ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE BIBLIOTHÈQUE

Data Repositories : a publication guide

by

Jan Krause

December 23, 2015

Contents

1 Abstract

This document is a **guide to data repositories for publication purposes**. It offers specific information for EPFL researchers.

Some sections of this this document were reproduced or adapted from "A Selection of Research Data Management Tools Throughout the Data Lifecycle" [22], which adopts a more generalist point of view on research data management tools.

2 Publishing your data

Setting up a data management plan (DMP) is a way to covers prepare for all the issues related to a project data lifecycle management. At EPFL, you can get support at datamanagementplan@epfl.ch. Regarding data publication specifically, the following elements are of importance:

- **data format:** see section 4.2,
- **metadata format:** see section 4.3,
- **data licence:** see section 4.3.3,
- **data workflow:** see section 4.3.4.

For working data or unpublished data **EPFL offers institutional storage 3.4**.

Domain repositories are generally recommended because they offer adapted functionality (e.g. specific data and metadata adaptations) and are well know by the related community, thus maximiziong content's discoverability and reuse. To find a domain specific repository, **see section 3.2**. In addition, a way to increase your datasets' impact are their description in data papers, see section 3.3.

However, **all repositories are not to be trusted**: some might not last in long or even the mid term, be technically unreliable (broken links, bit rot...), and so on. For more information on the elements to **evaluate a repository trustworthiness, see section 4**.

Is no suitable domain data repositories is appropriated for a specific datastet, **generalist data repositories are available, see section 3.1**. Another option, may be to fall back on an institutional repository. Today, there is no such tool at EPFL, but Infoscience may be used to publish small datasets.

3 Data repositories and storage

The goal of this section is to help find a repository to publish datasets:

- **Generalist data repositories:** publish any type of data.
- **Data repositories directories:** find the most appropriated repository for a specific dataset.
- **Data papers and data journal:** data papers are publications describing datasets, which may be published in a repository or in a journal as supplementary material.
- **Data storage at EPFL:** institutional storage infrastructures and related pricing.

In addition, in the next section (section 4, on page 5) significant elements to consider for datasets publications are discussed.

3.1 Generalist data repositories

ZENODO [50], DRYAD [8] and FIGSHARE [13] 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.

3.2 Data repositories directories

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 [36].

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 [19].

3.3 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 [9]
- Trac's multidisciplinary data journals list[39]
- Nature Publishing Group's scientific data website [20]
- DataShare's sources of dataset peer review list of data journals [5]
- GigaScience [17]

Some discipline specific data journals exist too, for example:

- Wiley's Geoscience Data Journal and Earth System Science Data [3]
- UpMetaJournal's Open Health Data [41]
- Pensoft's Biodiversity Data Journal [30]
- UpMetaJournal's Journal of open archaeology data [40]

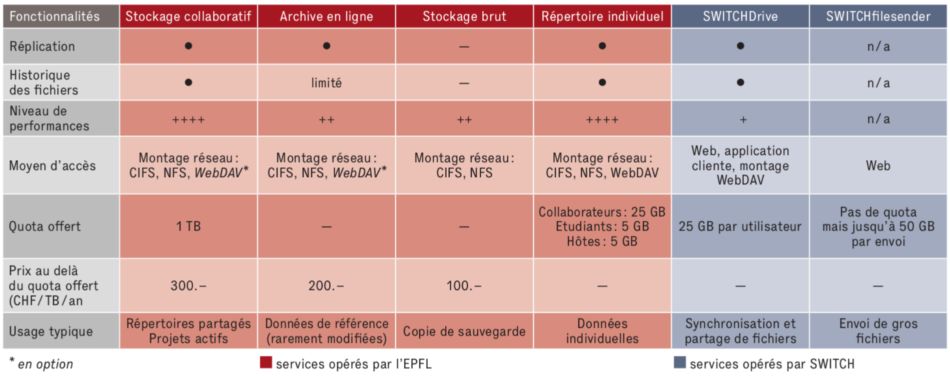
In addition, a list of JOURNALS DATA POLICIES compiled by Dryad may be of interest [7].

3.4 Data storage at EPFL

EPFL offers three types of storage, according to the Flash magazine, Number 4, 2015:

- **Collaborative storage [Stockage collaboratif]:** The standard storage for lab. Each lab has by default 1TB offered, and may extend the capacity for 300CHF/TB/year. This service offers good performances and high resiliency (high ability to continue operating or recover if there is a failure). Indeed, data are replicated on a second geographical site and file history is preserved, allowing recovery after accidental modification or deletion. It may be used as a shared drive.
- **On-line archive [Archive en ligne]:** This service is suited for data that are not modified frequently. It has the same degree of resiliency as the former service, but with lesser performances and version history. Its cost is of 200CHF/TB/year.
- **Raw storage [Stockage brut]:** This storage may be viewed as a resilient virtual disk. It may be used as backup location for local data. The cost is of 100CHF/TB/year.

All the above storage solutions are located in Switzerland, and subject to Swiss law.



Fonctionnalités	Stockage collaboratif	Archive en ligne	Stockage brut	Répertoire individuel	SWITCHDrive	SWITCHfilesender
Réplication	●	●	–	●	●	n/a
Historique des fichiers	●	limité	–	●	●	n/a
Niveau de performances	++++	++	++	++++	+	n/a
Moyen d'accès	Montage réseau : CIFS, NFS, *WebDAV**	Montage réseau : CIFS, NFS, *WebDAV**	Montage réseau : CIFS, NFS	Montage réseau : CIFS, NFS, WebDAV	Web, application cliente, montage WebDAV	Web
Quota offert	1 TB	–	–	Collaborateurs : 25 GB Etudiants : 5 GB Hôtes : 5 GB	25 GB par utilisateur	Pas de quota mais jusqu'à 50 GB par envoi
Prix au delà du quota offert (CHF/TB/an	300.–	200.–	100.–	–	–	–
Usage typique	Répertoires partagés Projets actifs	Données de référence (rarement modifiées)	Copie de sauvegarde	Données individuelles	Synchronisation et partage de fichiers	Envoi de gros fichiers

* *en option* 🟥 services opérés par l'EPFL ■ services opérés par SWITCH

Source: EPFL Flash, Number 4, 2015.

4 Significant elements for data publication

The purpose of this section is to highlight and describe the diferent key elements to consider when publishing data.

- **Standards and certification:** Good practices ensure the visibility, accessibility, availability and integrity of datasets. They include attribution of persistent identifiers, data curation in general (such as bit preservation) and an appropriated infrastructure. Data repositories certifications guarantee a sufficient quality.

- **Data formats:** The use of a standard data format increases datasets' interoperability. It is hence crucial to foster their reuse with other software or platforms, through time, and by other people as well.

- **Metadata Formats:** The use of a standard metadata format (or descriptive data) increases datasets' visibility in search engines, and the potential for their reuse.

- **Data Worflows:** Reproducible research requires a good description of data provenance and computation workflow.

- **Data and code licenses:** Licenses allow to define precisely what others may do with your data, and what they may not do.

In addition, in the next section (section 4, on page 5) significant elements to consider for datasets publications are discussed.

4.1 Standards and certifications

4.1.1 Peristant identifiers

Persistent identifiers (PID) allow to identity unambiguously, resolve and retrieve datasets (and other digital objects). The most commonly used PIDs

is the DOI (see below). PIDs avoid the loss of access to datasets due to broken links, which is quite frequent: A Plos One study showed in 2014 that more than 60% of links to datasets are broken after 10 years [31]. Another Plos One 2014 article showed that the bibliography of one article out of five is impacted by that phenomenon [37].
The DIGITAL OBJECT IDENTIFIER (DOI) " are characters strings (a "digital identifier") used to uniquely identify an object such as an electronic document. Metadata about the object is stored in association with the DOI name and this metadata may include a location, such as a URL, where the object can be found. The DOI for a document remains fixed over the lifetime of the document, whereas its location and other metadata may change. Referring to an online document by its DOI provides more stable linking than simply referring to it by its URL, because if its URL changes, the publisher need only update the metadata for the DOI to link to the new URL" [44].

4.1.2 Bit preservation

Bitrot is a phenomenon more frequent than we think. Causes are complex and varied: disk errors, RAID errors, memory errors... [21, 28]. To avoid this, some repositories offer bit preservation, or, even better are fully OAIS compliant [47].

4.1.3 Certifications

DATA SEAL OF APPROVAL and WORLD DATA SYSTEM are the two main data repositories certifications. Certified repositories may be trusted, unfortunately only few repositories are certified at that time.

4.2 Data formats

4.2.1 Generalist data formats

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" [18]. 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)"[49]. SQL is well suited to store and share relational data. RDBMS are a great help in maintaining dateset's coherence and enforce data constraints. Several multi-platform open source RDBMS are available, such as MariaDB (MySQL) [24] or PosgreSQL [33].

4.2.2 Subject specific 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 [23].

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 [6].

4.3 Metadata formats

4.3.1 Generalist metadata formats

DUBLINCORE is a vocabulary consisting of only 15 basic elements, such as Creator, Title, Date, Description, Format, Rights, or Subject [10]. 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. The basic DublinCore may be extended using qualifiers.

DATACITE METADATA SCHEMA is a standard designed with datasets in mind, and hence more adapted then DublinCore mentioned just above. For example, GeoLocation, ReserarchGroups, Collections, Videos or Workflows have their own specific resource types [4].

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.

4.3.2 Subject specific metadata formats

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 [25].

The Semantic Web Resource Description Format (RDF) “is a standard model data interchange on the Web”[43]. RDF is a very general format, however it my be used to describe precisely most types of data. Indeed, many OWL [42] RDF based ontologies exist. Ontologies are “formal naming and definition of types, properties and interrelationship” of data [48].

4.3.3 Data and code licences

Creative Commons BY licenses enable to choose exactly what is allowed to do with your datasets, text and multimedia documents. In addition to the CC0 (see below), the CC-BY offers 6 variants[2]:

- cc BY (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.”
- cc BY SA (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.”
- cc BY ND (Attribution-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.”
- cc BY NC (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.”
- cc BY NC SA (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.”
- cc BY NC ND (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" [2].

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 [16]:

- 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 Lisence, zlib/libpng License and Zope Public Lisence. [45]

GNU Lesser General Public licence (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" [15].

GNU Affero Genreal Public licence (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" [14].

4.3.4 Data Workflows

myExperiment "is a social website for sharing [...] scientific workflows [...]" and integrates many tools such as Taverna and Bioclipse [26, 27].

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” [1]. This tool is developed at EPFL. AiiDA’s core is free software, some of its plugins are licensed for non-commercial use [32].

Pegasus and Taverna are open source workflow management systems, both able to execute applications [29, 38].

Kepler is a free software system for designing, executing, reusing, evolving, archiving, and sharing scientific workflows [34, 35].

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 [46]. For example, at EPFL over 25 labs use the SLIMS software and ETHZ is developing an open source tool: openBIS [12, 11].

Index

Bibliography

[1] AiiDA.net. *AiiDA*. 2015. URL: http://www.aiida.net/ (visited on 08/12/2015).

[2] creativecommons.org. *Creative Commons*. 2015. URL: https://creativecommons.org/ (visited on 08/03/2015).

[3] Earth System Science Data. *ESSD*. 2015. URL: http://www.earth-system-science-data.net/ (visited on 07/30/2015).

[4] DataCite. *DataCite Schemas repository*. 2015. URL: https://schema.datacite.org/ (visited on 07/30/2015).

[5] 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).

[6] DataTypeRegistry. *Data Type Registry*. 2015. URL: http://www.typeregistry.org/registrar/ (visited on 07/29/2015).

[7] Dryad. *Joint Data Archiving Policy - Dryad*. 2014. URL: http://datadryad.org/pages/jdap (visited on 09/26/2014).

[8] Dryad. *Dryad Digital Repository - Dryad*. 2015. URL: http://datadryad.org/ (visited on 07/29/2015).

[9] Dryad. *Journal instructions - The Dryad data repository wiki*. 2015. URL: http://wiki.datadryad.org/Journal_instructions (visited on 07/30/2015).

[10] DublinCore. *Dublin Core Metadata Element Set, Version 1.1*. 3013. URL: http://dublincore.org/documents/dces/ (visited on 07/30/2015).

[11] EPFL. *LSIS | SV-IT*. 2015. URL: http://sv-it.epfl.ch/slims (visited on 08/03/2015).

[12] ETHZ. *ETH - CISD - openBIS*. 2015. URL: http://www.cisd.ethz.ch/software/openBIS (visited on 08/03/2015).

[13] Figshare. *figshare - credit for all your research*. 2015. URL: http://figshare.com/ (visited on 07/29/2015).

[14] FSF. *AGPL - Licence publique générale GNU Affero, v3.0*. 2015. URL: http://www.gnu.org/licenses/agpl.html (visited on 08/03/2015).

[15] FSF. *LGPL - Licence publique générale GNU amoindrie, v3.0*. 2015. URL: http://www.gnu.org/licenses/lgpl.html (visited on 08/03/2015).

[16] FSF. *GPL - Licence publique générale GNU, v3.0*. 2105. URL: http://www.gnu.org/licenses/gpl.html (visited on 08/03/2015).

[17] GigaScience. *GigaScience*. 2015. URL: http://www.gigasciencejournal.com/ (visited on 07/30/2015).

[18] HDF Group. *HDF5*. 2015. URL: https://www.hdfgroup.org/HDF5/ (visited on 09/14/2015).

[19] Nature Publishing Group. *Recommended Repositories : Scientific Data*. 2014. URL: http://www.nature.com/sdata/data-policies/repositories#q1 (visited on 06/27/2014).

[20] Nature Publishing Group. *Scientific Data*. 2015. URL: http://www.nature.com/sdata/ (visited on 07/30/2015).

[21] Robin Harris. *Data corruption is worse than you know*. 2007. URL: http://www.zdnet.com/article/data-corruption-is-worse-than-you-know/ (visited on 12/21/2015).

[22] Jan Krause. "A Selection of Research Data Management Tools Throughout the Data Lifecycle". In: (2015). URL: http://infoscience.epfl.ch/record/211157?ln=en (visited on 12/23/2015).

[23] LibraryOfCongress. *Sustainability of Digital Formats: Planning for Library of Congress Collections*. 2015. URL: http://www.digitalpreservation.gov/formats/ (visited on 07/29/2015).

[24] MariaDB. *MariaDB*. 2015. URL: https://mariadb.org/ (visited on 09/15/2015).

[25] MetadataDirectory. *Metadata Directory*. 2015. URL: http://rd-alliance.github.io/metadata-directory/ (visited on 07/29/2015).

[26] 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).

[27] myExperiment.org. *myExperiment*. 2015. URL: http://www.myexperiment.org/home (visited on 08/03/2015).

[28] Bernd Panzer-Steindel. *Data integrity*. 2007. URL: https://indico.cern.ch/event/13797/session/0/contribution/3/attachments/115080/163419/Data_integrity_v3.pdf.

[29] Pegasus. *Pegasus | Workflow Management System*. 2015. URL: http://pegasus.isi.edu/ (visited on 08/12/2015).

[30] Pensoft. *Biodiversity Data Journal*. 2015. URL: http://biodiversitydatajournal.com/ (visited on 07/30/2015).

[31] Alberto Pepe. *How Do Astronomers Share Data? Reliability and Persistence of Datasets Linked in AAS Publications and a Qualitative Study of Data Practices among US Astronomers*. 2014. URL: about:reader?url=http%3A%2F%2Fjournals.plos.org%2Fplosone%2Farticle%3Fid%3D10.1371%2Fjournal.pone.0104798 (visited on 12/21/2015).

[32] 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).

[33] PostgreSQL. *PostgreSQL: The world's most advanced open source database.* 2015. URL: http://www.postgresql.org/ (visited on 09/15/2015).

[34] 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).

[35] kepler project.org. *The Kepler Project — Kepler.* 2015. URL: https://kepler-project.org/ (visited on 08/03/2015).

[36] re3data. *re3data.org | Registry of Research Data Repositories.* 2015. URL: http://www.re3data.org/ (visited on 07/29/2015).

[37] Herbert Van de Sompel. *Scholarly Context Not Found: One in Five Articles Suffers from Reference Rot.* 2014. URL: http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0115253 (visited on 12/21/2015).

[38] Taverna. *Taverna - open source and domain independent Workflow Management System.* 2015. URL: http://www.taverna.org.uk/ (visited on 08/12/2015).

[39] 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).

[40] UpMetaJournals. *Journal of Open Archaeology Data.* 2015. URL: http://openarchaeologydata.metajnl.com/ (visited on 07/30/2015).

[41] UpMetaJournals. *Open Health Data.* 2015. URL: http://openhealthdata.metajnl.com/ (visited on 07/30/2015).

[42] W3C. *OWL - Semantic Web Standards.* 2009. URL: http://www.w3.org/2001/sw/wiki/OWL (visited on 07/30/2015).

[43] W3C. *RDF - Semantic Web Standards.* 2014. URL: http://www.w3.org/RDF/ (visited on 07/30/2015).

[44] Wikipedia. *Digital object identifier.* en. Page Version ID: 695617760. 2014. URL: https://en.wikipedia.org/w/index.php?title=Digital_object_identifier&oldid=695617760 (visited on 12/22/2015).

[45] 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).

[46] 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).

[47] Wikipedia. *OAIS*. de. Page Version ID: 127450705. 2015. URL: https://de.wikipedia.org/w/index.php?title=OAIS&oldid=127450705 (visited on 12/22/2015).

[48] 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 (visited on 07/30/2015).

[49] Wikipedia. *SQL*. 2015. URL: https://en.wikipedia.org/wiki/SQL (visited on 09/15/2015).

[50] Zenodo. *Zenodo*. 2015. URL: https://zenodo.org/ (visited on 07/29/2015).