

# Scholarly communication in the digital age

A practical introduction to Open Science and Research Data  
Management

**Francesca Di Donato**

 [orcid.org/0000-0003-0144-8934](https://orcid.org/0000-0003-0144-8934)

 @ederinita

[f.didonato@gmail.com](mailto:f.didonato@gmail.com)

# Outline of this course

1. Scholarly communication: an introduction. From the traditional paradigm to Open Science
2. The Open Access publishing paradigm and FAIR (Open) Data
3. **Research Data Management and Data Management Plans**
4. Research Infrastructures, the EOSC, OS skills and training, Science evaluation and OS workflows

3.

# Research Data Management and Data Management Plans\*

\*

The following slides are extracted from E. Lazzeri, F. Di Donato, Research Data Management and DMP - SSH, 10.5281/zenodo.4326563. The number between [ and ] indicates the original slide number. When not indicated, slides are new.

# Data in the SSH.. a fluid definition

## CO-OPERAS GOFAIR

IN organized 5 workshops to gather views and needs of Social Sciences and Humanities researchers regarding the FAIRification of their data and publications.

## CO-OPERAS publishes a variety of workshop reports on FAIRification efforts in the SSH

POSTED ON 28 AUGUST 2020

Arnaud Gingold from the **CO-OPERAS IN** reports:

The CO-OPERAS IN published the reports of all the workshops organized in the past year on the topic of “FAIR data for Social Sciences and Humanities” on Zenodo:

As a first step of its roadmap, CO-OPERAS intended to listen to the communities through a series of workshops held in national languages, in order to facilitate richer and deeper discussions. Between September 2019 and March 2020, CO-OPERAS thus organized five workshops to gather views and needs of Social Sciences and Humanities researchers regarding the FAIRification of their data and publications. The workshops were held in Italy, Portugal, Germany, France and a last international workshop was held in Brussels. The **reports are now available on Zenodo:**

- Turin: <https://doi.org/10.5281/zenodo.3999349>
- Porto: <https://doi.org/10.5281/zenodo.3999344>
- Coimbra: <https://doi.org/10.5281/zenodo.3999357>
- Göttingen: <https://doi.org/10.5281/zenodo.3999330>
- Paris: <https://doi.org/10.5281/zenodo.3999351>
- Brussels: <https://doi.org/10.5281/zenodo.3999359>

# Data in the SSH.. a fluid definition

“DATA IN THE SSH - Some of the proposed definitions:

- data as **a process**: dynamic and diachronic, data represents and preserves a process.
- data as **word**, it provides the basis of knowledge. Distinguish between data and abstraction, on which you can build specialistic analysis. By “process” we should consider a) access to data b) participation in the creation of knowledge
- data are **anything you can formalize through a language**, should “language” be any ontology, mark-up, method that formalize (e.g.TEI encoding schema). Data and metadata are indivisible
- “data” is **misleading**, as nowadays we have devices able to record as “data” anything which used to be unobserved/unimportant. Let’s talk instead of
  - weak documents, mere factual recordings
  - strong documents, whenever a deliberate act generates it

.....

...

- data are “**documents**” only when and if we have a method to collect/obtain
- data as **expression of a method**: if the method was tailored for a specific project/research question, interoperability is at stake
- “paradata” as “upon what method have the data been generated” is an important conceptual step
- data are **functional to information**: it’s something I can provide my interpretation on
- “record” is a more suitable term than data
- data can only exist if anyone has created them: e.g. in archaeology the mere recording seems objective, as you start describing it, subjectivity is immediately involved
- we have a few “data” as interpretation and subjectivity come early into play
- we have a few “**raw data**” as to be analyzed they have to be treated somehow
- ***data are usually discussed in the SSH, while in the hard sciences they often come from a machine, so there is no debate***
- **data and their quality depend also on your research perspective**: you need a perfect reproduction of a manuscript if you need to locate it by the depth/characteristics of the line, whereas if you are only interested in the content any reproduction will do
- books and articles are **research outputs** but at the same time they can be “data” for new researches
- in history knowledge is **created through a cycle non affected by the digital world**; most of the evidence is still analogic. Metadata are crucial to discover it
- in Economics you can find standardized data in primary sources as official financial statements, but whenever you introduce a choice of variables and sources you have to clearly describe your method and choices. If not, data are not reusable....”

# Some conclusions

1. The definition of data is **dependent on the disciplinary area**
2. Data have different natures
3. It is **almost impossible to find a definition on which everybody can agree**

CO-OPERAS, & Borges, Maria Manuel. (2020, August 25). CO-OPERAS - SSH FAIR data wksp report - Coimbra 06122019. Zenodo. <http://doi.org/10.5281/zenodo.3999358>

## *The concept of research data in the SSH is still not unambiguous*

"Participants' input clearly confirmed that at an epistemological level, there is a broader diversity of conceptions on what constitutes data in the SSH. In several contexts, we saw **a certain degree of reluctance towards the term 'data'**: some scholars claim that they do not work with 'data', but with texts, images, etc. This scepticism seemed to be especially prevalent in disciplines adopting often hermeneutic approaches. For these scholars, the term data carries an expression of objectifying and a forceful application of methods, vocabulary and narratives of the natural sciences, which is in principle foreign to them. In other words, it would be in the spirit of a FAIR RDM to recognize that **there is no unified, cross-disciplinary definition of data and that data is also a construct.**"

CO-OPERAS, Bertino, Andrea, & Tóth-Czifra, Erzsébet. (2020, August 25). CO-OPERAS-SSHOC - Research data in the SSH - wksp report - Goettingen 30012020. Zenodo. <http://doi.org/10.5281/zenodo.3999331>



Despite these  
difficulties,  
you need to properly  
manage your  
research data



What  
happens  
if you do  
not do it  
properly?

# Data can get lost

The authors of the study, which is published today in *Current Biology*<sup>1</sup>, looked for the data behind 516 ecology papers published between 1991 and 2011. The researchers selected studies that involved measuring characteristics associated with the size and form of plants and animals, something that has been done in the same way for decades. By contacting the authors of the papers, they found that, whereas data for almost all studies published just two years ago were still accessible, the chance of them being so fell by 17% per year. Availability dropped to as little as 20% for research from the early 1990s.

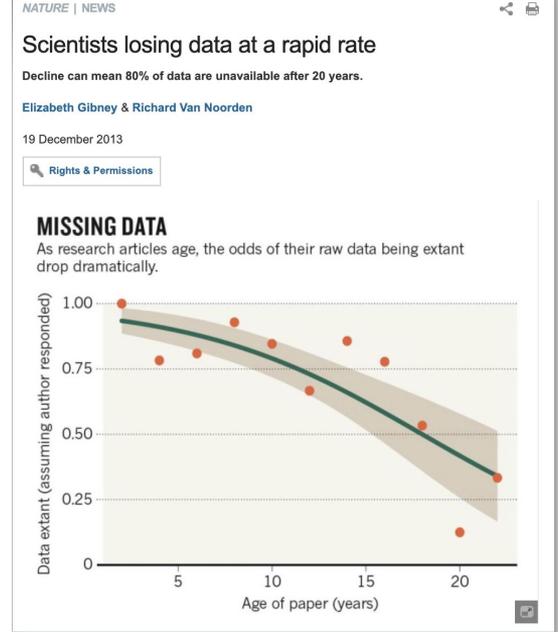
"Most of the time, researchers said 'it's probably in this or that location', such as their parents' attic, or on a zip drive for which they haven't seen the hardware in 15 years," says Timothy Vines, the lead author on the study and an evolutionary ecologist at the University of British Columbia in Vancouver. "In theory, the data still exist, but the time and effort required by the researcher to get them to you is prohibitive."

Another challenge was simply tracking down authors and receiving a response, something at which the team was successful in just 37% of cases. The likelihood of being able to find a working e-mail address, even after an extensive online search, declined by 7% per year. Meanwhile, only around half of the authors with valid addresses responded to the requests, however old the paper.

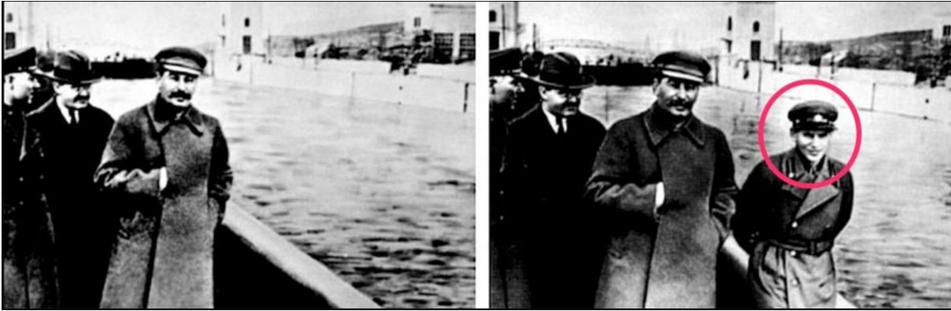
Gibney E. & Richard Van Noorden, Scientists losing data at a rapid rate. Decline can mean 80% of data are unavailable after 20 years, (19 December 2013)

[https://www.nature.com/news/scientists-losing-data-at-a-rapid-rate-1.](https://www.nature.com/news/scientists-losing-data-at-a-rapid-rate-1.14416)

14416



# Data can be manipulated



or altered..

Nikolai Ivanovich Yezhov was head of the People's Commissariat for Internal Affairs until fell from Stalin's favor and power. Among art historians, he also has the nickname "The Vanishing Commissar" because after his execution, his likeness was retouched out of an official press photo; he is among the best-known examples of the Soviet press making someone who had fallen out of favor "disappear".

[The Newseum \(1 September 1999\). "The Commissar Vanishes" in The Vanishing Commissar".](#)

[Archived from the original on 8 February 2007.](#)

# What about data out of its context?

HOME / ATTUALITÀ

## Lilli Gruber furibonda con Matteo Salvini: mette in giro una foto falsa di me e Conte



26 novembre 2020

a a a

NON FANNO NULLA



Scuole senza sicurezza e il governo se ne frega. Così non riapriranno

CASO IN PROCURA



# Science relies on the idea of a substantive debate

Experiment in a nursery school in Scandicci (Firenze) made by Carlo Bernardini, a University professor of physics who in the 80's had a teaching experience.

Bernardini observed that children between 3 and 5 years of age have a much greater non-specific scientificity than adults, for 3 reasons:

1. because they know **how to change their opinion in the face of evidence of facts;**
2. because they know **how to question themselves**
3. and because they have **no interest in altering the experiments.**

Carlo Bernardini, [Adulti e bambini](#), in Carlo Bernardini (a cura di), Il tempo, le cose, la natura, II Ed. 2009.

Daniela Sgobino e Simona Barbetti (a cura di), [La traccia di una innovazione. Esperienze di educazione scientifica nel Centro 0-6 "Torri" di Scandicci](#).

# Research Integrity



Consiglio Nazionale delle Ricerche

IT | EN

Search



Society



Companies



Schools



Researchers



Journalists



Personnel

Biomedical sciences

Earth and environment

Physics and matter

Bio and agri-food

Chemistry and materials technology

Engineering, ICT, energy and transportation

Human sciences and cultural heritage

HOME

ABOUT US ▾

ORGANIZATION ▾

ACTIVITIES ▾

SERVICES AND UTILITIES ▾

NEWS

EVENTS

Home / Organization / Organisms and Committees / Research Ethics and Integrity Committee / Research Integrity

RESEARCH  
ETHICS AND  
INTEGRITY  
COMMITTEE

Documents

Activities

Ethical Clearance

Research  
Misconduct

Research Integrity

## Research Integrity

By Research Integrity we intend the body of principles and ethical values, deontological obligations and professional standards that form the basis of the responsible and correct conduct of those who carry out, finance or evaluate scientific research, as well as the institutions that promote and perform it. The application of principles and values, and the respect for deontology and professional ethics and standards guarantee the quality of the research and enhance the reputation and public image of science, greatly contributing to its advancements and to progress in society.

Below are the most relevant international websites and documents on Research Integrity.

### Websites:

- ENRIO 2020 Congress on Research Integrity Practice
- European Network of Research Integrity Offices (ENRIO)
- USA - Office of Research Integrity (ORI)
- ALL European Academies (ALLEA)
- ENERI
- World Conferences on Research Integrity Foundation (WCRIF)
- Comité d'éthique du CNRS (COMETS)
- German Research Ombudsman

Research Integrity, definition on the CNR webpage: <https://www.cnr.it/en/research-integrity>

# Research Integrity: we have a problem

58 articles published by Diederik Stapel were withdrawn because they were based on **invented data**. His papers had been published in scientific journals considered prestigious (very high IFs!).

Following reports from three doctoral students, the Dutch university for which he worked had started an investigation. Stapel then admitted that he had fabricated the data on numerous occasions.

**If he had shared his data before, he probably wouldn't have been able to fabricate fakes for so long.**

This case led the Netherlands become one of the pioneer countries in Open Science policy and practices

REPORT

## Coping with Chaos: How Disordered Contexts Promote Stereotyping and Discrimination

Diederik A. Stapel<sup>1,\*</sup>, Siegwart Lindenberg<sup>1,2,\*</sup>

+ See all authors and affiliations

Science 08 Apr 2011:  
Vol. 332, Issue 6026, pp. 251-253  
DOI: 10.1126/science.1201068

Article

Figures & Data

Info & Metrics

eLetters

PDF

This article has been retracted. Please see:  
[is retracted by - December 02, 2011](#)

### Abstract

Being the victim of discrimination can have serious negative health- and quality-of-life-related consequences. Yet, could being discriminated against depend on such seemingly trivial matters as garbage on the streets? In this study, we show, in two field experiments, that disordered contexts (such as litter or a broken-up sidewalk and an abandoned bicycle) indeed

Paul Jump, [A Star's Collapse. Dutch begin documenting and trying to explain top psychologist's massive fraud.](#) Times Higher Education,

November 28, 2011

# Data are used by policy makers

Decisions also based on research  
evidence

This is why you should properly  
manage and share your latest  
research results



[Home](#) / [Covid-19 - Situazione in Italia](#)

## Covid-19 - Situazione in Italia



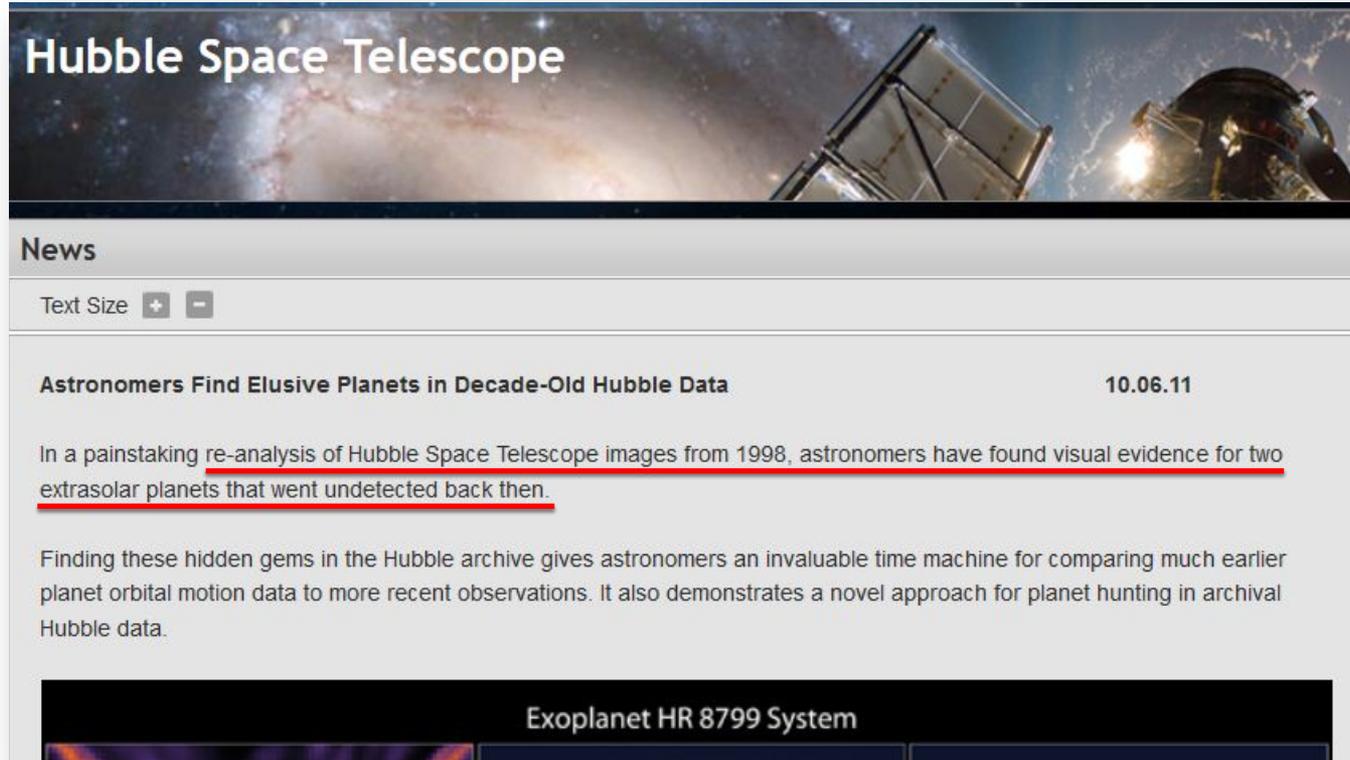
[Home](#) / [Emergencies](#) / [Diseases](#) / [Coronavirus disease \(COVID-19\)](#)

## Coronavirus disease (COVID-19) pandemic

[Advice for the public](#)

[Country and technical  
guidance](#)

# Standing on the shoulders of giants



The screenshot shows a news article from NASA's website. At the top, there is a banner image of the Hubble Space Telescope in space, with the text "Hubble Space Telescope" overlaid. Below the banner is a "News" section with a "Text Size" control. The main article title is "Astronomers Find Elusive Planets in Decade-Old Hubble Data" with a date of "10.06.11". The article text states: "In a painstaking re-analysis of Hubble Space Telescope images from 1998, astronomers have found visual evidence for two extrasolar planets that went undetected back then." Below this is a paragraph: "Finding these hidden gems in the Hubble archive gives astronomers an invaluable time machine for comparing much earlier planet orbital motion data to more recent observations. It also demonstrates a novel approach for planet hunting in archival Hubble data." At the bottom of the article, there is a section titled "Exoplanet HR 8799 System" with a corresponding image.

## Hubble Space Telescope

### News

Text Size

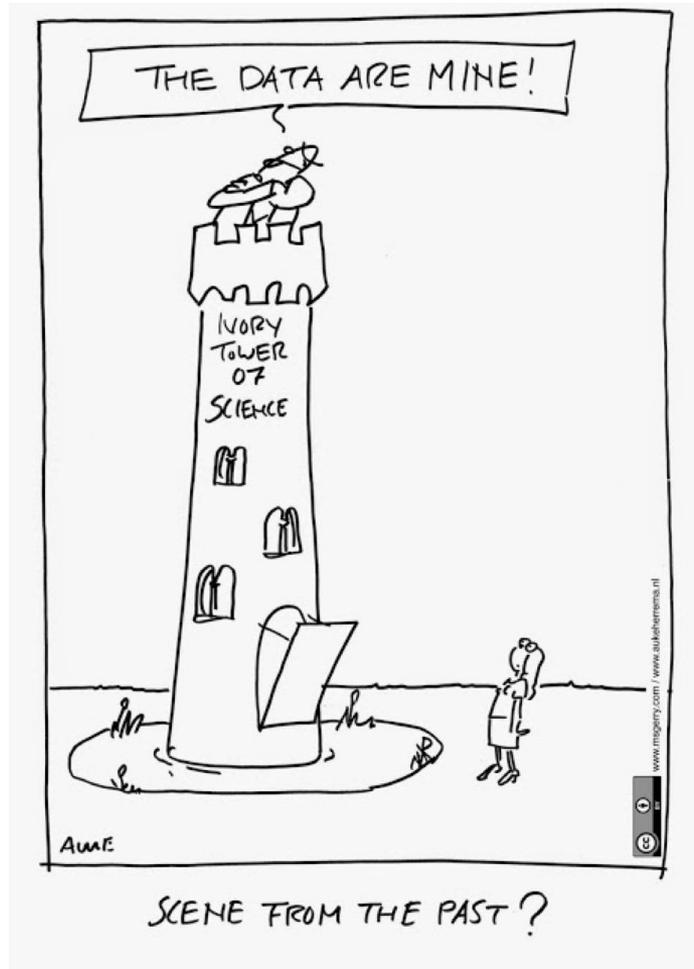
**Astronomers Find Elusive Planets in Decade-Old Hubble Data** 10.06.11

In a painstaking re-analysis of Hubble Space Telescope images from 1998, astronomers have found visual evidence for two extrasolar planets that went undetected back then.

Finding these hidden gems in the Hubble archive gives astronomers an invaluable time machine for comparing much earlier planet orbital motion data to more recent observations. It also demonstrates a novel approach for planet hunting in archival Hubble data.

**Exoplanet HR 8799 System**

Astronomers Find Elusive Planets in Decade-Old Hubble  
Data, [https://www.nasa.gov/mission\\_pages/hubble/science/elusive-planets.html](https://www.nasa.gov/mission_pages/hubble/science/elusive-planets.html)



SCENE FROM THE PAST ?





**Data is not  
yours...that is  
why you  
should take  
care of it!**

# Data is not “yours”!

Data is not intellectual work, it is fact and information

**Copyright protection covers expressions** and not ideas, procedures, operating methods or mathematical concepts as such.

Protection is on databases and not on data. The data are protected only and especially when they are collected and organized in a database.

The sui generis property right (only in Europe) covers not only the reproduction and dissemination of the database, but also the extraction and reuse of substantial parts of the database.

Read complete article [here](#).

A similar content article in English [here](#).

OpenAIRE Guidelines on data protection: available [here](#)



11 Dicembre 2019

## Data governance: un dato non appartiene a nessuno... a meno che sia personale

di [Simone Aliprandi](#)

Quando un nostro dato è personale? Come è giusto tutelarli? La risposta deve comprendere due punti di vista, quello della proprietà intellettuale e quello della privacy.

CONDIVIDI



L'autore



**Simone Aliprandi**

Simone Aliprandi ha un dottorato di ricerca in Società dell'Informazione ed è un avvocato che si occupa di consulenza, ricerca e formazione nel campo del diritto d'autore e più in generale del diritto dell'ICT. È responsabile del progetto [copyleft-italia.it](#), è membro del network Array e collabora come docente con alcuni istituti universitari; ha pubblicato articoli e libri sul mondo delle tecnologie open e della cultura libera, [questo](#) e [questo](#).

## Non c'è solo la privacy, quando si parla di dati e di diritto

Si sente spesso parlare di tutela del dato o titolarità del dato, soprattutto in questi ultimi due/tre anni in cui temi come [big data](#) e open data sono diventati di pubblico dominio e in cui l'entrata in vigore del GDPR ([il nuovo regolamento europeo sui dati personali](#)) ha portato un'ondata – per certi versi ridondante – di corsi di formazione, consulenze, articoli sul tema della protezione dei dati.

Mi occupo di consulenza e formazione proprio in quest'ambito e mi rendo conto che spesso tra gli utenti non c'è piena consapevolezza di come il diritto considera e tratta i dati. Note soprattutto che alle parole tutela del dato o titolarità del dato la gente pensa automaticamente all'ambito della privacy, della tutela del dato personale.

Database definition in the law does not correspond to the scientific one

di alcun tipo di proprietà intellettuale. Questo in virtù di uno dei principi cardine del diritto d'autore secondo cui il diritto tutela non l'informazione, bensì la specifica forma espressiva con cui l'informazione è presentata. Basti leggere il testo dell'articolo 9, numero 2 dell'[Accordo TRIPS](#):

# Be aware...

Managing data is not  
easy

No «one size fits all»

So many aspects to be  
considered

Several tools to use

It takes a lot of time to  
manage data

But, hey, BENEFITS are  
enormous!

# What is the cost of not managing data properly?

Time spent, cost of storage, licence costs, research retraction, double funding, interdisciplinarity and potential economic growth.

Published: 2019-01-16

Corporate author(s): [Directorate-General for Research and Innovation \(European Commission\)](#) , [PwC EU Services](#)  
[Cost of not having FAIR research data](#)



## **Cost of not having FAIR research data**

Cost-Benefit analysis for FAIR research data

***Following this approach, we found that the annual cost of not having FAIR research data costs the European economy at least €10.2bn every year. In addition, we also listed a number of consequences from not having FAIR which could not be reliably estimated, such as an impact on research quality, economic turnover, or machine readability of research data. By drawing a rough parallel with the European open data economy, we concluded that these unquantified elements could account for another €16bn annually on top of what we estimated. These results relied on a combination of desk research, interviews with the subject matter experts and our most conservative assumptions.***

***Moreover, while building on top of other available studies and being heavily reliant on existing material, we have come to realise ourselves how important is to have FAIR research data. Not only the time invested in this study could have been reduced by a significant amount, but the content could have been enhanced if more material had been accessible and reusable.***

Why  
should I  
care?

# You should care because...

- If you manage it, you probably will not **lose** it
- Organising your data will make your work more **efficient**
- Some data is **unique and not reproducible** (data on excavations, annotations on texts)
- By correctly managing your data, you can improve **research integrity**
- By managing your data, you enable **validation and control**
- Someone else could use it in the future to **advance scientific progress**

# An example

Reuse of [DASI - Digital Archive for the Study of pre-Islamic Arabian Inscriptions](#) open data in the frame of the recently ANR-funded project [MAPARABIA \(CNRS-CNR\)](#)

I. ROSSI / A. DE SANTIS

Places of Ancient Arabia. Epigraphic data in the MAPARABIA gazetteer

**Places of Ancient Arabia.  
Epigraphic data in the MAPARABIA gazetteer**  
Epigraphy Info Workshop V - Louvain, November 3rd-6th, 2020

Irene Rossi<sup>1</sup> and Annamaria De Santis  
<sup>1</sup> Consiglio Nazionale delle Ricerche - Istituto di Scienze del Patrimonio Culturale

The Arabian Peninsula is still poorly represented in maps and digital reference tools concerning the geography of the ancient world. The ANR-funded project MAPARABIA (ANR-18-CE27-0013) wants to respond to this need, by supporting the description, discovery, understanding, and process of the consistent amount of data about pre-Islamic Arabia that have been generated over the past years.

Within this frame, a Gazetteer of Ancient Arabia has been conceived. As a thesaurus of ancient "places", the gazetteer focuses on geographic entities provided with at least a name, attested in primary or in secondary sources. It identifies, disambiguates, and describes them, and represents their environmental and semantic relations, as well as their changes over time.

The Gazetteer of Ancient Arabia's conceptual model is very similar to that of the Pleiades gazetteer (<https://pleiades.stoa.org/>): the main entity, the abstract concept of Place, consists of at least a Location, that is the physical manifestation, and/or a Name, mentioned in a Source. Places can be related to each other for geographic or historic-cultural reasons.

The population of the Gazetteer is automatically performed by importing on demand data from DASI, in particular: onomastic items (toponyms, tribe names, names of buildings) from the word-lists generated from the EpiDoc-encoded texts (implementing Name records in the Gazetteer); the related epigraphic sigla (implementing Source records); data on the archaeological sites, which are provenance or place of production of inscriptions, and related monuments (implementing the Location records); some vocabularies, above all those related to periodization.

Synchronization with DASI does not prevent from manually entering new Name, Source and Period records, or implementing the existing ones with new metadata, in order to provide a full description of a Place based on other sources than those catalogued in the DASI archive.

The Gazetteer of Ancient Arabia is not openly available yet. However, access is allowed to authorized users at

Toponyms  $\mu\sigma/\mu$  and  $\mu\sigma/\mu\eta$  in the onomastic word lists of the Corpus of South Arabian Inscriptions in DASI

al-Bayḍ / Nasha place record in the MAPARABIA "Gazetteer of Ancient Arabia" data entry interface: relations with names and relevant sources

# The Stick and the Carrot of RDM

## Requirements

- Compliance with policies (funder, institutional)
- Ensure your data is accessible and shareable (journal publishers requirements)
- Demonstrate responsible practice (improve integrity and validation of results)

## Benefits

- Keep your research safe and secure (think of when you will be writing your paper)
- Increase your research efficiency
- Make your research outputs more visible (curated data facilitate data sharing)
- Enable collaboration (within or outside your discipline)

So...  
What is  
Research Data  
Management?

# What is Research Data Management

Question: *What is Research data management?*

(answer on mentimeter)



Research data management is simply the effective handling of information that is created in the course of research.

[How and why you should manage your research data: a guide for researchers](#)  
[An introduction to engaging with research data management processes.](#)  
[Caroline Ingram, JISC Guides](#)



# Do not harm...

- Managing research data is usually an integral part of the research process, **so you probably already do it.** You only have to **reflect on** and to **improve your strategy.**
- Most of the activities should be familiar:
  - **naming files** so you can find them quickly;
  - keeping track of different **versions**, and deleting those not needed;
  - **backing up** valuable data and outputs;
  - controlling who has **access** to your data.

# Fifty shades of no (to data sharing)

Too expensive

There's no business case

There's no commercial value

It's private

It's secret

It's our data

We have invested a lot of money in this

Link enough data and one will arrive at sensitive private information

It's not data, it's information

It will never work

We don't know how to do this

We don't have the right people to do this

We need the money

It's not ours, and we don't know who's data it is

No idea what the quality of the data is

We don't know where to find it

It's not our job

It isn't in the right format

I am not authorised



Who is going to use this anyway

People are going to misuse it

Image damage for the minister

We are not ready for this

Image loss for Government

The data file is too big

Not enough bandwidth

This is a first step, we will see what we can do later

We can't find it

We have no access

It is out of date

We have it on paper

We don't know if it's legal

Management says no

We never did this before

No value in it

No time / no resources

We will open up (but adapt 90%)

It's incorrect

Commercially sensitive

It is dangerous when linked

People are going to make the wrong conclusions

This is going to start a wrong discussion

We can't say whether we have it or we don't

contracts

Our website cannot hold files this large

It's not ours and we don't have authorization from the data owner

We've already published the data (but it's unfindable/unusable)

People may download and cache the data and it will be out of date when they reuse it

We don't collect it regularly

Too many people will want to download it, which will cause our servers to fail

People would get upset

It's very sensitive information

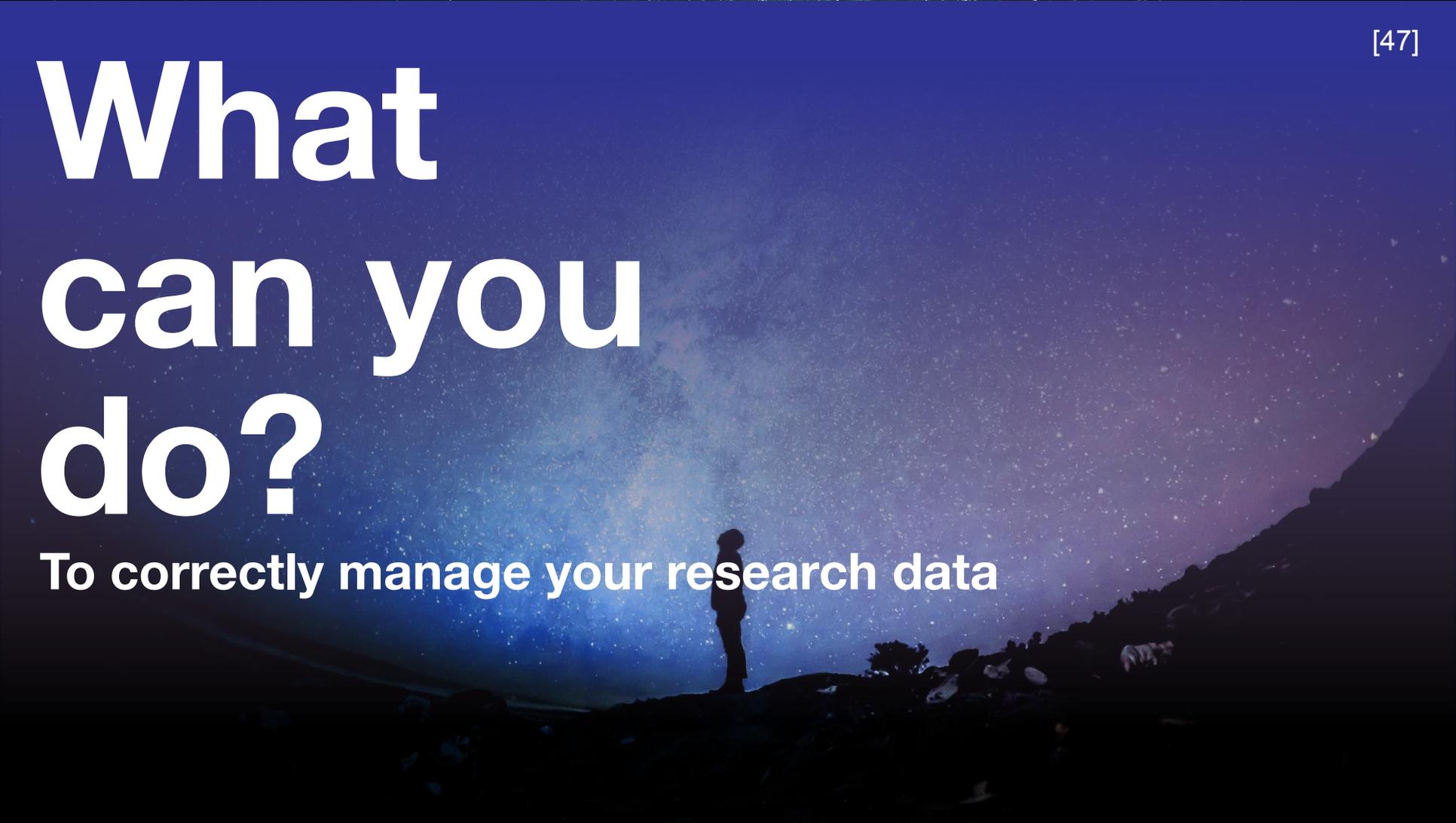
We are not ready for this

Tell us who is going to use it and we will make it open

I do not want to know how to share my data.  
I want to know how I can re-use other's data.

# What can you do?

To correctly manage your research data



# What are your options?

## Your community is already part of a RI

Get engaged, discover and use the RI services, developed tools, infrastructures, repositories, standards, best practices, etc

## Your community could be part of a RI

Get engaged with your reference RI, get your community together and liaise with other communities to understand how to apply existing best practices in your specific domain

## Your Institution could set up a data strategy

At the level of your Institution, several actions can be taken. You can learn from others' experience: set up a working group, adopt a data policy, design and develop a data registry/repository, give access to your contents



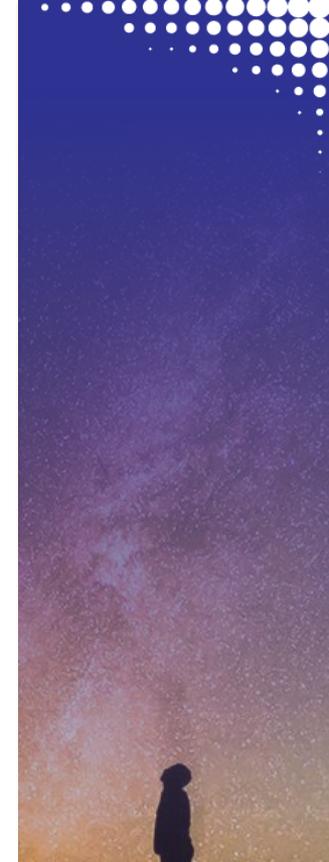
## None of the above

You can still apply the Open Science generic best practices that you will learn in this course to your projects and research workflow

Do not reinvent the wheel

Build on others' experience

Liaise with your and other communities



# Ask for Support



## OpenAIRE

Get support on  
general open  
science practices

[road-it@openaire.eu](mailto:road-it@openaire.eu)

[www.openaire.eu](http://www.openaire.eu)



## Research Infrastructures

Get support from your  
specific domain

[Research  
Infrastructure  
www.esfri.eu](http://www.esfri.eu)



## Competence Center

Be supported at  
national level from a  
team of experts in  
Open Science and  
EOSC related fields

Coming soon:

<https://www.icdi.it/it/attivita/tf-cc>



## Your Institution

In case your institution  
has a structured open  
science strategy and  
dedicated support!



# Management Planning



# Data Management Plan

- It is a **plan to manage the data within your project** (for every kind of project)
- It is a **living document** (so it needs updates)
- Since it is a plan, you should think it from the very beginning (ideally, **it is included in your project planning**)
- **Clear rules**, less mistakes from the beginning
- It is a **structured way to think of your data**

# What is normally a DMP about?

- **Identify** the data you are working with in your project
- **Decide the strategy** to organise your data and the standards you will use
- **Daily data management**
  - What is your plan for sharing your data?
  - Will you have issues sharing your data?
  - Will you need more resources/budget than expected?

# 1. Administrative data

- **Basic information** about your project: title, acronym, ID, reference numbers
- An **abstract** of your project highlighting the scope of data collection/creation
- Details related to **procedures and policies**

## Assignment 1- Create a DMP

(Cees, Elise, Francesca)

### Case

The project 'Veteran tapes' is allowed to start. The funding has been awarded and the funder expects a data management plan to be prepared.

### Project proposal summary:

Since the end of the Second World War, the Netherlands has participated in several missions abroad. "The navy, army, air force and military police help preserve peace and justice all over the world", said the Ministry of Defence.

The project 'Veteran tapes' is to conduct 600 interviews with (former) soldiers who have taken part in missions. This will record experiences that are of great importance to military-historical research, as the personal experiences in their own words, are a necessary addition to the documentation by the Ministry. Trial interviews show both a high degree of reflection on the mission, its societal reception and the role played; this often proves to be a sensitive matter. This also makes it relevant material for e.g. social science disciplines.

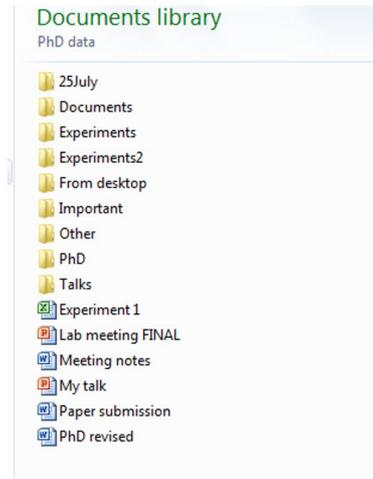
### Design and definition of the project:

- Interviewing 600 people; video-taped.
- 20 people who are doing the interviewing; alignment is important.
- Structured interviews to enable comparison.
- The working language is Dutch.
- The people being interviewed will have a chance to see the video and can decide whether to show the interview freely, or to use it for research purposes only.
- The project will run for two years.
- At the end there will be a portal where the videos can be viewed when allowed to be accessed publicly. It is the intention to make all interviews available for research purposes.
- The audio stream (speech) will be transcribed. In doing so, it will be possible to search for interviews and to refer to illustrative passages.

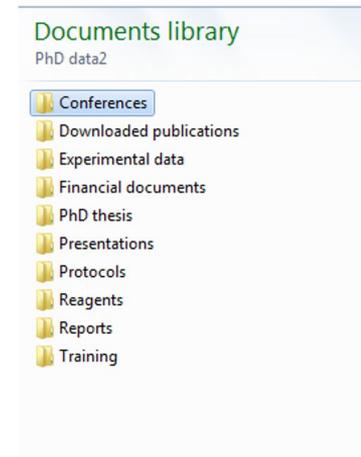
# 2. Data Collection

- Are you **using existing data**?
- What are the **standards or methodologies** that you will use to **collect** your data?
- Data **Formats** and Software
- How will you **structure and name** your files and folders?

Example-A¶



Example-B¶



Nome ↑	Proprietario	Ultima modifica
 00 - Assignments	io	25 ott 2018 io
 0 - My notes	io	25 ott 2018 io
 1 - Definitions	io	25 ott 2018 io
 2 - Planning phase	io	25 ott 2018 io
 3 - The research phase	io	25 ott 2018 io
 4 - User phase	io	25 ott 2018 io
 5 - Legislation and policy	io	25 ott 2018 io
 6- Data support	io	25 ott 2018 io
 2018-10-11 - First meeting material	io	25 ott 2018 io
 2018-11-21 - Second meeting material	io	25 ott 2018 io
 DMP examples	io	25 ott 2018 io
 Miscellanea	io	14:33 io
 000- Readme	io	26 ott 2018 io

# A good example: TILS document naming convention

## 3. Version

(upper case, max 4 chars, optional)

For documents that will continue in various versions use V followed by the version number. Use an underscore to indicate a decimal point if necessary.

Eg. PMF\_PRP\_ZenMonkeyProject\_V2\_20090607.docx

New versions should not be created for each iteration of the document, but rather at significant changes or when it has been reviewed or changed by another author.

Document naming for the TILS Division should follow this convention:

**GDL\_TILSDocNaming\_V1\_20090612.docx**

A prefix shows the document type

The document title describes the content

The version number

The date in the format yyyymmdd

Prefix	Meaning
AGD	Agenda
AGR	Agreement
GDL	Guideline
MEM	Memorandum
MIN	Minutes and Notes
PRE	Presentation
PRO	Procedure
PRP	Proposal
REP	Report
TEM	Template

## 2. Document title/ Description

(mixed case, max 30 chars, **no spaces**)

- Describes the purpose or “business” of the document. Acronyms, capitalisations, abbreviations can be used, keep in mind that descriptions should be **meaningful** to anyone reading the file name.
- In the case of project documentation use the **project name** or its usual abbreviation
- If possible Departmental Branch and/or Section should be integrated into this field to indicate origin / ownership of document.
- Use only alpha-numeric characters, plus the hyphen and underscore.
- **Do not use spaces.**

# 3. Metadata and supporting documentation

- Which **documentation and metadata** will support/describe your data?
- How will you create the supporting documentation and metadata?
- Which **metadata standards** will you use?

## Metadata Provision and Standards

The following types and levels of metadata will be produced and archived:

*Study-Level Metadata Record.* A summary record in an agnostic format will be created for inclusion in the searchable TRIPLE online catalogue. This record will be indexed with terms from the TRIPLE Thesaurus to enhance data finding.

*Data Citation with Digital Object Identifier (DOI).* A standard citation will be provided to facilitate attribution. The DOI provides permanent identification for the data and ensures that they will always be found at the URL specified.

*Variable-Level Documentation.* TRIPLE will tag variable-level information in the most relevant open standards for SSH i.e. in Data Documentation Initiative (DDI), Text Encoding Initiative (TEI), Metadata Encoding and Transmission Standard (METS), Metadata Object Description Schema (MODS).

Metadata records produced by TRIPLE will be published using the following standard vocabularies: *Component MetaData Infrastructure, Dublin Core Metadata Element Set and DCMI Metadata Terms.* Moreover, metadata records published in RDF will use the following linked open data vocabularies: *Data Catalog Vocabulary (DCAT), Open Digital Rights Language (ODRL), DDI-RDF Discovery Vocabulary (Disco).*

Annotation data will be published in compliance with W3C Web Annotation standards (data model, vocabulary and protocol).

Khan, Anas Fahad; Dumouchel, Suzanne; Gingold, Arnaud;  
Blotière, Emilie; Moranville, Yoann, TRIPLE Deliverable: D1.3  
Data Management Plan: <https://zenodo.org/record/4438626>

# So many ways to describe your data

How to create useful README files: <https://data.research.cornell.edu/content/readme>



```
Cornell AUTHOR_DATASET_ReadmeTemplate.txt

This DATSETNAMEreadme.txt file was generated on [YYYYMMDD] by [Name]

-----
GENERAL INFORMATION
-----

1. Title of Dataset

2. Author Information

Principal Investigator Contact Information
Name:
Institution:
Address:
Email:
```

# Discipline specific standards

Showing records 1 - 50 of 1482.

View as Table   View as Grid

Sort by: Domain

Recommended Records: Recommended

Associated Publication? No Publication   Has Publication

Claimed? No Maintainer   Has Maintainer

Record Status: Uncertain   Deprecated   In developer   Ready

Standard Type: Terminology Artifact   799

Registry	Name	Abbreviation	Type	Subject	Domain	Taxonomy
	Amphibian-Taxonomy Ontology	ATG	Standard	None		
	Traditional Medicine Constitution Value Set	TM-CONST	Standard		None	
	Traditional Medicine Other Factors Value Set	TM-OTHER-FACTORS	Standard		None	
	Ontology of Alternative Medicine, French	OntoMA	Standard		None	
	UniProt Taxonomy	UP Taxonomy	Standard			

### Metadata

RD4 | Metadata Directory

Edit this page

View the standards  
View the extensions  
View the tools  
View the use cases  
Browse by subject areas

Contribute  
Add standards  
Add extensions  
Add tools  
Add use cases

github  
 @twitter  
 linkedin

### Arts and Humanities

**Encoded Archival Description (EAD)**

A standard for encoding archival finding aids using XML in archival and manuscript repositories, implementing the recommendations of the International Council on Archives (ICA)(G); [General International Standard Archival Description](#).

**TEI - Text Encoding Initiative**

The guidelines of the Text Encoding Initiative make recommendations about suitable ways of representing those features of textual resources which need to be identified explicitly in order to facilitate processing by computer programs. They specify a set of XML tags in order to mark the textual metadata, text structure, relationship between images and transcriptions and other features of interest. The guidelines both describe the formal TEI encoding scheme, and document the TEI markup vocabulary. In their 30 years community driven development they have developed to a de-facto standard in the production of textual data in the humanities.

**DDI (Data Documentation Initiative)**

A widely used, international standard for describing data from the social, behavioral, and economic sciences. Two versions of the standard are currently maintained in parallel:

- DDI Codebook (or DDI version 2) is the simpler of the two, and intended for documenting simple survey data for exchange or archiving. Version 2.5 was released in January 2014.
- DDI Lifecycle (or DDI version 3) is richer and may be used to document datasets at each stage of their lifecycle from conceptualization through to publication and reuse. It is modular and extensible. Version 3.2 was published in March 2014.

Both versions are XML-based and defined using XML Schemas. They were developed and are maintained by the DDI Alliance.

**MARC (Machine-Readable Cataloging)**

MARC is a standard and serialization format for representing bibliographic metadata, originally designed as a way of exchanging bibliographic records between library catalogs. Various different versions have been defined, mostly with national or regional scope, of which MARC 21 is probably the most widely used. There also exists an XML serialization of MARC 21, known as MARCXML.

**MIDAS-Heritage**

<https://fairsharing.org/>



### Metadata

RD4 | Metadata Directory

Edit this page

View the standards  
View the extensions  
View the tools  
View the use cases  
Browse by subject areas

Contribute  
Add standards  
Add extensions  
Add tools  
Add use cases

github  
 @twitter  
 linkedin

### Social and Behavioral Sciences

**DDI (Data Documentation Initiative)**

A widely used, international standard for describing data from the social, behavioral, and economic sciences. Two versions of the standard are currently maintained in parallel:

- DDI Codebook (or DDI version 2) is the simpler of the two, and intended for documenting simple survey data for exchange or archiving. Version 2.5 was released in January 2014.
- DDI Lifecycle (or DDI version 3) is richer and may be used to document datasets at each stage of their lifecycle from conceptualization through to publication and reuse. It is modular and extensible. Version 3.2 was published in March 2014.

Both versions are XML-based and defined using XML Schemas. They were developed and are maintained by the DDI Alliance.

**MIDAS-Heritage**

A British cultural heritage standard for recording information on buildings, archaeological sites, shipwrecks, parks and gardens, battlefields, areas of interest and artefacts.

Sponsored by the Forum on Information Standards in Heritage, MIDAS Version 1.1 was released in October 2012.

**OAI-ORE (Open Archives Initiative Object Reuse and Exchange)**

The goal of these standards is to encode the rich content in aggregations of Web resources to applications that support authoring, deposit, exchange, visualization, reuse, and preservation. The standards support the changing nature of scholarship and scholarly communication, and the need for cyberinfrastructure to support that scholarship, with the intent to develop standards that generalize across all web-based information including the increasing popular social network of "Web 2.0".

**QUDEx (Qualitative Data Exchange Format)**

The QUDEx standard is an archive-ready format for qualitative data that preserves annotations of, and relationships between, data and other related objects. It can be viewed as the optimal baseline data exchange model for the archiving and interchange of data and metadata.

**SDMX (Statistical Data and Metadata Exchange)**

A set of common technical and statistical standards and guidelines to be used for the efficient exchange and sharing of statistical data and metadata.

Sponsoring institutions include BIS, ECB, ELIPSTAT, IMF, OECD, UN, and the World Bank. Technical Specification 2.1 was amended in May 2012.

### General Research Data

<http://rd-alliance.github.io/metadata-directory/>

# 4. Legal and ethical aspects

- Did you asked for an **informed consensus** to share the data and preserve them?
- How will you **protect personal data**?
- How about **data licencing**?



Attribution  
CC BY



Attribution-ShareAlike  
CC BY-SA



Attribution-NonCommercial  
CC BY-NC



Attribution-NoDerivatives  
CC BY-ND



Attribution-NonCommerical-ShareAlike  
CC BY-NC-SA



Attribution-NonCommercial-NoDerivatives  
CC BY-NC-ND

# 5. Data storage and backup

- Do you have enough space to **store your data** or should you include costs for additional services? Will the services be reliable/trusted?
- How will you **share your storage/backup** with your collaborators?
- Will you use **cloud solutions**?
- Will you **back your data up**? How?

**Do not leave it all to Google...**

Google services Terms of Use:

---

When you upload, submit, store, send or receive content to or through our Services, you give Google (and those we work with) a worldwide license to use, host, store, reproduce, modify, create derivative works (such as those resulting from translations, adaptations or other changes we make so that your content works better with our Services), communicate, publish, publicly perform, publicly display and distribute such content. The rights you grant in this license are for the limited purpose of operating, promoting, and improving our Services, and to develop new ones. This license continues even if you stop using our Services (for example, for a business listing you have added to

# You have better alternatives...



<https://cryptpad.fr/>

# ..and also

SWITCH

Services ▾

Stories ▾

About us ▾

Services → Share and sync files

## SWITCHdrive: store and share files online

Cloud-based storage services synchronise files automatically across several devices and make it easier to share them with other users. SWITCHdrive gives university members 50 GB of storage space. All files are stored securely in Switzerland.

SWITCHdrive offers the Swiss academic community a secure alternative to commercial cloud storage services. Files can be stored, synchronised, shared and worked on in collaboration with others – quickly and reliably. SWITCHdrive is easy to use with a browser, desktop client or mobile app. The academic cloud storage services differs from the commercial ones in one important respect: it runs entirely in the SWITCH cloud. It is connected to the academic network, and access is protected by AAI. All the IT resources and users' files are stored in SWITCH's data centres. This is the best way to meet university members' security needs.

<https://www.switch.ch/services/drive/>

SURF DRIVE

Home

Downloads

Tutorials

FAQ

About SURFdrive

Contact

## Personal cloud storage service for Dutch education and research



Log in to SURFdrive

### Why SURFdrive?

#### Secure file storage

Log in with your institutional account and obtain 250 GB right away.

#### Access anywhere, no matter where you are

Access to your files anywhere and anytime: from your smartphone, your tablet or your laptop.

### Latest news

17 NOV [Updates to SURFdrive](#)

17 NOV [Setting up WebDAV passwords](#)

09 OCT [From now on up to 250 gigabytes of storage](#)

[All news items](#)

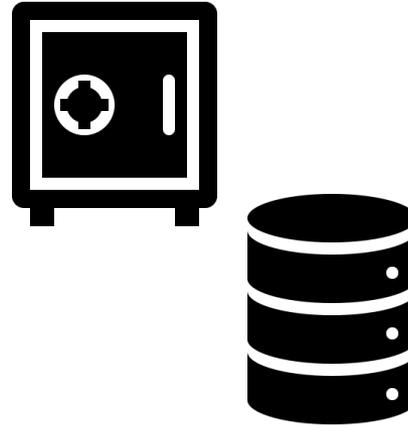
<https://www.surf.nl/en>



Your institution probably provides better alternative  
Ask your IT for support

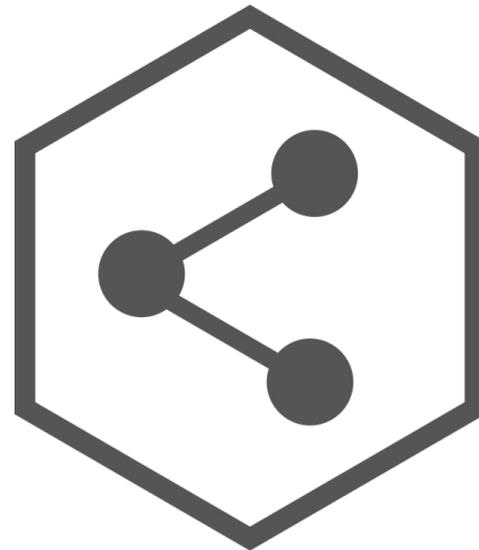
# 6. Select and preserve

- Which data shall be **preserved or destroyed** due to contractual, legal or administrative reason?
- What are the **envisaged uses** of your data for research purposes?
- Which data shall be preserved and potentially shared?
- What is your **long term preservation strategy**?
- Did you consider the **effort and costs** to prepare your data for sharing and preservation?



# 7. Data sharing

- **Who** will you share your data **with**?  
Under which **conditions**?
- **When** will you share your data?
- Will you need to apply any **access restriction**?
- Which actions do you foresee to avoid or reduce access restrictions?
- How will your potential users **find** your data?



# Funder policies on Data sharing

Sherpa Juliet

Browse

Search

Statistics

Our APIs

Suggest

Admin

## Search

Please enter a name or acronym of a funder.

Funder Name

Search

This quick search will find any items whose name or acronym (in any language) match any of the words entered.

# 8. Responsibilities and resources

- Who is **responsible to implement and revise the DMP?**
- How will you share responsibilities among partners in collaborative projects?
- Which resources will you need to implement your DMP?
- Will you need any specific **external experties or tools?**

# European Commission and H2020

A DMP describes the data management life cycle for the data to be collected, processed and/or generated by a Horizon 2020 project. As part of making research data findable, accessible, interoperable and re-usable (FAIR), a **DMP should include information** on:

- the handling of research data during & after the end of the project
- what data will be collected, processed and/or generated
- which methodology & standards will be applied
- whether data will be shared/made open access and
- how data will be curated & preserved (including after the end of the project).
- A DMP is required for all projects participating in the extended ORD pilot, unless they opt out of the ORD pilot. However, projects that opt out are still encouraged to submit a DMP on a voluntary basis.

# Towards Horizon Europe

## 1.2. Provisions on Research Data Management (RDM)

[81]

Horizon Europe	Horizon 2020
The governing principle will be to manage data responsibly, in line with FAIR and under the principle “as open as possible, as closed as necessary” (which means that data can remain closed provided there are good reasons for this, such as IP protection, security, etc.). In Horizon Europe the emphasis shifts from open research data to research data management.	The same governing principles are applying in H2020.
Projects generating research data cannot opt out from RDM. <sup>1</sup>	In H2020, it is possible to opt out (partially or entirely) from the Open Research Data pilot at any stage before or after signing the grant agreement.
All projects collecting or using data will have to update regularly the data management plan (DMP).	This is also strongly recommended in H2020, but not mandatory.
Beneficiaries will have to deposit data in a trusted repository. Valid repositories will be those that provide persistent identifiers for the data, and ensure rich metadata in line with FAIR. For some actions, there will be an obligation to deposit in a repository that is federated under the European Open Science Cloud (EOSC).	The specification of the characteristics of valid repositories is not present in H2020, and there was no reference to EOSC.
Beneficiaries will have to deposit and to ensure OA to data as soon as possible (as per DMP) and under CC-BY or CC0 or equivalent, unless exceptions apply that are duly justified in the DMP.	In H2020 these licenses are only recommended.
Information should be provided via the repository about any other research output or tool or instrument needed to re-use or validate the data, unless justified legitimate concerns/interests need to be safeguarded.	This was also requested under the H2020 Open Research Data pilot.
Costs for RDM will be eligible, but only during the duration of the project.	The same applies in H2020.

# FAIR self assessment

<https://fairaware.dans.knaw.nl/>

# DMP online

Guida di Elena Giglia: [Come scrivere un data management plan](#)  
e E. Giglia, [DMPonline passo dopo passo](#)



questions?

(answer on mentimeter)

# References

Aliprandi S., Data governance: un dato non appartiene a nessuno... a meno che sia personale, 11 Dicembre 2019, [https://www.apogeeonline.com/articoli/data-governance-un-dato-non-appartiene-a-nessuno-a-meno-che-sia-personale-simone-aliprandi/?fbclid=IwAR2LaXV6b6m8aaJ7SeFKVAidbTIB9C4\\_-mQPirLxs3m5DxAs0jG8gkn8ZKg](https://www.apogeeonline.com/articoli/data-governance-un-dato-non-appartiene-a-nessuno-a-meno-che-sia-personale-simone-aliprandi/?fbclid=IwAR2LaXV6b6m8aaJ7SeFKVAidbTIB9C4_-mQPirLxs3m5DxAs0jG8gkn8ZKg)

Carlo Bernardini, Adulti e bambini, in Carlo Bernardini (a cura di), Il tempo, le cose, la natura, II Ed. 2009.

CO-OPERAS, Bertino, Andrea, & Tóth-Czifra, Erzsébet. (2020, August 25). CO-OPERAS-SSHOC - Research data in the SSH - wksp report - Goettingen 30012020. Zenodo. <http://doi.org/10.5281/zenodo.3999331>

CO-OPERAS, & Borges, Maria Manuel. (2020, August 25). CO-OPERAS - SSH FAIR data wksp report - Coimbra 06122019. Zenodo. <http://doi.org/10.5281/zenodo.3999358>

CO-OPERAS, Giglia, Elena, CO-OPERAS - SSH FAIR data - wksp report - Turin 10092019, <https://doi.org/10.5281/zenodo.3999349>

Directorate-General for Research and Innovation (European Commission) , PwC EU Services, Cost of not having FAIR research data, 2019-01-16

Gibney E. & Richard Van Noorden, Scientists losing data at a rapid rate. Decline can mean 80% of data are unavailable after 20 years, (19 December 2013) <https://www.nature.com/news/scientists-losing-data-at-a-rapid-rate-1.14416>

Caroline Ingram, How and why you should manage your research data: a guide for researchers. An introduction to engaging with research data management processes. JISC Guides

Khan, Anas Fahad; Dumouchel, Suzanne; Gingold, Arnaud; Blotière, Emilie; Moranville, Yoann, TRIPLE Deliverable: D1.3 Data Management Plan: <https://zenodo.org/record/4438626>

Ignasi Labastida & Thomas Margoni, Licensing FAIR Data for Reuse, 2019, <http://eprints.gla.ac.uk/203694/1/203694.pdf>

# References

OpenAIRE Guidelines on data protection: <https://www.openaire.eu/how-do-i-know-if-my-research-data-is-protected>

I. Rossi / A. De Santis, Places of Ancient Arabia. Epigraphic data in the MAPARABIA gazetteer, [http://epigraphy.info/workshop\\_5\\_posters.html](http://epigraphy.info/workshop_5_posters.html)

Daniela Sgobino e Simona Barbetti (a cura di), [La traccia di una innovazione. Esperienze di educazione scientifica nel Centro 0-6 “Turri” di Scandicci.](#)

Paul Jump, A Star's Collapse. Dutch begin documenting and trying to explain top psychologist's massive fraud. Times Higher Education, November 28, 2011

## Links

[Astronomers Find Elusive Planets in Decade-Old Hubble Data,](#)  
[CO-OPERAS GOFAIR IN](#)

[DASI - Digital Archive for the Study of pre-Islamic Arabian Inscriptions](#)  
[ESFRI](#)

[Fifty Shades of No:](#)

[Google services Terms of Use](#)

[ICDI - Task Force Competence Center](#)

[MAPARABIA \(CNRS-CNR\)](#)

The Newseum (1 September 1999). ““The Commissar Vanishes” in [The Vanishing Commissar](#)”.

[OPENAire](#)

[README files template](#)

[RDA Metadata Standards Directory](#)

[Research Infrastructure](#)

[Research Integrity. CNR page](#)

[SWITCHDrive](#)

[SURE](#)

[TILS Document Naming Convention:](#)

## Credits

Slide 18: Source:

<https://www.fosteropenscience.eu/node/688>

Slide 19:

<https://www.fosteropenscience.eu/content/cartoon-data-sharing>