MISSION RDM: THE ULTIMATE QUEST BEFORE THE HOLIDAYS

Complementary guide

Serious game on research data management

Created by Myriam JOUHAR, Pauline MELLY, Alexia TROMBERT

Translated by Marie-Catherine BEAULIEU, Jolanda ELMERS, Pauline MELLY, Alexia TROMBERT

2024

Table of contents

Preamble

The information in this complementary guide is taken from the serious game "Mission RDM: The last quest before the holidays". During the game, the specialist role cards provide theoretical information about research data management but are not intended to be exhaustive. This information has been compiled in the present document for easy reuse and further learning.

As the game was originally designed in French, some of the sources are in French.

MISSION RDM: THE ULTIMATE QUEST BEFORE THE HOLIDAYS

Roadmap

Research data management: « refers to the activity of working with research data throughout the research process, from data collection, to data storage and backup, through to data sharing at the end of a research project» University of Edinburgh 2022

Research data: « any information that has been collected, observed, generated or created to validate original research findings» Library University of Leeds 2023

Research data lifecycle

From Research data lifecycle - UK Data Service
https://ukdataservice.ac.uk/manage-data/lifecycle

Step 1. Project design and planning

1.1 Writing the DMP
- A. Administrative information
- B. Data collection
- C. Documentation and metadata
- D. Ethics and legal compliance
- E. Storage and backup strategies
- F. Selection and conservation
- G. Data sharing
- H. Responsibilities and resources

1.2 File structure and naming conventions
1.3 Reflect on the data produced or reused
1.4 Consider legal and ethical aspects
1.5 Analyze storage costs

Step 2. Data collection/creation

2.1 Backup and storage strategies
2.2 Collect/obtain data
2.3 Document your project

Step 3. Data processing and analysis

3.1 Data cleaning
3.2 Data analysis
3.3 Data visualization

Step 4. Data preservation and archiving

4.1 Use an open file format
4.2 Identify data value and determine data preservation
4.3 Add appropriate metadata
4.4 Anonymize data (if necessary)

Step 5. Data sharing

5.1 Select a data repository
5.2 Determine access conditions
5.3 Choose a persistent identifier
5.4 Assign a copyright reuse license

Step 6. Data reuse

6.1 Evaluate data access and reuse
6.2 Prepare dataset citation

FAIR principles: data must be...

- **F**indable: easy to find thanks to the use of adequate metadata and a persistent identifier
- **A**ccessible: the conditions of access to the data are clearly stated (open, close, require authentication or authorization, etc.) and remain accessible in the long term
- **I**nteroperable: usable on non-proprietary (open source) software, while respecting norms and standards
- **R**eusable: data should be well-described so that it can be easily replicated and/or combined and, under which conditions

GoFAIR, 2022

STEP 1. PROJECT DESIGN AND PLANNING

Writing the DMP

The DMP or Data Management Plan is a document drafted in the preliminary phase of a research project and completed as the project progresses.

It helps organize and anticipate all the stages of the data lifecycle. It explains how the project's data is managed, from creation to sharing and archiving.

The DMP contains the following information:

- Administrative information
- Data collection
- Documentation and metadata
- Ethics and legal compliance
- Storage and backup strategies
- Selection and conservation
- Data sharing
- Responsibilities and resources

For more information

- Library of Vrije Universiteit Amsterdam – Research Data Management Guide – Data Management Plan: https://libguides.vu.nl/rdm/dmp
- Doranum - Plan de gestion de données : fiche synthétique: https://doranum.fr/plan-gestion-donnees-dmp/plan-de-gestion-des-donnees-fiche-synthetique_10_13143_cgv4-0k53/
- Médiathèques HES-SO Valais-Wallis - Research data management : https://hevs-ch.libguides.com/RDM

Reflect on the data produced or reused

They can be classified according to their origin:

- **Observation data**: captured in real time, usually unique, impossible to reproduce. E.g., questionnaires, interviews, neuroimaging
- **Experimental data:** obtained from laboratory equipment, often reproducible, but can be expensive. E.g., chromatograms, DNA chips, trials
- **Simulation data:** generated by computer or simulation models, reproducible if well documented. E.g., meteorological data, earthquake simulation data
- **Derived or compiled data**: from processing or combining raw data, often reproducible but expensive. E.g., text mining, MRI imaging, compiled databases

We could also classify them according to their nature or their form (text, digital...) or their level of elaboration (raw, processed, analyzed).

For more information

- Doranum - L'origine et la description des données de la recherche: https://doranum.fr/plan-gestion-donnees-dmp/origine-description-donnees-recherche/
- gTIGRE - Former les professionnels de l'information et de la documentation aux données de la recherche en 45 minutes : https://zenodo.org/record/4610514#.YUQ3ZLgzaSM
- University of Ediburgh, MANTRA Research Data Management Training – Research data in context: https://mantra.ed.ac.uk/researchdataincontext/

File structure and naming conventions

It is important to establish file structure and naming conventions in the early stages of the project. Here are some of the most used rules:

Naming conventions

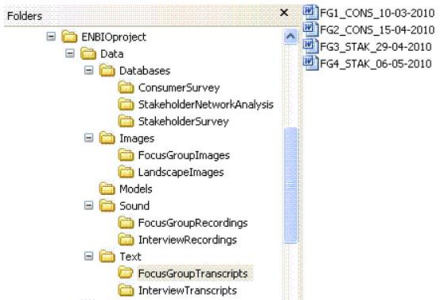
Choose a short and meaningful file name • Max. 32 characters • Start with the most important element • Use pre-defined acronyms to limit the number of characters • Should be easily understandable by all • Must contain: subjects, date, version	Meeting_20210902_AG_V01
Avoid special characters, punctuation, and accents	+, =, « », [ ], < >, $, %, &, é, è, ü, ç, ?, !, ., :, ;
Avoid spaces between words • Use underscores • Use capital letters	File_Naming_Conventions FileNamingConventions
Use standard date formats	YYYYMMDD 20210902 YYYY-MM-DD 2021-09-02
Include a pre-defined version control indicator	V01 V01Draft V03Final
File numbering • When using sequential numbering, use leading zeros to allow for multi-digit numbers	For a sequence of 1-10: 01-10 For a sequence of 1-100: 001-010-100

Source: authors of the game

Examples of naming convention

- [investigator]_[method]_[subject]_[YYYYMMDD]_[version].[ext]
- [project#]_[method]_[YYYYMMDD].[ext]
- [YYYYMMDD]_[version]_[subject]_[data collection].[ext]
- [file type]_[author]_[YYYYMMDD].[ext]

Also think about the tree structure of your files so that it is as coherent and understandable as possible.



Source: [1]

For more information

- [1]UK Data Service - Organising: https://ukdataservice.ac.uk/learning-hub/research-data-management/format-your-data/organising/
- UQAM - Recommandations pour le nommage de vos fichiers électroniques: https://uqam-ca.libguides.com/gestion-des-donnees-de-recherche/organisation-des-fichiers
- Médiathèques HES-SO Valais-Wallis - Research data management https://hevs-ch.libguides.com/RDM

STEP 2. DATA COLLECTION/CREATION

Backup and storage strategies

Here is a table that presents different storage solutions:

Storage solution	Security	Access	Cost	Recommendations
Personal or professional computer (local storage)	●●○○ Exposed to hacking, damage, and system failures	●○○○ - Only the person who has access to the computer can access the data - Collaborative work requires an external device or additional solution (mail, Cloud,...)	●●●● Little or no additional cost	- For temporary or short-term storage - Use in combination with encryption and password protection, especially if working with sensitive data
External devices	●○○○ - Easily lost or stolen - Limited lifespan (material deterioration)	●●●○ Easy to transport, they allow to transfer data from one computer to another	●●●● Little or no additional cost	- For temporary or short-term storage - Use in combination with encryption and password protection, especially if working with sensitive data - Potential quality control issues due to version confusion
Institutional network drive	●●●● - Data is centrally stored - Backups can be centrally managed and automated - Reliable, durable, and secure storage (better protected against theft, hacking, fire,...)	●●○○ Access for external project partners can be difficult or impossible	●●○○ Higher cost (but may be institutionally funded)	- For longer term storage - Suitable for the storage of sensitive data - Not provided by all institutions - Work with rights and permissions to ensure that not everyone has access to everything if this is not required
Cloud storage	●●○○ - Insufficient control over where the data is stored and how often it is backed up - Data can be lost if your account is suspended or deleted	●●●● Grants shared, remote and easy access to data to all involved in the project	●●○○ Fee-based once the set storage limit is reached	- To collaborate with individuals external to the institution - Not suitable for sensitive data - No control over the backup process

Source: [2] (translated and adapted)

Tip to avoid data loss: Use the 3+2+1+1+0 rule

- 3 copies on
- 2 different types of storage media with at least:
- 1 copy offsite
- 1 copy offline, and
- 0 errors after backup recoverability verification

For more information

- Doranum - Stockage, partage et archivage : quelles différences ?: https://doranum.fr/stockage-archivage/stockage-partage-archivage-quelles-differences/
- [2]Doranum - Stocker ses données de façon sécurisée : https://doranum.fr/stockage-archivage/stockage-donnees/
- Cessda Data Management Expert Guide, Chapter 4: Storage: https://dmeg.cessda.eu/Data-Management-Expert-Guide/4.-Store/Storage
- Nico Losschaert on Veeam Community: 3-2-1-1-0 Golden Backup Rule: https://community.veeam.com/blogs-and-podcasts-57/3-2-1-1-0-golden-backup-rule-569

Document your project

Datasets contain one or more data files and their documentation. This documentation will provide contextual information that will give meaning to the data, to facilitate its use, retrieval and management.

It is important to document your project as soon as possible. For each step of the research process, it is necessary to detail the procedures performed on the data, as well as the software and methods used. The project documentation should provide information about the project itself, the sampling,

the files, the structure and detail of the data within the files, and the content of the questions and answers.

The documentation can take different forms:

- Readme.txt file
- Codebook (variables dictionary)
- Electronic lab notebook (ELN)
- Field notebook
- Text file (.docx, .odt, .pdf)
- Directly integrated in the data file

For more information

- Library of Vrije Universiteit Amsterdam – Research Data Management Guide – Selecting Data: https://libguides.vu.nl/rdm/selecting-data-archiving
- UK data Service – Research data management: https://ukdataservice.ac.uk/learning-hub/research-data-management/

STEP 3. DATA PROCESSING AND ANALYSIS

Data cleaning

Before the data can be processed and analyzed, it must be cleaned to eliminate unnecessary fields and to facilitate analysis by ensuring standardization.

Which software should be used? It depends on the type of data, the quantity, the resources available and the team's experience. Here are some examples of free tools: R Studio, OpenRefine

How to clean up data?

- Identify essential data (e.g., delete unnecessary fields that were automatically added by the data collection software)
- Identify and eliminate duplicated data
- Resolve empty values
- Ensure data uniformity (data of the same type should have the same form)

Important: always keep a raw version of your file and keep track of all cleaning steps to ensure transparency and reproducibility!

For more information

- Talend – Pourquoi et comment nettoyer ses données d'entreprise ? https://www.talend.com/fr/resources/what-is-data-cleansing/
- Datascience.eu – 6 étapes pour le nettoyage des données et pourquoi c'est important: https://datascience.eu/fr/mathematiques-et-statistiques/6-etapes-pour-le-nettoyage-des-donnees-et-pourquoi-cest-important/
- Cessda Data Management Expert Guide, Chapter 3. Process: https://dmeg.cessda.eu/Data-Management-Expert-Guide/3.-Process

STEP 4. DATA PRESERVATION AND ARCHIVING

Use an open file format

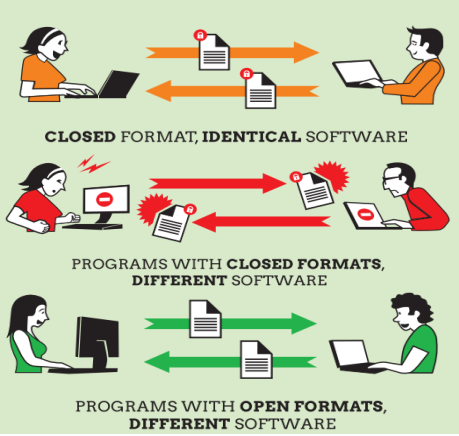
To ensure reuse, interoperability, and longevity of data, it is necessary to use an open file format. An open format is a file that can be read by multiple software applications. Here is a list of file formats extensions for reusability/preservation:

TYPE OF DATA	APPROPRIATE	ACCEPTABLE	DEPRECATED
Tabular (extensive metadata)	CSV – HDF5	TXT – HTML – TEX – FASTQ [3] – POR	
Tabular (minimal metadata)	CSV – TAB – ODS – SQL – TSV	XML (if appropriate DTD) – XLSX	XLS – XLSB
Textual / Presentation	TXT – PDF – ODT – ODM – TEX – MD – HTM – XML – EXTXYZ [4] – ODF	PPTX – RTF – DOCX – PDF (with embedded forms) – EPS – IPF	DOC – PPT – DVI – PS
Code / Computation	M – R – PY – IYPNB – RSTUDIO – RMD – NETCDF – AIML	SDD	MAT – RDATA
Image & Spectroscopy	TIF – PNG – SVG – JPEG – FITS	JCAMP – JPG – JP2 – TIF – TIFF – PDF – GIF – BMP – DM3 – OIR – LSM [5]	INDD – AIT – PSD – SPC
Audio	FLAC – WAV – OGG – MXL – MIDI – MEI – HUMDRUM	MP3 – AIF	
Video	MP4 – MJ2 – AVI – MKV	OGM – MP4 – WEBM	WMV – MOV – QT
Geospatial	NETCDF – tabular GIS attribute data – SHP – SHX – DBF – PRJ – SBX – SBN – POSTGIS – TIF – TFW – GEOJSON	MDB – MIF	
3D structures & images	X3D – X3DV – X3DB – PDF3D – POV – PDBML	DWG – DXF – PDB	PXP
Generic	XML – JSON – RDF		

Source: [3]

Using open formats serves two main objectives:

- To ensure access to your documents regardless of the software used by other people
- To ensure the longevity of your documents so that they remain accessible even if the software used to create them is not available or no longer exists



Source: [4 (edited)]

In short, when research teams get ready to archive their data, they should make sure to choose formats that are:

- open
- non-encrypted
- uncompressed
- used and recognized in their field
- interoperable

For more information

- [3]EPFL Library Research Data Management Fastguides: https://infoscience.epfl.ch/record/265349
- [4]April – Open formats, what for?: https://www.april.org/en/open-formats
- UK data Service – Recommended formats: https://ukdataservice.ac.uk/learning-hub/research-data-management/format-your-data/recommended-formats/
- Stanford Library - Data best practices and case studies: https://guides.library.stanford.edu/data-best-practices

Identify data value and determine data preservation

During a research project, a great quantity of data is created, and it is not possible to archive/preserve all of it for several reasons:

- Cost related to archiving
- The greater the volume of archived data, the more complicated it is to discover
- Cost of creating and maintaining metadata

Therefore, there is a need to identify data value and determine data preservation. You may be wondering how to evaluate and select what data should be archived?

It depends on many critera:

Criteria related to the research mission (value of the data):

- Requirements of the funder (e.g., SNSF)
- Legal requirements
- Requirements of the publisher
- Requirements of the institution
- The data supports a publication and research results
- The data has a unique character
- The data is related to the notion of intangible cultural heritage
- Originality of the data

Criteria related to the nature of the data:

- Raw data
- Processed data
- Data that supports a publication and research results
- Data that synthesizes research

Criteria related to the type of data:

- Observational data
- Experimental data
- Secondary data
- Negative data

Criteria related to the materials that complete the data:

- Physical samples
- Metadata and documentation
- Software used

For more information

- Library of Vrije Universiteit Amsterdam – Research Data Management Guide – Selecting Data: https://libguides.vu.nl/rdm/selecting-data-archiving
- Digital Curation Centre (DCC) UK - How to Appraise and Select Research Data for Curation: http://www.dcc.ac.uk/resources/how-guides/appraise-select-data
- Médiathèques HES-SO Valais-Wallis - Research data management : https://hevs-ch.libguides.com/RDM

Metadata

It is essential to add appropriate metadata to the stored data to ensure its findability. Metadata is usually described as "data about data".

There are several types of metadata:

- **Descriptive metadata**: title, subject, data creators,...
- **Administrative metadata**: file formats, versions, re-use options,...
- **Structural metadata**: relationships with other entities, list of variables,..

Unlike data documentation - which is intended to be interpretable by humans only - metadata must be machine-readable to allow discovery on the web. Metadata is therefore often converted in XML.

There are tools that can help teams to write their metadata in XML, such as the Datacite Metadata Generator. Metadata is standardized information. There are many schemes, both disciplinary and generic.

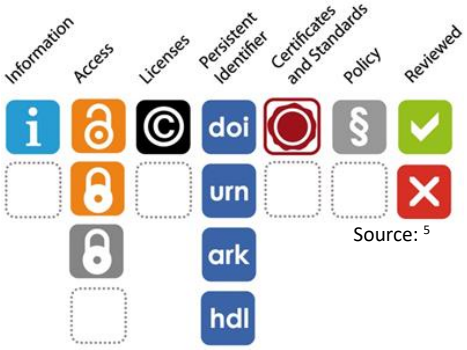
For more information

- NISO - Understanding Metadata: What is Metadata, and What is it For?: A Primer: http://www.niso.org/publications/understanding-metadata-2017
- University of Ediburgh, MANTRA Research Data Management Training – Preparing your data for archiving : https://mantra.ed.ac.uk/preparingyourdataforarchiving/
- University of Texas Libraries - Metadata basics: https://guides.lib.utexas.edu/metadata-basics/key-concepts
- Doranum - Les schémas de métadonnées : https://youtu.be/S-Hw_04ojCc
- Doranum - Datacite Metadata Generator https://doranum.fr/wp-content/uploads/datacite_metadata_generator_4.0.html
- Digital Curation Centre (DCC) UK – Digital curation standards: https://www.dcc.ac.uk/guidance/standards

STEP 5. DATA SHARING

Select a data repository

The submission of datasets to a data repository is important. It ensures the data to be stored in a distinct and separate environment from the scientific article.



Source: [5]

There are several types of repositories: disciplinary, multidisciplinary, institutional, publisher-specific, project-specific, ...

To decide which option was best suited for its project, the research team searched re3data.org, a data repository index that presents the following criteria:

Here is a comparison of three repositories:

Name	**SWISSUbase**	**OLOS**	**Zenodo**
Repository #	1	2	3
Organization	Fors, University of Lausanne and University of Zurich	DLCM	CERN & OpenAire
Servers	Switzerland	Switzerland	Switzerland and Hungary
Field	Social sciences, linguistics, other fields in development	Multidisciplinary	Multidisciplinary (specialized in physics)
Who may archive?	Researchers affiliated to Swiss research institutions and universities	Researchers affiliated to Swiss research institutions and universities	No limitation
Format	Tabular, textual, image, audio, video	Wide range of formats	No indication
License	Choice of licenses (CC)	Choice of licenses (CC)	Choice of licenses (CC)
Persistent IDs	DOI	DOI	DOI
Cost	Free of charge	Fee-based	Free of charge
Advantages	• Architecture based on the OAIS model • Validation of the metadata and documentation by an expert • Deposit and use agreements • Extensive project, dataset and documentation metadata	• Architecture based on the OAIS model • Custom metadata can be added • Choice of sensitivity level using DataTags • Choice of storage duration	• Altmetrics • Possibility to create personal collections • Easy to use
Disadvantages	• Complex submission process • Limited format list • Metadata limited to the social sciences context	• New repository • Not free • Unencrypted storage	• No data and metadata quality control • Limited dataset size • Only the submitter of the dataset is allowed to edit the entry

Source [6-7]

For more information

- [5]OpenAIRE – Select a data repository: https://www.openaire.eu/select-data-repository/
- [6]Melly, Pauline – Comparison OLOS and SWISSUbase (translated): https://sonar.rero.ch/hesso/documents/323070
- [7]Centre de ressources en information scientifique de l'Institut Pasteur – Zenodo: https://openscience.pasteur.fr/2022/12/07/zenodo/
- Doranum – Dépôt de données: https://doranum.fr/depot-entrepots/minute/
- Registry of research data repositories: re3data.org

Determine access conditions

- **Closed data**: A general description of the data is published, but access to the data itself is not possible. E.g., the dataset contains non-anonymized or pseudonymized sensitive data
- **Data on request**: A record of the data is published, and information is provided on how to demand access to the data. The demand is generally sent out to the researchers and approved by them. E.g., the dataset contains data with a high re-identification risk, it can only be shared under certain conditions.
- **Data under embargo**: A record of the data is published, but the data can't be accessed until the end of a specified period. After that period, the data becomes openly accessible E.g., the researcher wants to file a patent before publishing the data
- **Open data**: The data is freely available and can be accessed by anyone

For more information

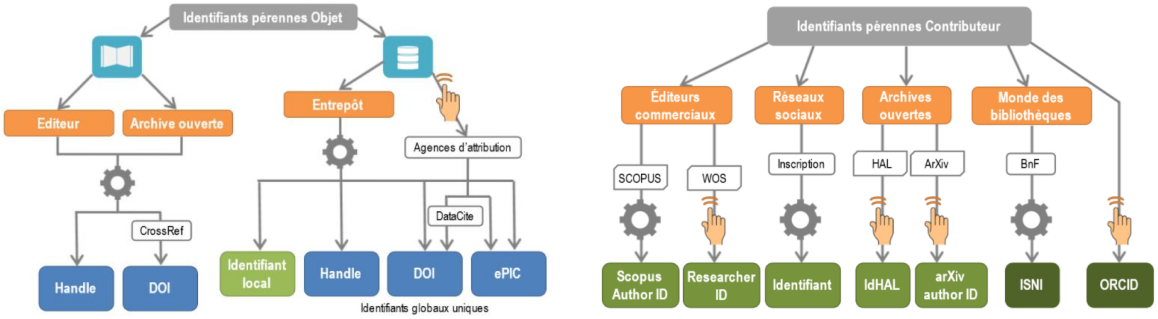
- University of Sidney - Data publication: Controlling access to published data: https://libguides.library.usyd.edu.au/datapublication/access

Choose a persistent identifier (PID)

A PID is a long-lasting digital reference to an object, contributor, or organization. There are two main types of identifiers:

- Object identifiers for publications and data
- Contributor identifiers for authors and organizations

They provide reliable and durable identification. PIDs promote discovery, sharing and reuse and provide a link between the researcher, their institution, their publications and their data.



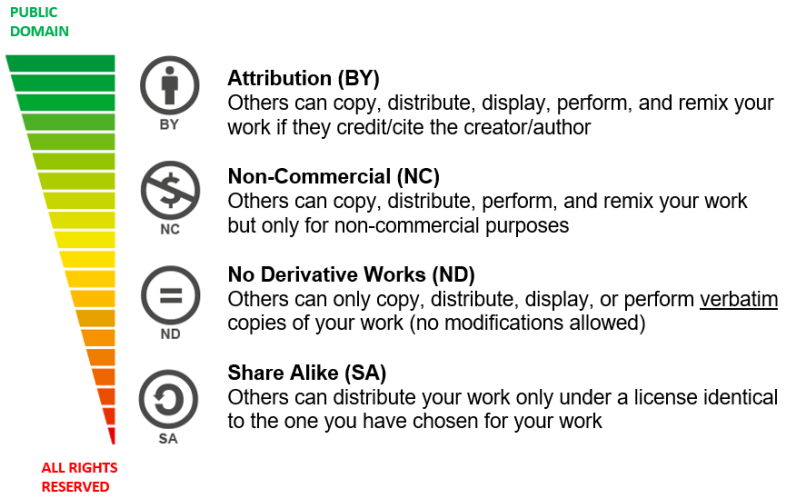
Source : [8]

For more information

- [8]Doranum – Les identifiants pérennes: un aperçu: https://doranum.fr/identifiants-perennes-pid/identifiants-perennes-apercu/
- Doranum – La minute identifiants pérennes : https://doranum.fr/identifiants-perennes-pid/la-minute-identifiants-perennes_10_13143_3jqw-zx31/
- Doranum - Identifiants pérennes : fiche synthétique: https://doranum.fr/identifiants-perennes-pid/identifiants-perennes-fiche-synthetique_10_13143_7gw1-b340/
- ELIXIR Belgium RDM Guide – Identifiers: https://rdm.elixir-belgium.org/identifiers

Assign a copyright reuse license

Creative commons licenses are distribution licenses that allow certain rights of use to be granted in advance on published materials. They are complementary to the author's copyright.



Source : [9]

For more information

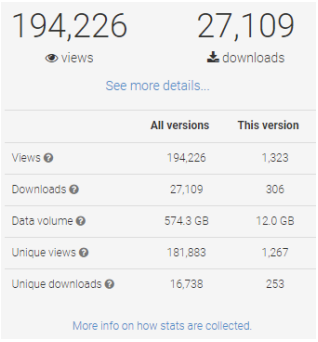
- [9]UQAM – Copyright: https://droit-auteur.uqam.ca/creative-commons/
- Creative Commons Licenses: https://creativecommons.org/
- Graduate Institute Geneva – Copyright: https://libguides.graduateinstitute.ch/droit-dauteur/creative-commons

STEP 6. DATA REUSE

Evaluate data access and reuse

Altmetrics (article-level metrics or alternative metrics) are bibliometric indicators.

They evaluate the impact of a publication or piece of information on the Internet, by observing its dissemination, the actions and interactions it generates in real time (e.g., number of downloads, citations, number of shares on social networks...)



Source: https://zenodo.org/record/5834780#.YeFVQf7MKUl
[accessed January 14, 2022]

Source: https://blog.zenodo.org/2020/07/09/2020-07-09-altmetric-badges/

For more information

- Coop IST – Que sont les altmetrics: https://coop-ist.cirad.fr/evaluer/les-altmetrics/1-que-sont-les-altmetrics
- Altmetric - What are Altmetrics? Capturing the online attention surrounding scholarly content: https://www.altmetric.com/about-altmetrics/what-are-altmetrics/

MORE USEFUL RESOURCES

- DoRANum – Données de la recherche apprentissage numérique: https://doranum.fr
- Data Management Skillbuilding Hub: https://dataoneorg.github.io/Education/
- University of Edinburgh – MANTRA: https://mantra.edina.ac.uk/
- Réseau PORTAGE : Services partagés pour les données de recherche (Association des bibliothèques de recherche du Canada) https://portagenetwork.ca/fr/
- La gestion des données de recherche en 180 secondes https://www.youtube.com/watch?v=FWzw2SqsJFo&ab_channel=Biblioth%C3%A8quesUdeM
- Research data management (RDM) open training materials – Zenodo Community https://zenodo.org/communities/dcc-rdm-training-materials/
- DLCM Research data management MOOC: https://numerique.hes-so.ch/?redirect=0&cid=207
- Médiathèques HES-SO Valais-Wallis - Research data management https://hevs-ch.libguides.com/RDM