

A Selection of Research Data Management Tools Throughout the Data Lifecycle

by

Jan KRAUSE

September 15, 2015



Contents

1	Abstract	4
2	Data lifecycle	5
2.1	Data lifecycle management phases	5
2.2	Data management planning tools	5
2.3	Data policies	6
3	Data management tools	6
3.1	Data discovery tools	6
3.1.1	Data repositories	6
3.1.2	Data papers and data journals	7
3.2	Data acquisition, format and description	7
3.2.1	Data acquisition	8
3.2.2	Data formats	8
3.2.3	Metadata formats	9
3.3	Data analysis tools	10
3.3.1	Statistical data analysis tools	10
3.3.2	Scientific and general data analysis tools	10
3.3.3	Mathematical symbolic analysis tools	11
3.4	Active data collaboration tools	11
3.4.1	Versioning	11
3.4.2	File sharing	11
3.4.3	Graphs	12
3.4.4	Geodata	12
3.4.5	3D	12
3.4.6	Papers and annotations	13
3.4.7	Reference management and sharing	13
3.4.8	Scientific workflows and laboratory management	14
3.5	Data publication tools	14
3.5.1	Data repositories	14
3.5.2	Interactive publications	14
3.5.3	Data licenses and Laws	15

Index

- Data Formats
 - Data Type Registry, 8
 - HDF5, 8
 - Naming Convention, 9
 - Recommended Data Formats, 8
 - Semantic Web (RDF), 9
 - Structured Query Language (SQL), 8
- Data Journals
 - Archaeology Data Journal, 7
 - Biodiversity Data Journal, 7
 - Data Papers, 7
 - Dryad's Data Journals List, 7
 - Earth System Science Data, 7
 - Geoscience Data Journal, 7
 - GigaScience, 7
 - Nature's Scientific Data, 7
 - Open Health Data, 7
 - Trac's atadData Journals List, 7
- Data Management Plan
 - Howto, 6
 - Online Tool, 6
- Data Management Policies
 - Humboldt University Berlin, 6
 - University of Bath, 6
 - University of Edinburgh, 6
 - University of Oxford, 6
 - University of Southampton, 6
- Data Repositories
 - Dryad, 7
 - Figshare, 7
 - Re3data, 6
 - Recommended Repositories, 6
 - Zenodo, 7
- EPFL
 - AiiDA, 14
 - Git, 11
 - Institutional File Share, 11
 - OwnCloud, 12
- Laws
 - CH, 15
 - EU, 15
- Licences
 - Creative Commons By, 15
 - Creative Commons Zero, 16
 - GNU Affero General Public Licence (AGPL), 17
 - GNU General Public Licence (GPL), 16
 - GNU Lesser General Public Licence (LGPL), 17
 - Open Data Commons (PDDL), 17
 - Open Data Commons Attribution License, 17
 - Open Data Commons Open Database License (ODbL), 17
- Metadata Formats
 - DataCite, 9
 - Directory, 9
 - DublinCore, 9
 - Semantic Web (OWL), 9
 - Textual description, 10
 - UML, 9
- Recommended
 - Data Formats, 8
 - Data Repositories, 6
 - HDF5, 8
 - Metadata Formats, 9
- Software
 - AiiDA, 14
 - Atlas, 13
 - Authorea, 13
 - Dia, 12
 - Git, 11
 - Git-annex, 12
 - Git-annex Assistant, 12
 - GitHUB, 11
 - Graphviz, 12

Hypothesis, 13
iPython Notebooks, 14
Kepler, 14
LIMS, 14
Maple, 11
Mathematica, 11
Matlab, 10
Mendeley, 13
MeshLab, 12
MyExperiment, 14
NumPy, 10
Octave, 10
OpenBIS, 14
OverLeaf, 13
OwnCloud, 11, 13
ParaView, 12
Pegasus, 14
Plotly, 12
PSPP, 10
Python, 10
QGIS, 12
R Project, 10
RopenSci, 8
Science Exchange, 8
SLIMS, 14
SPSS, 10
Taverna, 14
Zotero, 13

1 Abstract

In this document, several useful **research data management tools are listed and described** for each step of their research throughout the data lifecycle management.

This document is mostly generic from an institutional point of view, however it offers specific information intended for EPFL and Swiss researchers for some targeted points.

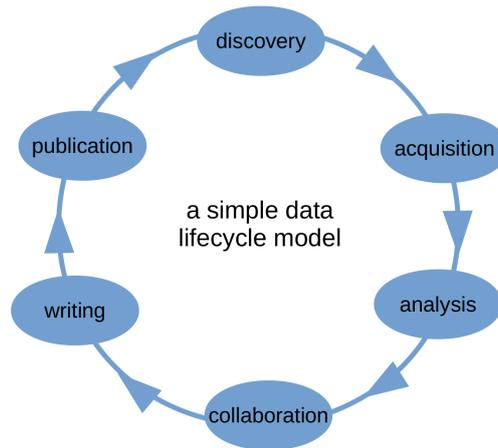
This selection aims to help researchers **make the most out of their data**, and especially:

- **save time** in the long run
- **collaborate efficiently** on their data
- promote **reproducible research**
- enhance the **visibility of their work**
- meet **fundings' data requirements** (Horizon 2020, SNF...)
- meet **publishers' data requirements** (Nature Publishing Group, PLoS...)
- **minimize the risks** around their data (such as data loss or corruption, data leak, etc.)
- **open or secure** their intellectual property privacy

2 Data lifecycle

2.1 Data lifecycle management phases

Many tools are available to help manage data throughout the research process; these can be categorized using data lifecycle management phases (illustrated below):



1. **discovery:** find useful, technically and legally reusable datasets
2. **acquisition:** maximize the potential of your data, its visibility, and reproducibility by choosing from the beginning appropriated data and metadata formats. Anticipating will be crucial, knowing that afterwards it might be too late to do so.
3. **analysis:** analyze the data
4. **collaboration:** share your data and collaborate with your colleagues and partners (locally or worldwide)
5. **writing:** use comprehensive and collaborative paper writing tools
6. **publication:** deposit your datasets in visible and trusted data repositories

2.2 Data management planning tools

Data management plans (DMP) are recognized as important means to produce data of good quality. DMPs are generally conceived from the beginning of research projects and updated throughout their duration. They aim to anticipate all data-related needs (such as storage capacities, collaborative tools, data licenses, data formats, metadata or description, sensitive data

anonymization, etc.) and avoid any problems (lacunar data, data loss, data corruption, intellectual properties issues, storage, etc.).

DMP ONLINE (UK) and DMP TOOL (US) are free online tools assisting in the conception of DMPs. Both will guide the user through the essential questions that must be answered to manage your data. In addition, various funders' requirements are built in and can be used out of the box (e.g. DMP Online supports the Horizon 2020 guidelines)[13, 14].

HOW TO DEVELOP A DATA MANAGEMENT is a more traditional guide describing the elaboration of a DMP [6].

GUIDELINES ON DATA MANAGEMENT IN HORIZON 2020 are the official requirements to follow when applying to a Horizon 2020 project subject to the data pilot program [7].

2.3 Data policies

The following institutional data policies, constitute together a good overview on this topic in Europe: University of Edinburgh, University of Oxford, Humbolt University Berlin, University of Southampton and University of Manchester [19, 58, 5, 42, 73].

3 Data management tools

In this section data management tools are listed along each step of the data lifecycle.

3.1 Data discovery tools

3.1.1 Data repositories

RE3DATA.ORG is a registry of data repositories. This tool indexes over a thousand archives which are both subject specific and generalist and can be browsed by disciplines, available repositories features such as persistent identifiers support (e.g. DOIs, which play a crucial for guarantying access to datasets, as web links tend to break after a few years), and other important information such as data licenses availability, standards and policies [70].

NATURE'S RECOMMENDED DATA REPOSITORIES is a set of disciplinary repositories covering the following fields: Biological sciences; Health sciences; Chemistry; Earth and environmental sciences; Physics, astrophysics & astronomy; Social sciences; and General science [36].

ZENODO [90], DRYAD [16] and FIGSHARE [23] are state of the art general-purpose data repositories. Zenodo is made available by Cern and OpenAire and is free for any researcher publishing his or her data openly. Dryad is a curated repository, maintained by a non profit organization. Figshare belongs to a for profit company, the MacMillan group, which also owns the Nature Publishing Group.

3.1.2 Data papers and data journals

Data papers are publications describing datasets. In other words, they constitute peer-reviewed searchable metadata, and they can be used to find or highlight datasets. Data papers can be found in pure data journals, or in journals mixed with traditional scholarly publications. An important point, is that these papers may be found through classical scholarly search engines. In addition, the following resources can help you find multi-disciplinary data-journals and data papers:

- Dryad’s examples of journal data policies lists journals that require data archiving and journals with data policies [17]
- Trac’s multidisciplinary data journals list[76]
- Nature Publishing Group’s scientific data website [37]
- DataShare’s sources of dataset peer review list of data journals [11]
- GigaScience [28]

Some discipline specific data journals exist too, for example:

- Wiley’s Geoscience Data Journal and Earth System Science Data [9]
- UpMetaJournal’s Open Health Data [79]
- Pensoft’s Biodiversity Data Journal [61]
- UpMetaJournal’s Journal of open archaeology data [78]

In addition, a list of JOURNALS DATA POLICIES compiled by Dryad may be of interest [15].

3.2 Data acquisition, format and description

In order to make the most out of your research data, it is important to use appropriated data and metadata standards. This has to be thought through at the beginning of the project and in any case before data acquisition. Indeed, it is often to late to correct, complete or correct data after the project is well started. Using good data and metadata standard help collecting

coherent data and avoid missing some points. In addition, badly described data will not be re-usable by others at all, nor will it be easy to find. A standard and open data format will allow more people to access and re-use a dataset because of its good compatibility across software and platforms. Finally, open standards will maximize the chances to access your results in the future because they are supported longer.

3.2.1 Data acquisition

SCIENCEEXCHANGE is a platform where scientists can order data for specific experiments (including from their own design): it is an on-line scientific experiment marketplace [72].

ROPENSCI offers packages providing an easy access to data repositories through the R statistical programming environment. R is a free software available on all platforms (Windows, Mac, Linux). These packages cover data access in the following domains: primary data, full-text of journal articles, altmetrics, data-publication, reproducibility and data visualization. Many other data analysis R packages are available through CRAN [71, 65].

3.2.2 Data formats

A directory of RECOMMENDED DATA FORMATS is maintained by the US Library of Congress. It covers the following categories: still images, sounds, moving images, textual documents, web archives, datasets, geospatial data as well as generic data [41].

The DATATYPERegistry is a generic open source data type description platform. It allows in particular to combine already described units or data types to create new ones. Data types are labeled with unique identifiers. In addition to a web interface an automated access is allowed through the API [12].

HDF5 “is a data model, library, and file format for storing and managing data. It supports an unlimited variety of datatypes, and is designed for flexible and efficient I/O and for high volume and complex data. HDF5 is portable and is extensible, allowing applications to evolve in their use of HDF5. The HDF5 Technology suite includes tools and applications for managing, manipulating, viewing, and analyzing data in the HDF5 format” [35]. HDF5 is supported form within many environments such as Matlab, Octave, Python (H5py, PyTables), GNU-R, Java, C++, Fortran and Mathematica.

The STRUCTURED QUERY LANGUAGE (SQL) “a special-purpose programming language designed for managing data held in a relational database management system (RDBMS)” [87]. SQL is well suited to store and share

relational data. RDBMS are a great help in maintaining dataset's coherence and enforce data constraints. Several multi-platform open source RDBMS are available, such as MariaDB (MySQL) [44] or PostgreSQL [64].

The SEMANTIC WEB Resource Description Format (RDF) “is a standard model data interchange on the Web”[81]. RDF is a very general format and may be used to encode most types of data. A significant advantage of RDF over other data formats resides in its interoperability capabilities with other datasources, allowing to extend analyses beyond given datasets by connecting them to other data sources. In practice, this is done using the SPARQL query language on triple stores such as Virtuoso, Jena or 4store.

Sometimes, a simple but clear datasets NAMING CONVENTION can help a lot. Consider e.g. the pattern: Project-Institution-Group-DataSetName-Version-YYYYMMDD.DataFormat, which could in practice become something like FictiveProject-UniversityX-SignalProcessingLab-SolarIntensity-Version2.3-20151131.csv .

3.2.3 Metadata formats

DUBLINCORE is a vocabulary consisting of only 15 basic elements, such as Creator, Title, Date, Description, Format, Rights, or Subject [18]. It is not specifically designed for dataset description, but widely used in scholarly communication. For that reason, it is a minimalist solution, and we recommend one of the solutions listed below instead.

DATA CITE METADATA SCHEMA is a standard designed with datasets in mind, and hence more adapted than DublinCore mentioned just above. For example, GeoLocation, ResearchGroups, Collections, Videos or Workflows have their own specific resource types [10].

A directory of RECOMMENDED METADATA FORMATS is available in this open source collaborative platform. They are indexed by discipline, extensions, associated tools and associated use cases. General available subject categories are Art and Humanities, Engineering, Life Sciences, Physical Sciences and Mathematics, Social and Behavioral Sciences and General Research Data. Dozens of subcategories are also available [48].

The SEMANTIC WEB Resource Description Format (RDF) “is a standard model data interchange on the Web”[81]. RDF is a very general format and may be used to encode most types of data. Indeed, many OWL [80] RDF based ontologies exist. Ontologies are “formal naming and definition of types, properties and interrelationship” of data [86].

UNIFIED MODELING LANGUAGE (UML) is a complete modeling language. The class diagrams are particularly useful to describe data structures [77].

Sometimes, a simple but clear TEXTUAL DESCRIPTION of a dataset can help a lot (even its own creator in the future). For instance, if it is not explicit in the dataset, it is a means to describe how, when, where and with what device the data has been gathered. In addition, the meaning of the labels (e.g. the column headers) are generally of interest, and so are the physical units and their accuracy.

3.3 Data analysis tools

3.3.1 Statistical data analysis tools

The R STATISTICAL PROGRAMMING ENVIRONMENT is a free software available on all platforms (Windows, Mac, Linux). In addition, RopenSci offers packages allowing an easy access to data repositories through R. These packages cover data access in the following domains: primary data, full-text of journal articles, altmetrics, data-publication, reproducibility and data visualization. Many other data analysis R packages are available through CRAN [71, 65].

SPSS is a tool for statistical analysis, popular in social sciences. It is a proprietary software available on all platforms (Windows, Mac, Linux). IBM SPSS Modeler is a companion product for data mining. PSPP is an open source multi-platform SPSS clone. Many SPSS features are supported by PSPP [39, 33].

3.3.2 Scientific and general data analysis tools

MATLAB is an high-level interactive environment specializing in engineering and scientific computations. Matlab is a proprietary software available on all platforms (Windows, Mac, Linux). It covers in particular signal and image processing, communications, and control systems. GNU OCTAVE is an open source multi-platform Matlab clone. Many Matlab features are supported by Octave [32, 45].

NUMPY is a scientific package for computing with Python. It is a free software available on all platforms (Windows, Mac, Linux). It may be completed with many other scientific packages that may be found in the Python Package Index (PyPI), PyData and/or SciPy. For example: Matplotlib (plotting), Sympy (symbolic mathematics), Pandas (data structures and analysis) [51, 24, 68].

3.3.3 Mathematical symbolic analysis tools

MATHEMATICA and MAPLE are both mathematics, scientific and engineering oriented computing environments. They are based on symbolic mathematics and their interfaces take the form of interactive documents. Both are proprietary multi-platform software [43, 89].

3.4 Active data collaboration tools

3.4.1 Versioning

Versioning is a crucial tool for reproducible research, especially in the context of data and computer code: reproducing results generally requires at least to be able to combine the exact version of a dataset and the matching version of the computer code that exploits it. It is however not always sufficient, as in some cases the computing environment and hardware may also play a part in reproducibility.

GIT is a free multi-platform revision control system. Git allows to track versions of large textual datasets (such as computer code or \LaTeX documents), and enables to restore, compare and merge files across their various branches and versions. Git supports decentralized work and nonlinear workflows. GITHUB is a git hosting platform, including open and private repositories, and provides additional features [29, 30]. Many institutions offer their own Git server and include access rights (e.g. the EPFL Git server). Using an institutional server generally ensures that your data is safely stored and it avoids data leaks.

3.4.2 File sharing

The INSTITUTIONAL FILE SHARE is usually a good option if you are not working with external people. Using an institutional server generally ensures that your data is safely stored and it avoids data leaks. For example, at EPFL, the default on-line storage can be monitored from the MyNAS Web interface, in which it is possible to request more space and find the access configuration for each operating system. In addition, the EPFL-IT services offers a range of storage options, depending on the quantity and the access modalities required for your data. For more information please contact 1234@epfl.ch.

OWNCLOUD is a free software: it synchronizes and shares data on several computers (supported systems: Windows, Linux, Mac, Android and iOS). OwnCloud functionality may be extended through many plugins such as calendar share, collaborative editing, music player, pictures galleries, password storage, and so on. Institutions often offer an ownCloud instance to their members. Using an institutional server generally ensures that your

data is safely stored and avoids data leaks. For example, EPFL members have free access to SWITCHDRIVE, an instance of ownCloud opened to most Swiss universities, and therefore, ideal for sharing data among the national academic community[56, 74]. If an institutional instance is not available or unsuitable for your needs, many ownCloud providers are available[57], including some in Switzerland, e.g Woelkli.com[88].

GIT-ANNEX is an extension of Git, the versioning software described in the previous section. *Per se* Git is not well suited to manage non-textual (binary) files. This tool extends Git in that regard, and comes with the git-annex assistant, allowing to create an easy to use decentralized file share. In other words, this is comparable sharing file via ownCloud, however without the need for a central server [3].

3.4.3 Graphs

PLOTLY is a collaborative “suite of data visualization and collaboration tools for engineers and data scientists” [63].

DIA is a multiplatform open source diagram editor that supports among others Unified Modeling Language (UML) formalism [31].

GRAPHVIZ “Graphviz is open source graph visualization software. Graph visualization is a way of representing structural information as diagrams of abstract graphs and networks. It has important applications in networking, bioinformatics, software engineering, database and web design, machine learning, and in visual interfaces for other technical domains” [34].

3.4.4 Geodata

QGIS is a free multi-platform Geographic Information System (GIS) software with data editing, analysis and viewing functionality [69].

3.4.5 3D

PARAVIEW “is an open-source, multi-platform data analysis and visualization application. ParaView users can quickly build visualizations to analyze their data using qualitative and quantitative techniques. The data exploration can be done interactively in 3D or programmatically using ParaView’s batch processing capabilities. ParaView was developed to analyze extremely large datasets using distributed memory computing resources” [59].

MESHLAB “MeshLab is an open source, portable, and extensible system for the processing and editing of unstructured 3D triangular meshes” [47].

MeshLab can convert mesh files to STereoLithography (STL) files, widely used by 3D applications.

3.4.6 Papers and annotations

AUTHOREA is collaborative L^AT_EXbased writing tool: it handles group authoring. It supports text, images, mathematical formulas, computer code, bibliographic reference management, citation and bibliography, footnotes, commentaries, iPhython Notebooks, and many other features. A third alternative to Atlas and Authorea is OVERLEAF [4, 54].

Atlas is another collaborative writing tool similar to Authorea, but WYSIWYG (WYSIWYG: what you see is what you get, in practice similar to a classical word processing application). The Atlas platform is a free software and could therefore be installed locally on your server or more generally at your institution [55, 4, 54].

HYPOTHESIS is collaborative annotation tool, with which you can take personal notes, discuss, collaborate and organize your research [38].

OWNCLOUD is a collaborative software, particularly useful for file sharing (see section 3.4.2). In addition, many plugins, such as text editor with real-time collaborative editing are available [56].

3.4.7 Reference management and sharing

Reference management is generally integrated in collaborative authoring tools (e.g. Authorea, Atlas, Overleaf), which were discussed in the previous section.

ZOTERO is a free mutlplatform reference management software: it manages bibliographic references and associated research materials (PDF files, Web pages captures, etc.). A specificity of Zotero is its integration with Web browsers, which allows to naturally capture research materials and references in one click while surfing. The software is extensible via plugins, in particular writing tools for citation and bibliography management. The tool boasts an excellent integration with Libre Office, Open Office, and Neo Office Microsoft Word, and BibTex is supported. The Zotero social and collaborative features enable the synchronization of bibliographies among public or private groups. Ressource annotations are shared as well. Zotero does not require the users to upload any data to its servers [91].

MENDELEY is another multiplatform desktop reference management software and an online social researcher network. Contrarily to Zotero, all basic

metadata is sent to its servers and this cannot be turned off. Mendeley is a proprietary software belonging to Elsevier [46].

3.4.8 Scientific workflows and laboratory management

MYEXPERIMENT “is a social website for sharing [...] scientific workflows [...]” and integrates many tools such as Taverna and Bioclipse [49, 50].

AiiDA, the Automated Interactive Infrastructure and Database for Computational is a “flexible and scalable informatics’ infrastructure to manage, preserve, and disseminate the simulations, data, and workflows of modern-day computational science. Able to store the full provenance of each object, and based on a tailored database built for efficient data mining of heterogeneous results, AiiDA gives the user the ability to interact seamlessly with any number of remote HPC resources and codes, thanks to its flexible plugin interface and workflow engine for the automation of complex sequences of simulations” [2]. This tool is developed at EPFL. AiiDA’s core is free software, some of its plugins are licensed for non-commercial use [62].

PEGASUS and TAVERNA are open source workflow management systems, both able to execute applications [60, 75].

KEPLER is a free software system for designing, executing, reusing, evolving, archiving, and sharing scientific workflows [66, 67].

LABORATORY INFORMATION MANAGEMENT SYSTEMS are software that take modern laboratory operations in charge. Their features often include workflow management, data tracking, sample tracking, data exchange interfaces and enterprise resource planning [85]. For example, at EPFL several labs use the SLIMS software and ETHZ is developing an open source tool: openBIS [21, 20].

3.5 Data publication tools

3.5.1 Data repositories

Data repositories are listed in section 3.1.1, page 6.

3.5.2 Interactive publications

Interactive publications are a new generation of publications: in addition to the iPython Notebooks discussed just below, Maple and Mathematica use the concept interactive publication too. These software are however more mathematics-oriented (see section 3.3 on page 10).

IPYTHON NOTEBOOKS (IPYNB) are interactive publications making use of the Python language. Hence, these are based on free and multiplatform software. Such notebooks have two advantages in academia: they promote research reproducibility (because they constitute an excellent way to comment the code, run on all platforms and require only free software) and are excellent pedagogical tools. A student may easily play with the code illustrating a concept, for example he or she can change parameters to get a practical understanding of theories and algorithms. Practically, in addition to the interpreted-code parts, IPYNB can be structured using various header-levels, rich text (markdown), and mathematical formulas (\LaTeX). Since they are Python based, these notebook can take advantage of many scientific packages that may be found in the Python Package Index (PyPI), PyData and/or SciPy. For example: NumPy (numeric analysis), Matplotlib (plotting), Sympy (symbolic mathematics), Pandas (data structures and analysis) [51, 24]. In addition, iPython notebooks may be shared or hosted through Web based tools such as the collaborative writing platform Authorea (see section 3.4.6 page 13) or Wakari.io [84, 40].

3.5.3 Data licenses and Laws

Personal Data Protection Laws

The EUROPEAN UNION has a personal data legal protection framework. A significant part of this is DIRECTIVE 95/46/EC [22], which will probably be superseded by the GENERAL DATA PROTECTION DIRECTIVE (GDPR) [83].

The SWITZERLAND has a personal data legal protection framework. A significant part of this is LOI FÉDÉRALE SUR LA PROTECTION DES DONNÉES (LPD) [1] .

Text and multimedia licenses

CREATIVE COMMONS BY licenses enable to choose exactly what is allowed to do with your text and multimedia documents. In addition to the CC0 (see below), the CC-BY offers 6 variants[8]:

-  (Attribution : CC-By) “lets others distribute, remix, tweak, and build upon your work, even commercially, as long as they credit you for the original creation. This is the most accommodating of licenses offered.”
-  (Attribution-ShareAlike) “lets others remix, tweak, and build upon your work even for commercial purposes, as long as they credit you and license their new creations under the identical terms. [...] All new works based on yours will carry the same license, so any derivatives will also allow commercial use.”

-  (Attribution-NonCommercial-NoDerivs) “This license allows for redistribution, commercial and non-commercial, as long as it is passed along unchanged and in whole, with credit to you.”
-  (Attribution-NonCommercial) “lets others remix, tweak, and build upon your work non-commercially, and although their new works must also acknowledge you and be non-commercial, they don’t have to license their derivative works on the same terms.”
-  (Attribution-NonCommercial-ShareAlike) “lets others remix, tweak, and build upon your work non-commercially, as long as they credit you and license their new creations under the identical terms.”
-  (Attribution-NonCommercial-NoDerivs) “This license is the most restrictive of [the Creative Commons] six main licenses, only allowing others to download your works and share them with others as long as they credit you, but they can’t change them in any way or use them commercially.”

CREATIVE COMMONS ZERO , contrarily to CC-By the “CC0 enables scientists, educators, artists and other creators and owners of copyright- or database-protected content to waive those interests in their works and thereby place them as completely as possible in the public domain, so that others may freely build upon, enhance and reuse the works for any purposes without restriction under copyright or database law” [8].

Computer code licenses

GNU GENERAL PUBLIC LICENCE (GPL)  is to software what the Creative Commons Attribution-ShareAlike is to documents. It is the typical free software copyleft license and guarantees four freedoms to the users [27]:

- the freedom to use the software for any purpose,
- the freedom to change the software to suit your needs,
- the freedom to share the software with your friends and neighbors, and
- the freedom to share the changes you make.

Many licences are compatible with GPLv3, notably: Apache Licence 2.0, Artistic Licence 2.0, Berkley Database Licence, Modified BSD License, Boost Software License, CeCILL, CreativeCommons Zero, Educationnal Community Licence, AGLP, LGPL, IBM Public Licence, Intel Open source License, ISC License, MIT License / X11 License, Python Software Lisence,

W3C Software Notice and LicenseXFree86 License, zlib/libpng License and Zope Public License. [82]

GNU LESSER GENERAL PUBLIC LICENSE (LGPL) “The GNU Project has two principal licenses to use for libraries. One is the GNU Lesser GPL; the other is the ordinary GNU GPL. The choice of license makes a big difference: using the Lesser GPL permits use of the library in proprietary programs; using the ordinary GPL for a library makes it available only for free programs” [26].

GNU AFFERO GENERAL PUBLIC LICENSE (AGPL) “The GNU Affero General Public License is a modified version of the ordinary GNU GPL version 3. It has one added requirement: if you run a modified program on a server and let other users communicate with it there, your server must also allow them to download the source code corresponding to the modified version running there. The purpose of the GNU Affero GPL is to prevent a problem that affects developers of free programs that are often used on servers” [25].

Database licenses

The OPEN DATA COMMONS ATTRIBUTION LICENSE (ODC-BY) allows to share data (copy, distribute and use), to create (produce works using the data), to adapt the data (modify, adapt and build upon the data), as long as users attribute, i.e. “You must attribute any public use of the database, or works produced from the database, in the manner specified in the license. For any use or redistribution of the database, or works produced from it, you must make clear to others the license of the database and keep intact any notices on the original database” [53].

The OPEN DATA COMMONS OPEN DATABASE LICENSE (ODBL) allows to share, create and adapt as just above. As long as users attribute (as above for ODC-By). In addition to ODC-By, users have to share-alike (as in creative commons) and keep it open (if DRM are used, an open version of the database must also be redistributed) [52].

The OPEN DATA COMMONS PUBLIC DOMAIN DEDICATION AND LICENSE (PDDL) allows to share, create and adapt as just above, but without any restriction. It is thus similar to the CC0 license (see above in section 3.5.3).

Bibliography

- [1] admin.ch. *RS 235.1 Loi fédérale du 19 juin 1992 sur la protection des données (LPD)*. 1992. URL: <https://www.admin.ch/opc/fr/classified-compilation/19920153/> (visited on 09/15/2015).
- [2] AiiDA.net. *AiiDA*. 2015. URL: <http://www.aida.net/> (visited on 08/12/2015).
- [3] git annex. *git-annex*. 2015. URL: <http://git-annex.branchable.com/> (visited on 07/31/2015).
- [4] Authorea. *Authorea: Write research documents online, together.* - YouTube. 2015. URL: https://www.youtube.com/watch?v=Tz1Bh6JR_wA (visited on 07/31/2015).
- [5] Humboldt University Berlin. *Research data management policy — Forschungsdatenmanagement*. de. Seite. 2015. URL: <https://www.cms.hu-berlin.de/de/ueberblick/projekte/dataman/policy/policy-en/rdm-eng-policy> (visited on 08/05/2015).
- [6] Digital Curation Centre. *How to Develop a Data Management and Sharing Plan*. 2015. URL: <http://www.dcc.ac.uk/resources/how-guides/develop-data-plan> (visited on 07/30/2015).
- [7] European Commission. *Guidelines on Data Management in Horizon 2020*. 2013. URL: http://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-data-mgt_en.pdf (visited on 07/29/2015).
- [8] creativecommons.org. *Creative Commons*. 2015. URL: <https://creativecommons.org/> (visited on 08/03/2015).
- [9] Earth System Science Data. *ESSD*. 2015. URL: <http://www.earth-system-science-data.net/> (visited on 07/30/2015).
- [10] DataCite. *DataCite Schemas repository*. 2015. URL: <https://schema.datacite.org/> (visited on 07/30/2015).
- [11] DataShare. *Sources of dataset peer review - datashare - Wiki Service*. 2015. URL: <https://www.wiki.ed.ac.uk/display/datashare/Sources+of+dataset+peer+review> (visited on 07/30/2015).
- [12] DataTypeRegistry. *Data Type Registry*. 2015. URL: <http://www.typeregistry.org/registrar/> (visited on 07/29/2015).
- [13] DMPonline. *DMPonline*. 2015. URL: <https://dmponline.dcc.ac.uk/> (visited on 07/30/2015).
- [14] DMPTool. *DMPTool*. 2015. URL: <https://dmptool.org/> (visited on 07/30/2015).
- [15] Dryad. *Joint Data Archiving Policy - Dryad*. 2014. URL: <http://datadryad.org/pages/jdap> (visited on 09/26/2014).

- [16] Dryad. *Dryad Digital Repository - Dryad*. 2015. URL: <http://datadryad.org/> (visited on 07/29/2015).
- [17] Dryad. *Journal instructions - The Dryad data repository wiki*. 2015. URL: http://wiki.datadryad.org/Journal_instructions (visited on 07/30/2015).
- [18] DublinCore. *Dublin Core Metadata Element Set, Version 1.1*. 3013. URL: <http://dublincore.org/documents/dces/> (visited on 07/30/2015).
- [19] University of Edinburgh. *Research Data Management Policy | Policies and Regulations* |. 2015. URL: <http://www.ed.ac.uk/schools-departments/information-services/about/policies-and-regulations/research-data-policy> (visited on 08/05/2015).
- [20] EPFL. *LSIS | SV-IT*. 2015. URL: <http://sv-it.epfl.ch/slims> (visited on 08/03/2015).
- [21] ETHZ. *ETH - CISD - openBIS*. 2015. URL: <http://www.cisd.ethz.ch/software/openBIS> (visited on 08/03/2015).
- [22] EurLex. *9546EC*. 1995. URL: <http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=CELEX:31995L0046:en:HTML> (visited on 09/15/2015).
- [23] Figshare. *figshare - credit for all your research*. 2015. URL: <http://figshare.com/> (visited on 07/29/2015).
- [24] Python Software Foundation. *Python.org*. 2015. URL: <https://www.python.org/> (visited on 07/29/2015).
- [25] FSF. *AGPL - Licence publique générale GNU Affero, v3.0*. 2015. URL: <http://www.gnu.org/licenses/agpl.html> (visited on 08/03/2015).
- [26] FSF. *LGPL - Licence publique générale GNU amoinerie, v3.0*. 2015. URL: <http://www.gnu.org/licenses/lgpl.html> (visited on 08/03/2015).
- [27] FSF. *GPL - Licence publique générale GNU, v3.0*. 2105. URL: <http://www.gnu.org/licenses/gpl.html> (visited on 08/03/2015).
- [28] GigaScience. *GigaScience*. 2015. URL: <http://www.gigasciencejournal.com/> (visited on 07/30/2015).
- [29] *Git*. 2015. URL: <https://git-scm.com/> (visited on 07/29/2015).
- [30] GitHub. *GitHub*. 2015. URL: <https://github.com> (visited on 07/29/2015).
- [31] GNU. *Dia - GNOME*. 2015. URL: <https://wiki.gnome.org/Apps/Dia/> (visited on 09/14/2015).
- [32] GNU. *GNU Octave*. 2015. URL: <https://www.gnu.org/software/octave/> (visited on 07/29/2015).
- [33] GNU. *PSPP - GNU Project - Free Software Foundation*. 2015. URL: <https://www.gnu.org/software/pspp/> (visited on 07/29/2015).

- [34] Graphviz. *Graph Visualization Software*. 2015. URL: <http://www.graphviz.org/> (visited on 09/14/2015).
- [35] HDF Group. *HDF5*. 2015. URL: <https://www.hdfgroup.org/HDF5/> (visited on 09/14/2015).
- [36] Nature Publishing Group. *Recommended Repositories : Scientific Data*. 2014. URL: <http://www.nature.com/sdata/data-policies/repositories#q1> (visited on 06/27/2014).
- [37] Nature Publishing Group. *Scientific Data*. 2015. URL: <http://www.nature.com/sdata/> (visited on 07/30/2015).
- [38] hypothes.is. *Hypothesis*. 2015. URL: <https://hypothes.is/> (visited on 07/31/2015).
- [39] IBM. *IBM SPSS software*. 2015. URL: <http://www-01.ibm.com/software/analytics/spss/> (visited on 07/29/2015).
- [40] iPython.org. *The IPython Notebook — IPython*. 2015. URL: <http://ipython.org/notebook.html> (visited on 08/03/2015).
- [41] LibraryOfCongress. *Sustainability of Digital Formats: Planning for Library of Congress Collections*. 2015. URL: <http://www.digitalpreservation.gov/formats/> (visited on 07/29/2015).
- [42] University of Manchester. *University of Manchester Research Data Management Policy*. 2015. URL: http://www.miss.manchester.ac.uk/?page_id=425 (visited on 08/05/2015).
- [43] Maplesoft. *Maple 2015 - Technical Computing Software for Engineers, Mathematicians, Scientists, Instructors and Students - Maplesoft*. 2015. URL: <http://www.maplesoft.com/products/maple/> (visited on 07/29/2015).
- [44] MariaDB. *MariaDB*. 2015. URL: <https://mariadb.org/> (visited on 09/15/2015).
- [45] Mathworks. *Matlab*. 2015. URL: <https://ch.mathworks.com/products/matlab/> (visited on 07/29/2015).
- [46] Mendeley.com. *Mendeley*. 2015. URL: <https://www.mendeley.com/> (visited on 07/31/2015).
- [47] meshlab.sf.net. *MeshLab*. 2015. URL: <http://meshlab.sourceforge.net/> (visited on 09/15/2015).
- [48] MetadataDirectory. *Metadata Directory*. 2015. URL: <http://rd-alliance.github.io/metadata-directory/> (visited on 07/29/2015).
- [49] myExperiment. *myExperiment*. en. Page Version ID: 595344437. Feb. 2014. URL: <https://en.wikipedia.org/w/index.php?title=MyExperiment&oldid=595344437> (visited on 08/03/2015).

- [50] myExperiment.org. *myExperiment*. 2015. URL: <http://www.myexperiment.org/home> (visited on 08/03/2015).
- [51] NumPy. *NumPy*. 2015. URL: <http://www.numpy.org/> (visited on 07/29/2015).
- [52] OpenDataCommons. *ODbL : Open Data Commons Open Database License (ODbL) | Open Data Commons*. 2015. URL: <http://opendatacommons.org/licenses/odbl/> (visited on 08/05/2015).
- [53] OpenDataCommons. *ODC-By : Open Data Commons Attribution License | Open Data Commons*. 2015. URL: <http://opendatacommons.org/licenses/by/> (visited on 08/05/2015).
- [54] O'Reilly. *Getting-Started-with-Atlas · GitHub*. 2015. URL: <https://github.com/oreillymedia/Getting-Started-with-Atlas> (visited on 07/31/2015).
- [55] O'Reilly. *Atlas*. 2015. URL: <https://atlas.oreilly.com/> (visited on 07/31/2015).
- [56] ownCloud. *ownCloud.org*. 2015. URL: <https://owncloud.org/> (visited on 07/29/2015).
- [57] ownCloud.org. *ownCloud Providers*. 2015. URL: <https://owncloud.org/providers/> (visited on 09/14/2015).
- [58] University of Oxford. *University of Oxford Policy on the management of research data and records*. 2015. URL: <http://researchdata.ox.ac.uk/university-of-oxford-policy-on-the-management-of-research-data-and-records/> (visited on 08/05/2015).
- [59] paraview.org. *ParaView*. 2015. URL: <http://www.paraview.org/> (visited on 09/15/2015).
- [60] Pegasus. *Pegasus | Workflow Management System*. 2015. URL: <http://pegasus.isi.edu/> (visited on 08/12/2015).
- [61] Pensoft. *Biodiversity Data Journal*. 2015. URL: <http://biodiversitydatajournal.com/> (visited on 07/30/2015).
- [62] Giovanni Pizzi et al. "AiiDA: Automated Interactive Infrastructure and Database for Computational Science". In: *arXiv:1504.01163 [cond-mat, physics:physics]* (Apr. 2015). arXiv: 1504.01163. URL: <http://arxiv.org/abs/1504.01163> (visited on 08/12/2015).
- [63] plot.ly. *Plotly*. 2015. URL: <https://plot.ly/> (visited on 07/31/2015).
- [64] PostgreSQL. *PostgreSQL: The world's most advanced open source database*. 2015. URL: <http://www.postgresql.org/> (visited on 09/15/2015).
- [65] R Project. *R: The R Project for Statistical Computing*. 2015. URL: <https://www.r-project.org/> (visited on 07/29/2015).

- [66] kepler project.org. *Kepler scientific workflow system*. en. Page Version ID: 644507207. Jan. 2015. URL: https://en.wikipedia.org/w/index.php?title=Kepler_scientific_workflow_system&oldid=644507207 (visited on 08/03/2015).
- [67] kepler project.org. *The Kepler Project — Kepler*. 2015. URL: <https://kepler-project.org/> (visited on 08/03/2015).
- [68] PyData. *PyData.org | Downloads*. 2015. URL: <http://pydata.org/downloads/> (visited on 08/06/2015).
- [69] qgis.org. *QGIS project*. 2015. URL: <http://www.qgis.org/en/site/> (visited on 09/15/2015).
- [70] re3data. *re3data.org | Registry of Research Data Repositories*. 2015. URL: <http://www.re3data.org/> (visited on 07/29/2015).
- [71] rOpenSci.org. *rOpenSci - Open Tools for Open Science*. 2015. URL: <https://ropensci.org/> (visited on 07/29/2015).
- [72] ScienceExchange. *Science Exchange - Order experiments from the world's best labs*. 2015. URL: <https://www.scienceexchange.com/> (visited on 07/29/2015).
- [73] University of Southampton. *University of Southampton Reserach Data Management Policy*. 2015. URL: <http://www.calendar.soton.ac.uk/sectionIV/research-data-management.html> (visited on 08/05/2015).
- [74] SWITCHdrive. *SWITCHdrive*. 2015. URL: <https://drive.switch.ch/> (visited on 07/31/2015).
- [75] Taverna. *Taverna - open source and domain independent Workflow Management System*. 2015. URL: <http://www.taverna.org.uk/> (visited on 08/12/2015).
- [76] trac. *Blog: A list of Data Journals (in no particular order) – PREPARDE*. 2013. URL: <http://proj.badc.rl.ac.uk/preparde/blog/DataJournalsList> (visited on 07/30/2015).
- [77] uml.org. *Unified Modeling Language (UML)*. 2015. URL: <http://www.uml.org/> (visited on 09/15/2015).
- [78] UpMetaJournals. *Journal of Open Archaeology Data*. 2015. URL: <http://openarchaeologydata.metajnl.com/> (visited on 07/30/2015).
- [79] UpMetaJournals. *Open Health Data*. 2015. URL: <http://openhealthdata.metajnl.com/> (visited on 07/30/2015).
- [80] W3C. *OWL - Semantic Web Standards*. 2009. URL: <http://www.w3.org/2001/sw/wiki/OWL> (visited on 07/30/2015).
- [81] W3C. *RDF - Semantic Web Standards*. 2014. URL: <http://www.w3.org/RDF/> (visited on 07/30/2015).

- [82] Wikipedia. *Comparison of free and open-source software licenses*. en. Page Version ID: 671857753. July 2015. URL: https://en.wikipedia.org/w/index.php?title=Comparison_of_free_and_open-source_software_licenses&oldid=671857753 (visited on 08/05/2015).
- [83] Wikipedia. *General Data Protection Regulation*. 2015. URL: https://en.wikipedia.org/wiki/General_Data_Protection_Regulation (visited on 09/15/2015).
- [84] Wikipedia. *IPython - Notebooks*. en. Page Version ID: 673833812. July 2015. URL: <https://en.wikipedia.org/w/index.php?title=IPython&oldid=673833812> (visited on 08/03/2015).
- [85] Wikipedia. *Laboratory information management system*. en. Page Version ID: 673350789. July 2015. URL: https://en.wikipedia.org/w/index.php?title=Laboratory_information_management_system&oldid=673350789 (visited on 08/03/2015).
- [86] Wikipedia. *Ontology (information science)*. en. Page Version ID: 669326401. June 2015. URL: [https://en.wikipedia.org/w/index.php?title=Ontology_\(information_science\)&oldid=669326401](https://en.wikipedia.org/w/index.php?title=Ontology_(information_science)&oldid=669326401) (visited on 07/30/2015).
- [87] Wikipedia. *SQL*. 2015. URL: <https://en.wikipedia.org/wiki/SQL> (visited on 09/15/2015).
- [88] Woelkli. *Secure Cloud Storage in Switzerland*. 2015. URL: <https://woelkli.com/en> (visited on 09/14/2015).
- [89] Wolfram. *Wolfram Mathematica: L'outil de calcul technique le plus abouti*. 2015. URL: <http://www.wolfram.com/mathematica/> (visited on 07/29/2015).
- [90] Zenodo. *Zenodo*. 2015. URL: <https://zenodo.org/> (visited on 07/29/2015).
- [91] Zotero.org. *Zotero | Home*. 2015. URL: <https://www.zotero.org/> (visited on 07/31/2015).