|  |  |  |
| --- | --- | --- |
| **Term** | **Explanation** | **More information** |
| Concepts | | |
| open source | Computer code where the source code is available for inspection, and may be freely re-used and distributed. R, Python and GNU/Linux are all open source. | https://opensource.org/osd |
| open access | Access to research products, such as publications and datasets, without financial or copyright barriers, but such that authors have control over the integrity of their work and the right to be acknowledged and cited. One approach is to publish in open access journals, such as PLOS ONE, another approach is to submit manuscripts of published papers to institutional repositories where they are freely available to the public. | http://www.budapestopenaccessinitiative.org/read |
| reproducibility | A study is reproducible if there is a specific set of computational functions/analyses (usually specified in terms of code) that exactly reproduce all of the numbers and data visualizations in a published paper from raw data. Reproducibility does not require independent data collection and instead uses the methods and data collected by the original investigator. | https://osf.io/s9tya/ |
| replicability | A study is replicated when another researcher independently implements the same methods of data collection and analysis with a new data set. | http://languagelog.ldc.upenn.edu/nll/?p=21956 |
| provenance | The origin of data and code, including any transformations occurring along the way. |  |
| File formats | | |
| CSV | A common file format for collecting, sharing and archiving tabular data. This is a plain text file where variables (columns) are separated by commas. Thus the name, 'comma separated variables', it is closely related to TSV, 'tab separated variables' | http://www.digitalpreservation.gov/formats/fdd/fdd000323.shtml |
| plain text | A file that contains simple text characters and no formatting (e.g. margins) or embedded images. Use of plain text files is not dependent on specific programs, so they can be created, read, and edited by almost any program, regardless of operating system and computer architecture. Using plain text formats allows a high degree of interoperability between computational environments, and ensures that your files can be read by other people with minimum effort. Most programming script files are plain text files. | http://www.linfo.org/plain\_text.html |
| binary | A file that must be interpreted by a specific program before it is human-readable and editable. For example, PDF, Microsoft Word doc and Excel xls files are binary files, and can only be read and edited by those programs. Many commercial programs use proprietary binary file formats. This limits their interoperability and archival value. Images, video and audio files are also binary files. |  |
| Licenses for data and code | | |
| CC0 | Public domain, no rights reserved. This license allows for the greatest freedom for reuse. Used for data by major online repositories such as Dryad, Figshare, Zenodo. Good scientific practices assure proper credit is given via citation, which enforced through peer review. Marking data with CC0 sends a clear signal of zero barriers to reuse. | https://creativecommons.org/licenses/ |
| CC-BY | Allows for reuse only if attribution is given to the author, in the manner specified by the author. Often used for copyrightable materials such as journal articles in open access publications, for example PLOS ONE, BioMed Central, and Nature Communications. |  |
| CC-NC | Allows for reuse only for non-commercial purposes (for example, a Cultural Heritage Management business would not be allowed to use CC-NC data or code). Not recommended for most research output. |  |
| MIT | A license especially for software that places very few restrictions on the use of the software, and disclaims the author of any responsibility for problems arising from others using the software. It is one of the most popular licenses for open source software. | http://opensource.org/licenses/MIT |
| Data archiving | | |
| DOI | DOI stands for 'digital object identifier', a persistent (but not permanent) label that stores information about the online location of a electronic file. A DOI also includes metadata, for example in the case of journal article it might include the author, title, date of publication, etc. The online location and metadata of a file may change, but its DOI remains fixed. This means that a DOI is generally a more reliable link to an online document than a URL. | http://www.doi.org/ |
| figshare | A commercial online digital repository where research output can be freely archived and openly accessed. Issues DOIs for individual files or groups of files. | http://figshare.com/ |
| zenodo | Similar to figshare, but a non-profit service operated by European Organization for Nuclear Research (known as CERN) | https://zenodo.org/ |
| tDAR | The Digital Archaeological Record (tDAR) is a digital repository for the digital records of archaeological investigations. Fees are charged for archiving files, but access to open files is free. | https://www.tdar.org/ |
| Open Context | A data publishing and archiving service. It is aimed at maximizing the integration of data with other services (such as maps, media, and other data sets). Similar to tDAR, there are fees to upload but accessing open data is free. | http://opencontext.org/ |
| Archaeological Data Service | An open data repository focused on output from research and commercial archaeology in the UK. There are fees to upload but accessing open data is free. | http://archaeologydataservice.ac.uk/ |
| CLOCKSS | A not-for-profit joint venture between several academic publishers and research libraries to build a sustainable, geographically distributed dark archive with which to ensure the long-term survival of Web-based scholarly publications. | https://www.clockss.org/ |
| Document markup languages | | |
| markdown | A simple, minimal language for formatting plain text files so that they can be converted into richly formatted HTML, PDF and Microsoft Word documents. Scholarly requirements such as citations, captions and cross-referencing can be enabled with a small amount of HTML or LaTeX and use of Pandoc. | http://daringfireball.net/projects/markdown/syntax |
| R markdown | A variant of markdown that extends it to allow chunks of R code to be embedded among the text. This results in a simple system for literate programming. For example, an R markdown document might have several paragraphs of text, then a chunk of R code that generates a figure, then several more paragraphs of text. Suitable for journal-article-length documents that include narrative text and output from statistical analysis. | http://rmarkdown.rstudio.com/ |
| LaTeX | A complex document preparation system optimized for producing technical and scientific documentation. Suitable for large multi-part documents such as complex journal articles, books and theses. Literate programming with R code interwoven among text is enabled via the knitr package. | https://latex-project.org |
| pandoc | An open source program for converting documents between a very wide variety of formats. Often used to convert markdown, R markdown and LaTeX documents to HTML (for web publication), PDF and Microsoft Word documents. It is built into RStudio. | http://pandoc.org/ |
| Scientific programming | | |
| script | A plain text file containing instructions for a computer written in a programming language, for example in R or Python |  |
| R | A free and open source programming language with strengths in data analysis and visualization. Most effective when used in combination with RStudio, a free and open source integrated development environment for R. | https://www.r-project.org/ |
| Python | A free and open source programming language with a reputation for ease of use and being suitable for a wide range of scientific and commercial applications. | https://www.python.org/ |
| MATLAB | A commercial programming language known for numerical and symbolic computing capabilities. The algorithms are proprietary, which means you cannot easily see the code of the algorithms and have to trust that MATLAB implemented it correctly. The proprietary nature also makes it hard, if not impossible, for others to extend or create tools for MATLAB. | http://www.mathworks.com/products/matlab |
| Version control | |  |
| Git | Open source software for version control and collaboration. It can handle any file type, but is most effective on plain text files such as scripts and markdown/LaTeX documents. | https://git-scm.com/ |
| GitHub | A popular commercial web service that provides collaboration tools and free public hosting of files in git repositories. Private repositories are available for a fee. Similar services include GitLab and Bitbucket, both of which have the advantage of unlimited free private repositories. | https://github.com/ |
| commit | A Git command to record changes in files to the Git repository. A sequence of commits creates a history of how the files have changed during your work on them. | http://git-scm.com/book/en/v2/Git-Basics-Recording-Changes-to-the-Repository |
| Computational environments | | |
| virtual machine | The use of software to emulate an entire operating system (such as GNU/Linux, Microsoft Windows or Apple OS X) within another computer. For example, you might use a virtual machine to use a GNU/Linux operating system on a laptop where the main operating system is Microsoft Windows. Virtual machines are convenient for reproducing an entire computational environment, but they can consume a lot of hard disk space which makes sharing and archiving challenging. |  |
| GNU/Linux | A free and open source computer operating system (i.e., an alternative to Microsoft Windows and Apple OS X). Commonly used for scientific computing, internet servers, supercomputers and Android phones and tablets. Popular distributions of GNU/Linux in academia include Ubuntu and Debian. | http://www.linux.org/ |
| Linux container | A system for running multiple isolated Linux systems (containers) on a single Linux control host. Isolation means that the dependencies can be well understood and documented. In a research context, containers are useful for encapsulating the all of the diverse components of a complex data analysis system. Containers take up less disk space than a virtual machine, and so are more efficient for sharing and archiving. | https://linuxcontainers.org/ |
| Docker | A free and open source system that simplifies the creation, use, sharing and archiving of Linux containers. In a research context Docker makes it easy to document and share computational environments so you can ensure that others have exactly the same software versions as you used. | https://www.docker.com/ |
| Communities | | |
| Software Carpentry | An international non-profit volunteer organization focusing on teaching researchers basic software skills. Prioritizes the use of free and open source software tools, encourages researchers to use permissive licenses for their research products. Target audience is novices with little or no prior computational experience. | http://software-carpentry.org/ |
| Data Carpentry | Similar to Software Carpentry, but focuses more on domain-specific training covering the full lifecycle of data-driven research. | http://www.datacarpentry.org/ |
| rOpenSci | A collaboration of volunteers from academia and industry developing R-based tools for making scientific research, data and publication freely accessible to the public. They also conduct workshops to train researchers to use R and related tools. | https://ropensci.org/ |