasoning, discussion or calculation.

Data can be classified in different ways. Examples include observations, experiences, published and unpublished sources, bibliographic references, text, images, all created and/or collected in digital form, as well as other digital outputs of research, such as 3D models and source code.

Data types also vary according to the research **domain.**



In the field!

My research is theoretical in nature, and I do not produce any data. *Do these Guidelines apply to me too?*

Yes. All type of research produces or reuses data (in a broad sense), even though every disciplinary domain has its own specificities. You almost certainly use primary or secondary sources to answer your research questions and, in doing so, you produce a collection of bibliographic metadata, organised more or less systematically. In this context, always remember to use the Persistent Identifiers (PIDs) of the resources you cite and think about how you can exploit this output, for example by publishing it online as open data.

Types of data: how to categorise them

Knowing and classifying your research data allows you to choose the most effective strategies to manage them responsibly, avoid data loss or corruption, and select the most appropriate methods for data collection, archiving and analysis.

Research data can be **digital, digitised or non-digital.** While digital or digitised data management necessarily follows computerised protocols, non-digital data can be managed both digitally and non-digitally.

Whether digital or non-digital, data can be described based on its **content** – numerical, textual, audio, video, etc.

Dati con lo stesso contenuto possono avere forme diverse e quindi la loro struttura dal punto di vista digitale può cambiare. Ad esempio, dati testuali possono essere raccolti tanto nella forma di fogli di calcolo quanto nella forma di documenti di testo.

- Dati con lo stesso contenuto e raccolti nella stessa forma possono avere formati (e quindi estensioni) differenti. Ad esempio, dati numerici possono essere raccolti in un foglio di calcolo che può essere scritto in formato file "comma-separated values" (CSV) con estensione del file .csv, così come in formato OpenDocument Spreadsheet (ODS), con estensione del file `.ods`, o ancora in formato Microsoft Excel, con estensione del file `.xls` o `.xlsx`.

To ensure that your data remains accessible and reusable, it is recommended that you collect, save, share and deposit it in open, non-proprietary formats, rather than in closed, proprietary ones.

- **Proprietary format**: a format that is owned and developed by a specific company or organisation.
  - **Proprietary and closed format**: the developer of the format decides what software can use that format. For example, `.indd` for Adobe InDesign files, a software produced by Adobe for the publishing industry.
  - **Proprietary and open format**: the developer of the format has not restricted its use to a certain software. For example, MP3 exists as an open format for audio files but is patented in some countries. The XLS format was once a closed format – i.e. it could only be run by Microsoft's proprietary software, Microsoft Excel – but has since been opened. The same goes for the `.xlsx` format that is based on XML (open format) and can also be used by other software, such as LibreOffice Calc.
- **Non-proprietary and open format**: the format specifications are openly available, and anyone can create software for it. For example, CSV files for tabular data can be opened by a variety of different software.

Examples of open formats for the most common data types include:

- **Quantitative and qualitative tabular data**: SPSS (`.sav`), Stata (`.dta`), CSV (`.csv`).
- **Geospatial, vector and raster data**: ESRI Shapefile (essential – `.shp`, `.shx`, `.dbf`, optional – `.prj`, `.sbx`, `.sbn`), Geo-referenced TIFF (`.tif`, `.tfw`), CAD data (`.dwg`), e Tabular GIS attribute data.
- **Qualitative textual data**: eXtensible Mark-up Language (XML), Rich Text Format (`.rtf`), Plain text data, ASCII (`.txt`).
- **Images, audio and video**: TIFF (`.tif`, `.tiff`), JPEG (`.jpeg`, `.jpg`), Adobe Portable Document Format

(PDF/A, PDF) (`.pdf`), PNG (`.png`), Free Lossless Audio Codec (FLAC) (`.flac`), MPEG-1 Audio Layer 3 (`.mp3`), Audio Interchange File Format (`.aif`), Waveform Audio Format (`.wav`), MPEG-4 (`.mp4`), MOV (`.mov`), Windows Media Video (WMV) (`.wmv`).

In the field!

I am a researcher working with data organised in tables. *What are the most used formats?*

The most used formats for tabular data are:

- 'Comma-Separated Values' (CSV, `.csv`): a non-proprietary, textual format in which data is usually separated by commas.
- 'OpenDocument Spreadsheet' (ODS, `.ods`): an open standard format for spreadsheets, which stores data in cells arranged into rows and columns. Also, .ods files can be opened in Microsoft Excel and saved as XLS or XLSX files.
- 'Excel Workbook' (XLS/XLSX, `.xls`/`.xlsx`): the Excel format is a proprietary yet very common format that allows users to create, handle and analyse tabular data in a spreadsheet.

For my research, I need to collect data through surveys. *What tool can I use?*

Surveys can be conducted through interviews or questionnaires in person, over the phone or online.

Depending on the population you want to sample, the size of the sample itself and the sample design – which can be simple or complex, longitudinal or cross-sectional – it may be necessary to accompany these techniques and tools with the support services offered for data management, privacy and/or ethics.

Examples of online survey tools include Microsoft Forms, Google Forms, LimeSurvey, SurveyMonkey, Qualtrics. If you collect personal data, you must use a tool such as Microsoft Form, supplied by the University, or check any licences available through your Department (LimeSurvey, SurveyMonkey, Qualtrics), rather than using personal licences.

In the case of a cross-sectional survey for which you won't need to contact the same person twice (or several times), you can choose to implement privacy-by-design techniques to anonymise data at the source, thus avoiding privacy issues.

For my research, I work with biomedical imaging data. *How do I choose the format for saving and archiving it?*

Digital Imaging and Communications in Medicine (DICOM) is the standard for transmission and management of medical images and related information. In addition to the image, a DICOM file includes a heading that contains all the metadata acquired with the image itself (patient data, tumour location, duration and amount of radiation, etc.).

TIFF is another appropriate format to store and share medical images – it is a raster graphic file format that supports lossless compression and, as such, is suitable for archiving and printing high-resolution images and photos. All the relevant metadata can be saved in a separate TXT file.

On data collection and methodologies

In research practice, data collected or generated by third parties can be reused instead of, or in addition to, **generating new data**.

Provided the data is of good quality, **reusing existing data** saves time and resources. **Online digital archives for long-term data preservation**, sometimes specific to a certain disciplinary domain, can be accessed to browse and download relevant data **Repositories**.

Before reusing data, regardless of their origin, you need to make sure you are legally and contractually able to do so **Copyright** **Respecting privacy**.

Generare o raccogliere i dati può comportare pratiche molto diverse tra loro. Ad esempio, i dati possono essere di natura **sperimentale**, quando ottenuti tramite esperimenti e dimostrazioni che seguono un metodo scientifico. Oppure possono essere di natura **osservativa**, quando vengono raccolti attraverso l'osservazione critica, con l'eventuale aiuto di strumenti. Quando la ricerca è **compilativa**, i dati vengono raccolti in forma derivata/compilata da altre fonti.

Indipendentemente dalle pratiche di generazione o raccolta dati, gli **strumenti**, **software e metodi utilizzati devono essere registrati per consentire la riproducibilità della ricerca** **Gestire il software**.

Inoltre, sempre a prescindere dai metodi di raccolta o generazione dei dati, è necessario assicurarsi di essere conformi alle normative sulla privacy e sull'etica.

Se hai intenzione di sfruttare commercialmente i tuoi dati, perché possono essere utili, per esempio, per depositare una domanda di brevetto, pianifica in anticipo delle strategie di gestione dei dati che possano garantirti adeguata protezione.

In the field!

I have developed a software for analysing and displaying the results of my research.
Do I have to manage it in the same way as research data?

Yes, it is advisable that you plan software development and use tools to document it, as this allows you to exploit the software itself as an asset and the main output of your research, as well as facilitating reuse in future research.

Some tools, such as cloud notebooks, can help you document code development and every step of its algorithm. By running your code in cloud, you can view how every single part is run and the corresponding input and output data.

Once your code reaches a stable executable version, it is recommended that your deposit it in a disciplinary repository together with suitable documentation and specific metadata, to ensure that it is preserved in the long term. An example of disciplinary repository for source code is Software Heritage, which uses CodeMeta as metadata schema, and regularly and automatically harvests the most common forges for development, such as GitHub. **Managing software** **Repositories**.

I work with cultural heritage and my data is mostly text and images, often from archives, museums and libraries.
What do I do?
Contact the institution that holds the sources you wish to use in your work to find out what you need to do. Even though they are no longer copyrighted, they may still be protected as cultural heritage, and you may need permission to reproduce them. If the sources you work with are still in copyright, you will need permission from the rights holders **Copyright**.

Useful links

More on formats: https://www.loc.gov/preservation/resources/rfs/TOC.html | https://www.dicomstandard.org/

https://ukdataservice.ac.uk/learning-hub/research-data-management/format-your-data/recommended-formats/

Useful tools for interviews:

https://forms.office.com | https://www.qualtrics.com/it/ | https://www.limesurvey.org/it

Useful tools for software:

https://datasciencenotebook.org/ | https://www.softwareheritage.org/ | https://codemeta.github.io/codemeta-generator/ | https://github.com/