



Corso di formazione a distanza  
Scienza aperta e gestione dei dati  
per le scienze umane e  
del patrimonio culturale


20 maggio – 9 giugno 2021

Photo by Martin Adams on Unsplash

# Gestione dei dati

CNR, maggio  
2021

Elena Giglia  
Università di Torino  
elena.giglia@unito.it

 @egiglia



In questo modulo impareremo:

1. come gestire i propri dati correttamente e rendere la ricerca più efficace

2. la differenza fra dati FAIR e Open

## MESSAGGI CHIAVE

- gestire bene i dati è nell'interesse di chi fa ricerca
- solo dati gestiti bene possono essere resi FAIR e se possibile aperti
- NON ci sarà una ricetta per DMP, ma strumenti utili (da imparare)

[DMP]

# ... dati e scienze umane??



Are artistic/humanities disciplines/methodologies only considered valid if we call them 'data'?  
In other words, can we not accept different kinds of research methodologies as valid \*on their own terms\*, rather than on those imported from (or imposed by) other fields? #munin2018



USIAMO I DATI ANCHE  
SE NON LO SAPPIAMO.

UNA VOLTA  
REALIZZATO, LA  
VOSTRA RICERCA  
CAMBIERÀ PER  
SEMPRE

In the humanities, we all use research data, although we may not be aware of it. It is like in the case of Monsieur Jourdain, the title character of Molière's *Le Bourgeois gentilhomme*, who learnt, to his great satisfaction, that unwittingly he had been speaking prose all his life. With research data in the humanities it is exactly the same: you are using it, even if you don't know it, and once you realise it, it will affect your research workflow forever.

<http://www.insideoutproject.com>

« On ne sait pas  
ce que nous réserve  
l'avenir,  
alors profitons »

<http://www.insideoutproject.com>

« Laissez  
être heure

# Dati e scienze umane?



Anthropologists. One reason for the adoption of the term could be that in the humanities, “[w]e resist the blanket term ‘data’ for the very good reason that we have more and precise terminology (e.g. primary sources, secondary sources, theoretical documents, bibliographies, critical editions, annotations, notes, etc.) available to us to describe and make transparent our research processes” (Edmond & Tóth-Czifra, 2018:1). The resistance to ‘data’ in the humanities, as an oversimplifying abstraction of complex phenomena, was voiced by many critics, most notably by Johanna Drucker (2011), who opposed the objectifying term ‘data’ (something given) and proposed to use ‘capta’ (something captured, taken) instead. This

fact that data in the humanities are also an effect of operationalisation and interpretive processes.

LE «FONTI» SONO DATI  
C'È SEMPRE INTERPRETAZIONE  
PROBLEMI CON I DIRITTI

## What is Data, Anyway?

- Examples for Humanities data: primary sources (texts, pictures), secondary sources, theoretical texts, digital tools (software), annotations, etc.
- most “sources” are research data and their management has in fact always been part of the scientific process, digitization only adds complexity
- digitized sources, born digital sources
- various formats and types (pictures, texts, multimedia, measurements, etc.) [Wuttke 2018](#)

## Are Humanities and Cultural Heritage data special?

- Yes and No!
- Humanities are a very broad research discipline, many specific research contexts, but also increasingly interdisciplinary research
- Humanities research lives from enrichment of data (layers of interpretation)
- Problematic to distinguish between primary data (raw data) and secondary data
- Issues with ownership of the data (cultural heritage institutions, publishers)
- But: Many issues and solutions apply to the broader field (and beyond Humanities and Heritage Science!)

# Cosa sono i dati?

DATI NON SONO MAI  
«GREZZI»

DATI SONO SEMPRE  
ESPRESSIONE DI UN  
METODO

C'È SEMPRE UNA SCELTA  
[METODOLOGICA,  
EPISTEMOLOGICA,  
POLITICA...]

C'È SEMPRE  
INTERPRETAZIONE  
[DATI NON SONO GENERATI  
DA UNO STRUMENTO]

DATI SONO DISCUSSI

ANCHE I LIBRI SONO DATI PER  
FUTURE INTERPRETAZIONI

DATI=TUTTO CIÒ CHE PUÒ  
ESSERE FORMALIZZATO  
CON UN LINGUAGGIO

MEGLIO «RECORD»  
DI DATO?

DATI= DOCUMENTI  
- DEBOLI (MERE  
REGISTRAZIONI)  
- FORTI (INTERVENTO  
UMANO)

DATI SONO UN PROCESSO,  
DINAMICO E DIACRONICO

# Perché trattarli come «dati»?



Finally, let us briefly consider the issue of why it is so important that we recognise our assets as research data and act according to the guidelines presented in the next sections. Let us conclude with an example. In *Graphs, Maps and Trees*, Franco Moretti (2005) discusses a complex evolution of British novelistic genres (1740-1900), using an elaborate graph to show that the lifecycle of most of those genres spanned only one generation. Yet, in order to come up with this conclusion, he had to amass and collate the research material previously collected and presented - not always explicitly - in numerous analytical studies on individual genres. So, the availability of data generated in multiple previous studies allowed entirely new insights to appear. This demonstrates the huge potential of the dispersed, lower-scale data that are sitting in our publications or hard-drive folders, which - when made accessible and aggregated - can open the path to new, original research and could be reused by others.

RESI ACCESSIBILI E  
AGGREGATI APRONO  
A NUOVE RICERCHE  
INEDITE E POSSONO  
ESSERE RIUSATI

# Perché?



## RECOMMENDATIONS

» Think of all your research assets as research data that could be potentially reused by other scholars. Consider how useful it would be for your own work if others shared their data.

» Familiarise yourself with the FAIR Data Principles before you start collecting data and building corpora e.g. [FORCE11: the FAIR Data Principles](#), [GO-FAIR: FAIR Data Principles](#) and discuss with colleagues and experts to build a better understanding.

» Digitally document all your research and data collection work -- at the beginning of a project it is difficult to judge which information of the research process will be important and valuable later on.

- PENSATE A TUTTI I MATERIALI DELLA VOSTRA RICERCA COME «DATI» CHE POTREBBERO ESSERE RIUSATI
- E A QUANTO SAREBBE UTILE PER VOI SE ALTRI CONDIVIDESSERO I LORO
  - PER QUESTO OCCORRE DOCUMENTARLI DIGITALMENTE DALL'INIZIO: NON SI PUÒ MAI SAPERE COSA SARÀ UTILE

Perché occuparci dei dati?





• ...vi è mai capitato...

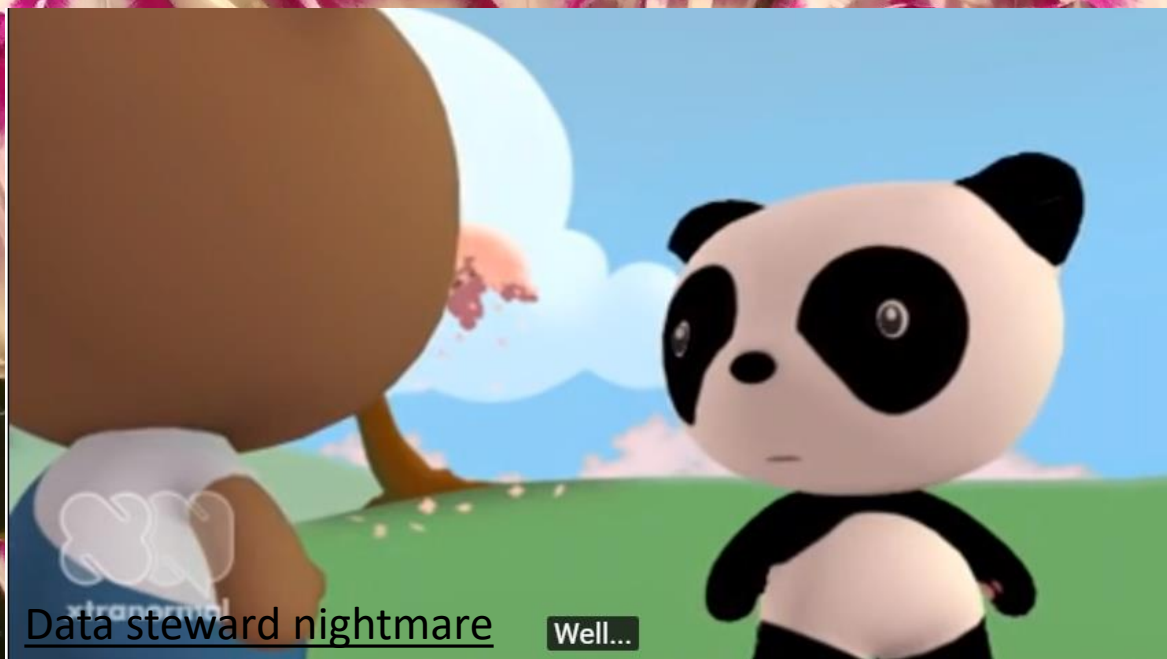
DI AVERE I DATI SUL COMPUTER  
DEL PhD CHE POI SE NE È  
ANDATO??

DI PERDERE DATI?

DI CHIEDERE DATI DOPO AVER  
LETTO UN ARTICOLO E IL  
COLLEGA NON LI TROVA PIÙ??

DI APRIRE DATI ALTRUI E NON  
RIUSCIRE A LEGGERLI??

[Video]



... È L'INCUBO DEL DATA STEWARD:



- NESSUN BACKUP
- NESSUN SOFTWARE DI ACCOMPAGNAMENTO
- NESSUNA LEGENDA DATI

... E IN PIÙ:

- DATI PRODOTTI CON FONDI PUBBLICI
- PUBBLICATI SU SCIENZE CHE LI RICHIEDE
- UTILI A UNA RICERCATRICE DI AREA DIVERSA

# Perché occuparci dei dati?

## Journal open-data policies

 This list is part of the [Open Access Directory](#) 

UN NUMERO CRESCENTE  
DI RIVISTE LO RICHIEDE

1. NELLA VALUTAZIONE EX ANTE (SCHEMA DI DMP)
2. NELLA DISSEMINAZIONE «FAIR, AS OPEN AS POSSIBLE»



V.1 Feb 2021



Horizon Europe (HORIZON)  
Euratom Research and Training Programme (EURATOM)

General Model Grant Agreement  
EIC Accelerator Contract

IT

2019

Gazzetta ufficiale dell'Unione europea

DIRETTIVA (UE) 2019/1024 DEL PARLAMENTO EUROPEO E DEL CONSIGLIO

del 20 giugno 2019

relativa all'apertura dei dati e al riutilizzo dell'informazione del settore pubblico

DATI DELLA RICERCA  
INCLUSI NELLA DIRETTIVA

EOSC Association: Advancing  
Open Science to accelerate the  
creation of new knowledge, inspire  
education, spur innovation and  
promote accessibility and  
transparency

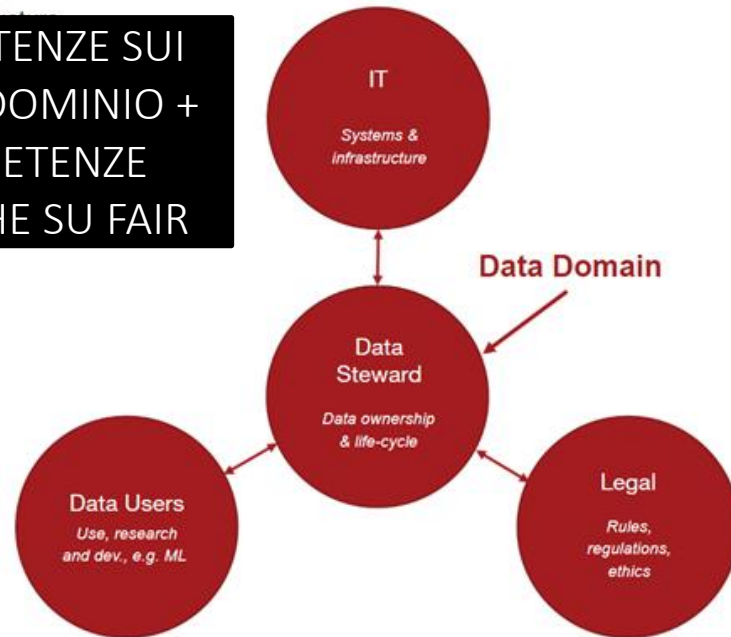
EOSC SI BASA SU  
DATI FAIR



1. PERCHÉ DOBBIAMO.

[EOSC signifi  
stewards]

COMPETENZE SUI  
DATI DI DOMINIO +  
COMPETENZE  
TECNICHE SU FAIR



Copenhagen Univ. June 17 2020

- WE NEED 500.00 DATA STEWARDS
- DATA STEWARDS ARE ONE OF THE CRITICAL SUCCESS FACTORS OF EOSC

Strategic Research and Innovation Agenda  
(SRIA)  
of the  
European Open Science Cloud (EOSC)  
SRIA 1.0 Version 1.0 15 February 2021

#### 7.4. Critical success factors

The developments and expected impacts described above will not happen spontaneously. For these benefits to materialise a number of critical success factors (CSFs) must be in place. The following CSFs have been identified for EOSC:

- Researchers performing publicly funded research make relevant results available as openly as possible;
- Professional data stewards are available in research-performing organisations in Europe to help implement FAIR principles and support Open Science;

# Perché occuparci dei dati?

 | D | C | C because good research needs good data

2. PERCHÉ SONO  
IL FONDAMENTO  
DI UNA SCIENZA  
SOLIDA E DI UNA  
RICERCA  
RESPONSABILE

## What are the Key Recommendations?

The RDA COVID-19 Recommendations and Guidelines are aimed at developing a systematic approach for data sharing in public health emergencies that supports scientific research and policymaking, including an overarching framework, common tools and processes, and principles that can be embedded in research practice.

- 1 Coordinate cross-jurisdictional efforts to foster global **Open Science** through policy and investment.
- 2 Incentivise early publication and release of data and software outputs.
- 3 Invest in state-of-the-art IT, data management systems **infrastructure, economies of scale, and people**.
- 4 Data, software and models should be **timely and FAIR: Findable, Accessible, Interoperable, Reusable**.
- 5 Require the use of **Data Management Plans**.
- 6 Use common generic as well as domain-specific **metadata standards, and persistent identifiers**.
- 7 Provide **documentation** of context, methodologies used to define, construct, and compile data, data cleaning and quality checks, data imputation, and data provenance.
- 8 Use **Trustworthy Data Repositories** committed to the long-term preservation and sustained access to their data holdings.
- 9 **Expedite** article and data review processes, **prioritising and fast-tracking data** at all stages.
- 10 **Balance ethics and privacy**, taking into account public interests and benefits while addressing the health crisis.
- 11 Access should be as **open as possible** and as **closed as necessary**.
- 12 Seek **technical solutions** that ensure anonymisation, encryption, privacy protection, and de-identification to **increase trust** in data sharing.
- 13 Provide **legal frameworks that promote sharing** of surveillance data across jurisdictions and sectors.

... E NELLA CRISI SI  
CAPISCE LA LORO  
IMPORTANZA

# Perché occuparci dei dati

**CASH REWARD**  
for returning my lost backpack



• Black [AK] Burton Rucksack  
• Lost on Friday 15. July at 8 pm in the Panton Arms pub 43, Panton St. Cambridge  
• Containing a laptop (white MacBook), a black external hard drive and scientific research documents

The external hard drive is VERY important to me as it contains 5 years of research data which are crucial for my PhD thesis!!!

If you found it, I would be extremely grateful if you could return it to the Panton Arms or contact me on: 07804430054 (ar456@cam.ac.uk)

Thank you!!  
PMRblog, 2011

 **IMC\_Leeds** @IMC\_Leeds Parthenos training

71% of you got this right. Quite a few pairs of sunglasses, but we have a \*lot\* of USB sticks! Let us know if you think you may have lost yours on campus.

**IMC\_Leeds @IMC\_Leeds**  
What do you think was the most common item of lost property handed in at #IMC2018? Get in touch if you're missing something!

5:04pm · 11 Jul 2018 · Twitter Web Client

MENU CERCA **10 marzo 2021** la Repubblica ABBONATI QUOTIDIANO R A

## OVH, dall'incendio del datacenter di Strasburgo disagi anche per i comuni italiani

di Alessandro Longo



A Pavia, Cattolica, Trapani e altre città finiscono offline siti e servizi pubblici dopo l'incidente nella città francese. L'Agenzia delle Entrate esclude che i rallentamenti registrati oggi siano legati all'episodio. Per gli esperti è la conferma che bisogna accelerare la migrazione a un cloud pubblico, come detto anche dal ministro all'innovazione Vittorio Colao ieri

**COPYRIGHT-Italia** Avv. Simone Alprandi, Ph.D. - Copyright-Italia.it | Array Law Firm www.copyright-italia.it - www.alprandi.org - www.array.eu **ARRAY**

### il backup: definizione (meno seria)

Il backup è quella cosa che andava fatta prima.

(fonte: Proverbio cinese)

— S.Alprandi, Sicurezza dati e privacy (le norme) 2017

**Scientists losing data at a rapid rate** 2013

Decline can mean 80% of data are unavailable after 20 years.

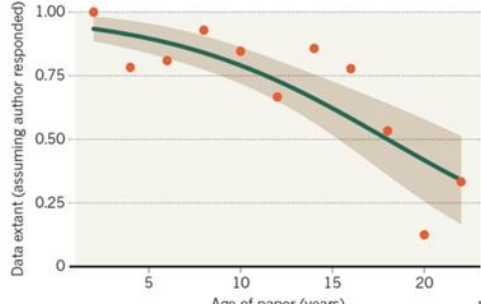
Elizabeth Gibney & Richard Van Noorden

19 December 2013

Rights & Permissions

### MISSING DATA

As research articles age, the odds of their raw data being extant drop dramatically.



Age of paper (years)	Data extant (assuming author responded)
0	1.00
5	0.85
10	0.70
15	0.55
20	0.35

**80% PERSI  
ENTRO 20  
ANNI**

**3. PERCHÉ I DATI SONO  
FRAGILI...**

**...ECCO A COSA SERVE IL  
DATA MANAGEMENT PLAN.  
NON È SOLO L'ENNESIMA NOIA  
BUROCRATICA**

# Parliamone [dal corso febbraio 2021]

Backup su Google Drive (Google suite INRIM) Alcuni dati >10 anni irrecuperabili per obsolescenza software Tutti i dati >20 anni irrecuperabili per obsolescenza hardware

Backup su circa 5 memorie diverse. Non mi è mai capitato di perderli. Non ho neanche avuto problemi di compatibilità (dati ASCII). A volte è stato un problema ricordarsi il significato delle colonne a causa di una insufficiente descrizione.

Li conservo nel mio pc o hard disk esterno. O/E NAS del laboratorio. Sì, è capitato, che dati non "backupati" si perdessero perché la macchina si era rotta...

# Perché occuparci dei dati?



4. PER GARANTIRE INTEGRITÀ.  
I DATI POSSONO ESSERE  
MANIPOLATI...È VOSTRO  
INTERESSE PRIMARIO EVITARLO

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.](#)



# Perché occuparci dei dati?



5. PERCHÉ ALCUNI SONO UNICI E  
IRRIPETIBILI (EVENTI SISMICI O  
METEOROLOGICI)

# Perché occuparci dei dati?



6. PERCHÉ POSSONO  
ESSERE RIUTILIZZATI

... SPESSO IN MODO  
INEDITO, FAVORENDO  
INNOVAZIONE

«THE COOLEST THING TO DO WITH YOUR DATA WILL BE THOUGHT OF BY SOMEONE ELSE» [R.POLLOCK]

## 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



# Perché occuparci dei dati?

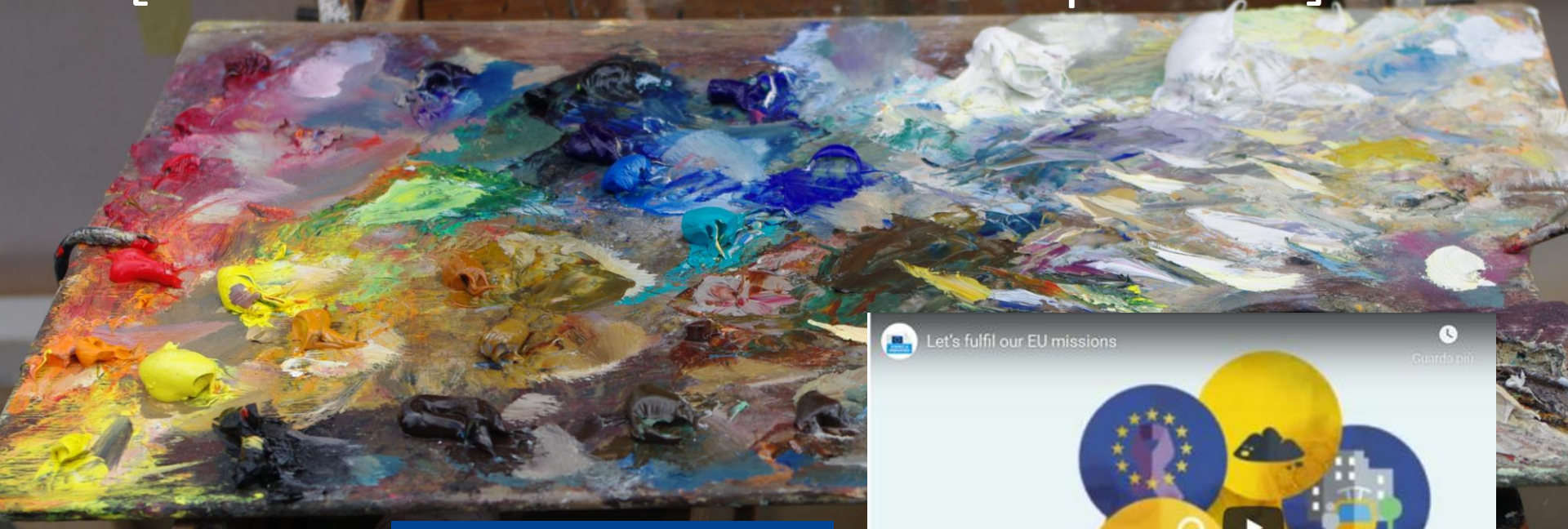
Data creates a bridge between traditional disciplines, spawning discovery and innovation from the humanities to the hard sciences. Data dissolves barriers, opening up new channels of communication, lines of research, and commercial opportunities. Data will be the engine, the spark to create a better world for all.

World Economic Forum 2012

7. I DATI CREANO PONTI  
FRA LE DISCIPLINE...

...E NON È INDIFFERENTE PER LE MISSIONS DI HORIZON EUROPE...

[missions sono interdisciplinari]



## Missions in Horizon Europe

### Mission areas

5 mission areas have been identified, each with a dedicated mission board and help specify, design and implement specific missions in Horizon Europe.

[Mission area: Adaptation to climate change including societal transformation](#)

[Mission area: Cancer](#)

[Mission area: Climate-neutral and smart cities](#)

[Mission area: Healthy oceans, seas, coastal and inland waters](#)

[Mission area: Soil health and food](#)

[Horizon Europe](#)



EU missions will

[Horizon Europe](#)

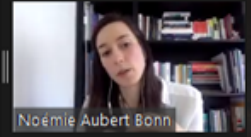
- be bold, inspirational and widely relevant to society
- be clearly framed: targeted, measurable and time-bound
- establish impact-driven but realistic goals
- mobilise resources on EU, national and local levels
- [link activities across different disciplines and different types of research and innovation](#)
- make it easier for citizens to understand the value of investments in research and innovation

We know there are **core problems with research systems** but approaches for integrity tend to focus on researchers

The way in which we measure **success is problematic** and could even lead to integrity issues

Indicators used to advance **research careers are misaligned** with indicators needed to advance **science**

DORA community call March 24, 2021



Noémie Aubert Bonn

# Perché occu

INTEGRITÀ SI VALUTA SUL  
PROCESSO NON SUL  
RISULTATO FINALE

**Welcome The Turing way**

The Turing Way is an open source community-driven guide to reproducible, ethical, inclusive and collaborative data science.

Our goal is to provide all the information that data scientists in academia, industry, government and the third sector need at the start of their projects to ensure that they are easy to reproduce and reuse at the end.

The book started as a guide for reproducibility, covering version control, testing, and continuous integration. However, technical skills are just one aspect of making data science research "open for all".

In February 2020, *The Turing Way* expanded to a series of books covering reproducible research, project design, communication, collaboration, and ethical research.

**Navigation:**  
 Guide for Reproducible Research  
 Guide for Project Design  
 Guide for Communication  
 Guide for Collaboration  
 Guide for Ethical Research  
 Community Handbook  
 Afterword

Visit our GitHub Repository  
 This book is powered by Jupyter Book

**ALLEA**  
ALL European Academies

**integrity** | in  
the quality of being b  
integrity. te of being w

**The European Code of Conduct for Research Integrity**  
REVISED EDITION

[Research Integrity](#)

8. PER ESSERE  
RIPRODUCIBILI

9. PER L'INTEGRITÀ  
DELLA RICERCA

# Perché occuparci di dati?



Il debito pubblico deprime la crescita? Il clamoroso errore di Carmen Reinhart e Kenneth Rogoff 2013

Publicato da keynesblog il 18 aprile 2013 in consigliati, Economia, ibt, Teoria economica



10. PER PERMETTERE  
VALIDAZIONI E  
CONTROLLI  
(E SCOPRIRE ERRORI)

Does High Public Debt Consistently Stifle Economic Growth? A Critique of Reinhart and Rogoff

Thomas Herndon\* Michael Ash Robert Pollin

April 15, 2013

Herndon, 2013

JEL CODES: E60, E62, E65

## Abstract

We replicate Reinhart and Rogoff (2010a and 2010b) and find that coding errors, selective exclusion of available data, and unconventional weighting of summary statistics lead to serious errors that inaccurately represent the relationship between public debt and GDP growth among 20 advanced economies in the post-war period. Our finding is

- ESCLUSIONE SELETTIVA DI DATI
- SCHEMA NON CONVENZIONALE DI PESATURA DEI DATI
- ERRORE NEL FOGLIO DI CALCOLO PER SELEZIONARLI

debt loads greater than 90 percent of GDP consistently reduce GDP growth.

Curiosamente, la posizione espressa dal professor Doshi è stata criticata non su basi scientifiche (che so: errori di metodo, ambiguità negli obiettivi da raggiungere, utilizzo di software e strumentazioni inadatte) ma dell'opportunità politica e del principio di autorità. Cioè su presupposti **diametralmente opposti** a quelli di una

## Le ragioni di Peter Doshi sui vaccini: "Fidati, ma verifica"

COVID-19/Filosofia

di Andrea Monti

Chi pratica il metodo scientifico ha la testarda abitudine (incomprensibile ai più) di trarre conclusioni dall'analisi di dati secondo i criteri di un'ipotesi di ricerca e applicando un metodo che consente la verificabilità intersoggettiva dei risultati. Questo atteggiamento mentale è diametralmente opposto a chi basa le proprie opinioni e—peggio— decisioni sulla "fiducia" (spesso tramutata in "fede") e dunque sull'autorità di eminenze varie. Non discuto questo atteggiamento nell'ambito religioso; ma in quello laico, quello della scienza, sì. Se un dogma esiste, nella pratica del metodo scientifico, è quello della **metodicità del dubbio**, insieme a quello dell'**assenza di certezze**. Un teoria s quando vale. È successo, tanto per fare un e di fuori dei laboratori, con la gravitazione d

thebmjopinion

Jan. 4, 2021 Latest

Authors ▾

Topic

Peter Doshi: Pfizer and Moderna's "95% effective" vaccines—  
we need more details and the raw data

January 4, 2021

Five weeks ago, when I [raised questions](#) about the results of Pfizer's and Moderna's covid-19 vaccine trials, all that was in the public domain were the [study protocols](#) and a [few press releases](#). Today, two [journal publications](#) and around 400 pages of summary data are available in the form of [multiple reports presented by](#) and [to the FDA](#) prior to the agency's emergency authorization of each company's mRNA vaccine. While some of the additional details are reassuring, some are not. Here I outline new concerns about the trustworthiness and meaningfulness of the reported efficacy results.

# Nessun dato?

Is withholding your data simply bad science, or should it fall under scientific misconduct?

22 comments | 5 shares

Estimated reading time: 5 minutes



A recent study sent data requests to 200 authors of economics articles where it was stated 'data available upon request'. Most of the authors refused. What does the scientific community think about those withholding their data? Are they guilty of scientific misconduct? **Nicole Janz** argues that if you don't share your data, you are

breaking professional standards in research, and are thus committing scientific misconduct.

Classifying data secrecy as misconduct may be a harsh, but it is a necessary step.



Alastair Dunning

@alastairdunning

Following

To me, data are like footnotes. I might not always read them, but I get suspicious if they are not there.

Traduci dalla lingua originale: inglese

12:49 - 27 feb 2018

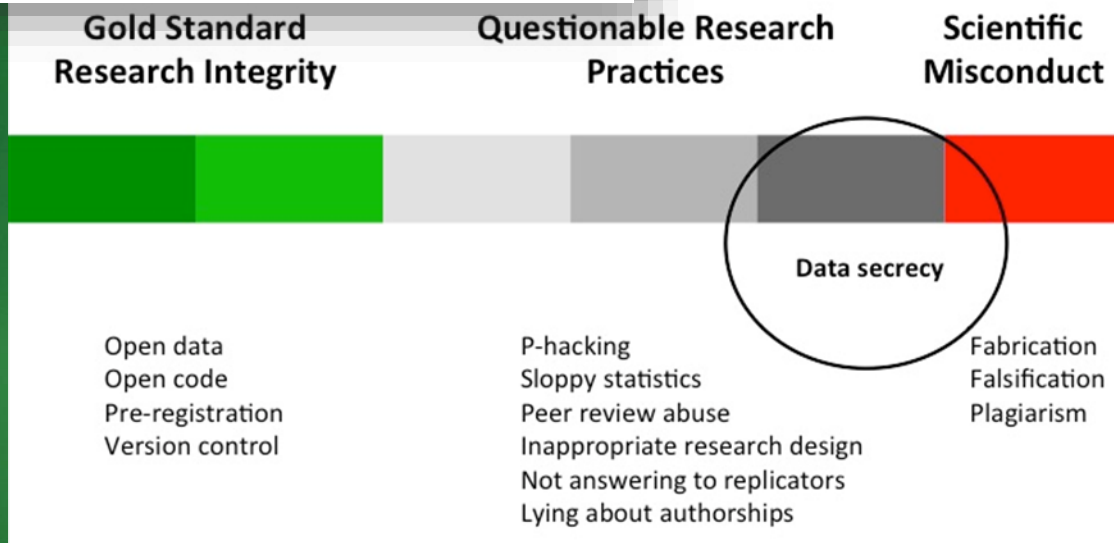
<https://twitter.com/alastairdunning/status/968453078218395648>

2 Retweet 8 Mi piace



NESSUN DATO?  
PIGRIZIA O FRODE?  
I DATI COME NOTE A PIE'  
PAGINA: POSSO NON  
LEggerLE, MA DIVENTO  
SOSPETTOSO SE NON CI SONO

2015





# Parliamo di dati

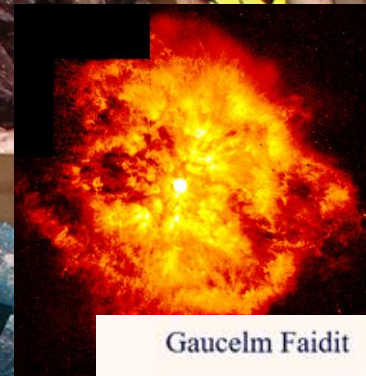
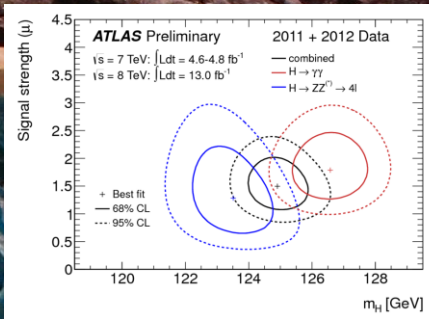


Table S1. Number of reads per prokaryotic operational taxonomic unit (OTU) and sample for

A04_TP1	A04_T1
25707	6
119	
73	2
382	
15	
1843	
5429	
29	
0	

## Gaucelm Faidit

I.

Ara nos sia guitz  
 lo vers dieus Iesu Cristz,  
 car de franca gen gaia  
 soi per Lui partitz,  
 on ai estat noiritz  
 et onratz e grazitz;  
 per so-l prec no-ill desplaia  
 s'ieu m'en vauc marritz.

A! gentils lemozis,  
 el vostr'onrat pais  
 lais de bella paria  
 seignors e vezis  
 e domnas ab pretz fis,  
 pros, de gran cortesia,  
 don plane e languis  
 e sospir nueg e dia.



**Wilma van Wezenbeek**

@wvanwezenbeek

Following

#osc2018 Wolfram Horstmann wants us to talk about datadiversity, like we do with biodiversity #openscience

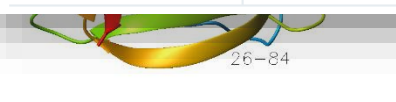
Traduci il Tweet

12:51 - 13 mar 2018

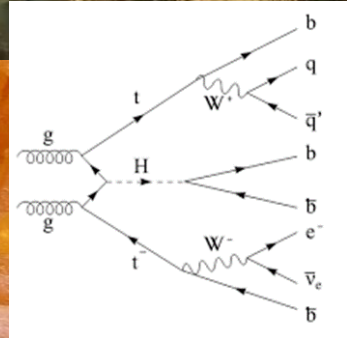
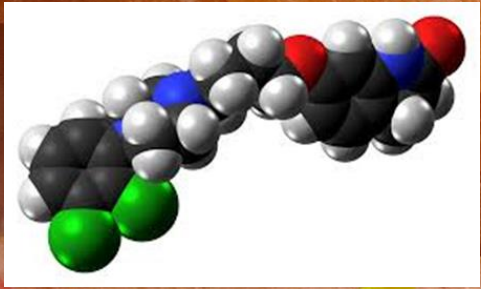
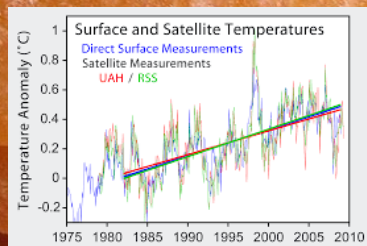
3 Retweet 1 Mi piace



<https://twitter.com/wvanwezenbeek/status/973527086685093893>



**B**atus vir,  
 qui non abit  
 in ostio im-  
 pioꝝ, et in via  
 peccatoꝝum  
 non stetit: et  
 in cathedra pestifeꝝie noꝝ sedet.  
 Sed i lege dñi volũtas ei:  
 et i lege ei inditabit die, ac nocte.  
 Et erit tamq̃s lignus, quod  
 plantatus eꝛ scens decursus  
 aquarũ: quod fructus suum  
 dabit in tempore suo.  
 Et foliũ eius non defluet: et  
 oia q̃cuq; faciet p̃sperabũt.  
 Non sic impij, non sic: sed



dati

## What is Data, Anyway? Wuttke 2018

**Do Humanities and Cultural Heritage researchers have data?**

- Yes, a lot, but they don't tend to use the word data
- Research Data are data that are produced in and used in scientific processes such as digitization, study of sources, experiments, measurements, interviews, and surveys

*We could then define data in the humanities broadly as all materials and assets scholars collect, generate and use during all stages of the research cycle. In this report we focus on digital assets.*



**DATI = TUTTO CIÒ CHE VIENE RACCOLTO,  
GENERATO E USATO NEL PROCESSO DI RICERCA**

[DMP]

# Le basi

rdnl re ESSENTIAL4DATA  
data  
netherlands

Essentials 4  
Data Support

ABOUT THE COURSE START THE COURSE LOGIN

## 5 MODI PER PENSARE I DATI:

- COME SONO RACCOLTI (ESPERIMENTI, SIMULAZIONI...)
- COME SI PRESENTANO (TESTI, QUESTIONARI, VIDEO...)
- IL LORO FORMATO ELETTRONICO (.TXT, .MKV...)
- IL LORO VOLUME (BIG DATA...)
- IN CHE FASE SONO DEL CICLO (RAW DATA...)

### ▣ The way the data is collected.

- ▣ By experimenting, simulations, observations, derived data, reference data.

### ▣ The data forms.

- ▣ For example text documents, spreadsheets, lab journals, logs, questionnaires, software code, transcripts, code books, audio and video recordings, photos, samples, slides, artefacts, models, scripts, databases, metadata, etc.

### ▣ The formats for electronic storage of the research data.

### ▣ The size (volume) of the data files.

### ▣ The *research lifecycle* phase the data is in.

RICHIEDONO  
STRUMENTI E  
TRATTAMENTI  
DIVERSI

## Part I

## Five Ways To Think About Research Data

Science has progressed by 'standing on the shoulders of giants' and for centuries research and knowledge has been shared through the publication and dissemination of books, papers and scholarly communications. Moving forward much of our understanding builds on (large scale) data sets which have been collected or generated as part of this scientific process of discovery. How will this be made available for future generations? How will we ensure that, once collected or generated, others can stand on the shoulders of the data we produce?

Deciding on how to look after data depends on what your data looks like and what needs to be done with it. You should find out if your discipline already has standard practices and use them. We hope that this brief introduction will give some templates of what is already being done in a few disciplines and enable you to start thinking about what you might do with your research data to make it accessible to others.

Further University of Southampton guidance can be found on the library's web site <http://library.soton.ac.uk/researchdata>. Any research data management questions can be emailed to [researchdata@soton.ac.uk](mailto:researchdata@soton.ac.uk).

This part of the guide introduces five ways of looking at research data.

## 1 Research data collection

The first way of thinking about research data is where it comes from (Research Information Network, 2008). Each of the case studies in Part II illustrates one of these categories.

**Reference data:** *Example: the reference human genome sequence in Case Study 1*  
A data set that can be used for validation, comparison or information lookup.

**Scientific experiments:** *Example: materials engineering fatigue test in Case Study 2*  
Data generated by, e.g. instruments during a scientific experiment.

**Models or simulations:** *Example: CFD helicopter rotor wake simulation in Case Study 3*  
Data generated on computer by an algorithm, mathematical model, or the simulation of an experiment. A computer simulation can help when experiments are too expensive, time consuming, dangerous or even impossible to perform.

**Derived data:** *Example: chemical structures in chemistry in Case Study 4*  
A data set created by taking existing data and performing some manipulation to it. Each data set requires careful curation because the original data may be needed to understand the new data.

**Observations:** *Example: archaeological dig in Case Study 5*  
Data generated by recording observations of a specific, possibly unrepeatable, event at a specific time or location.

## 2 Types of research data

Research can come in many different forms, some electronic and some physical. Here are some examples:

- Electronic text documents, e.g. text, PDF, Microsoft Word files
- Spreadsheets
- Laboratory notebooks, field notebooks and diaries
- Questionnaires, transcripts and codebooks
- Audiotapes and videotapes
- Photographs and films
- Examination results
- Specimens, samples, artefacts and slides
- Digital objects, e.g. figures, videos
- Database schemas
- Database contents
- Models, algorithms and scripts
- Software configuration, e.g. case files
- Software pre-process files, e.g. geometry, mesh
- Software post-process files, e.g. plots, comma-separated value data (CSV)
- Methodologies, workflows, standard operating procedures and protocols
- Experimental results
- Metadata (data describing data), e.g. environmental conditions during experiment
- Other data files, e.g. literature review records, email archives

## 3 Electronic storage

The third way to think about research data is how it is stored on a computer. Here are some of the categories of electronic data:

**Textual, e.g.:**

- Flat text files
- Microsoft Word
- PDF
- RTF

**Numerical, e.g.:**

- Excel
- CSV

**Multimedia, e.g.:**

- Image (JPEG, TIFF, DICOM)
- Movie (MPEG, AVI)
- Audio (MP3, WAV, OGG)

**Structured, e.g.:**

- Multi-purpose (XML)
- Relational (MySQL database)

**Software code, e.g.:**

- Java
- C

**Software specific, e.g.:**

- Mesh
- Geometry
- 3D CAD
- Statistical model

**Discipline specific, e.g.:**

- Flexible Image Transport System (FITS) in astronomy
- Crystallographic Information File (CIF) in chemistry

**Instrument specific, e.g.:**

- Olympus Confocal Microscope Data Format
- Carl Zeiss Digital Microscopic Image Format (ZVI)

Data can be born digitally, such as a simulation, or ingested into a computer, such as scanning a photograph. Some data can remain in a non-digital format.

# [il ciclo di vita dei dati]



## Services offerings throughout the research lifecycle

### Research Lifecycle



The research lifecycle refers to the (often iterative) process of conducting research, from the initial planning, funding, and research project design to publishing and disseminating the conclusions or work of scholarship. Although the research process varies across disciplines and research domains, it often includes validating a model or hypothesis by using information and data. In turn, the results from the data help improve the model and thus, gather additional data to validate the new model. On this site, we refer to data in the broadest sense of the word, including experimental, observational, acquired, and simulated data, as well as any relevant information, artifacts, and original sources. In recent years, the research lifecycle has also included publishing the data, code, and workflows to facilitate the reproducibility of the published results.

#### Planning:

Access & Reuse  
Plan & Design  
(14 service offerings)

#### Active Research:

Collect & Create  
Analyze & Collaborate  
(22 service offerings)

#### Dissemination & Preservation:

Evaluate & Archive  
Share & Disseminate  
(5 service offerings)



...un passo indietro...

# [il fondamento] [DMP]

## Information Guide: Introduction to Ownership of Rights in Research Data. CREATE, University of Glasgow, 2018

2018

Burrow, S. , Margoni, T.  and McCutcheon, V.  (2018) Information Guide: Introduction to Ownership of Rights in Research Data. CREATE, University of Glasgow, 2018. Documentation. University of Glasgow.



Guides for Researchers

How do I know if my research data is protected?

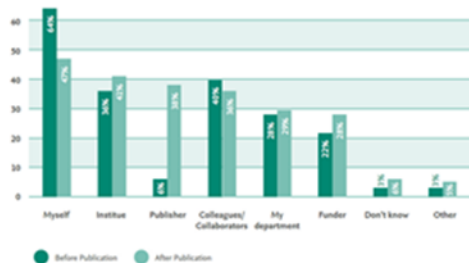
Learn more about what is research data and their protection by intellectual property rights

OpenAIRE

I DATI GREZZI  
NON SONO «MIEI»  
NON ESISTE COPYRIGHT  
PERCHÉ NON SONO CREATIVI

This time though it happened. What it was: 64% of researchers believe they own the data they generated for their research.

Figure 3. Research data ownership before and after publication (%; n=1162)



The result comes from a **solid piece of academic research** based on equally solid (open) data. The study and the report 'Open Data - the Researcher Perspective' were done by **CWTS / Leiden** and **Elsevier**. Credit giving, check.

Of course, the study reports other equally surprising results



Wainer Lusoli

@w\_lusoli

Following

repeat with me: [#researchdata](#) is NOT mine. I was paid to get it, I'll get a [#nobel](#) 4 it, but it's NOT mine [linkedin.com/pulse/repeat-m ...](https://www.linkedin.com/pulse/repeat-m...) [#opendata](#)

Traduci dalla lingua originale: inglese



### Repeat with me: research data is not mine

Seldom do I see something that truly shakes me at work. You know, work is work, I am no neurosurgeon, no médecin sans frontières nor am I a social

[linkedin.com](https://www.linkedin.com)

11:18 - 12 apr 2017

14 Retweet 18 Mi piace



[Lusoli, Apr.2017](#)

[DMP]

[webinar]

OpenAIRE 2019

SERVICES SUPPORT OPEN SCIENCE IN EUROPE ABOUT

More Information about the 2019 webinar series.  
data management plan | OA to research data | open science

**Aspetti legali nella gestione dei dati della ricerca**

Thomas Margoni  
University of Glasgow - CREATE  
OpenAIRE project

**Support**

- RESOURCES
  - Open Science Primers
  - Guides
  - Factsheets
  - Use cases
- HELPDESK
  - FAQs
  - Ask a Question
- TRAINING
  - Webinars
  - Workshops
  - Community of Practice

RDA RESEARCH DATA ALLIANCE EUROPE  
IT NATIONAL NODE  
OpenAIRE

- POSSONO ESSERCI ALTRE FORME DI PROTEZIONE DEI DATI (ES. CONTRATTI)
- PER DATI CHE RICADONO SOTTO GDPR VA SEMPRE ESPLICITATA LA BASE LEGALE SULLA QUALE SI CONDUCE LA RICERCA

2020

OpenAIRE Legal Policy Webinars

**Supporting researchers on the reuse of data: legal aspects to consider**

29th April and May 4th, at 2 PM CEST



[i tre passi fondamentali]

OPEN

FAIR

GESTITI

1. I DATI DEVONO ESSERE «AS OPEN AS POSSIBLE»

2. MA SE I DATI NON SONO «FAIR», APRIRLI COMPORTA RISCHI  
(USO SCORRETTO, CATTIVE INTERPETAZIONI, ...)

3. MA SE I DATI NON SONO CORRETTAMENTE GESTITI, RENDERLI  
«FAIR» COSTA TROPPO TEMPO E DENARO. CON EOSC, DATI GESTITI E  
DATI FAIR TENDONO A COINCIDERE, FAIR BY DESIGN

E GESTIRE I DATI CORRETTAMENTE È NELL'INTERESSE PRIMARIO DI CHI FA RICERCA,  
PERCHÉ L'INTERA RICERCA SCORRE PIÙ FLUIDA

Open non FAIR è Open???

# Shades of Open <sup>2019</sup>

Open consumption **“Can I use it?”**

**If there is no license, the legal default is that you cannot use it!**

- Open for analysis
- Open for reuse
- Open for redistribution
- Open to adapt
- Open for redistribution of adapted versions
- Open, but with obligation to cite
- Open, but not for commercial applications
- Open, in name only, without explicit permissions



Open access to data **“Can I get it?”**

# 1. I dati vanno gestiti



CONSERVAZIONE  
SUL LUNGO  
PERIODO

ASPETTI LEGALI

ORGANIZZAZIONE  
(file naming,  
folders,  
versioning...)

METADATI

BACKUP E STORAGE

Data management is an active process by which digital resources remain discoverable, accessible and intelligible over the longer term, a process that invests data and datasets with the potential to accrue value as assets enjoying far wider use than their creators may have anticipated. In the world of research, such a value-adding process is a significant contributor to the much desired achievement of impact.

## 2. I dati DEVONO essere FAIR

### To be Findable:

F1. (meta)data are assigned a globally unique and eternally persistent identifier.

F2. data are described with rich metadata.

F3. (meta)data are registered or indexed in a searchable resource.

F4. metadata specify the data identifier.

### TO BE ACCESSIBLE:

A1 (meta)data are retrievable by their identifier using a standardized communications protocol.

A1.1 the protocol is open, free, and universally implementable.

A1.2 the protocol allows for an authentication and authorization procedure, where necessary.

A2 metadata are accessible, even when the data are no longer available.

### TO BE INTEROPERABLE:

I1. (meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation.

I2. (meta)data use vocabularies that follow FAIR principles.

I3. (meta)data include qualified references to other (meta)data.

### TO BE RE-USABLE:

R1. meta(data) have a plurality of accurate and relevant attributes.

R1.1. (meta)data are released with a clear and accessible data usage license.

R1.2. (meta)data are associated with their provenance.

R1.3. (meta)data meet domain-relevant community standards.


Force 11

«ACCESSIBLE»

≠«OPEN»

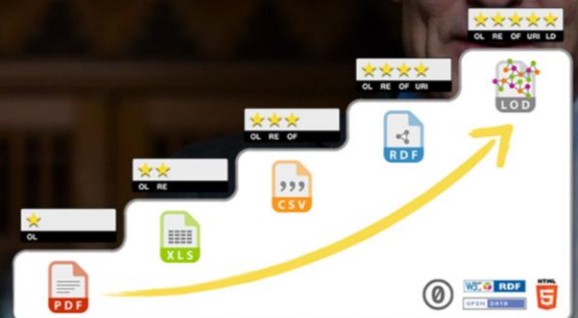
= DOVE E A QUALI  
CONDIZIONI  
I DATI SONO  
ACCESSIBILI

# 3. I dati POSSONO essere Open



## 5 ★ OPEN DATA

Tim Berners-Lee, the inventor of the Web and Linked Data initiator, suggested a 5-star deployment scheme for Open Data. Here, we give examples for each step of the stars and explain costs and benefits that come along with it.



The image shows two golden-brown, flaky pastries, possibly croissants or similar baked goods, resting on a white plate. The pastries are the central focus, with a textured, woven placemat visible in the background. The lighting is bright, highlighting the texture of the pastries.

[una premessa]

- ...DA QUI IN POI: PANORAMICA SUGLI STRUMENTI
- VANNO «ASSAGGIATI» E ADATTATI AL PROPRIO CONTESTO...
- IMPARARE A USARLI PER SUGGERIRLI (IMPENSABILE CHE UN RICERCATORE SCENDA COSÌ NEL DETTAGLIO)
- FONDAMENTALE IL SUPPORTO
- **FONDAMENTALE UNA POLITICA ISTITUZIONALE CHE CHIARISCA RUOLI E RESPONSABILITÀ E DEFINISCA IL LIVELLO DEI SERVIZI**
- **CREARE UNA RETE DI DATA STEWARDS**  
[COMPETENZE DI DOMINIO + TECNICHE]

Ci sono tre passaggi:

1. I dati vanno gestiti correttamente (nell'interesse del ricercatore: il lavoro risulta più fluido e si ri
2. I dati vanno resi FAIR by design
3. SE POSSIBILE, i dati vanno aperti

Perché investire sulla gestione dei dati (B.Mons. 2020)

#### GUIDE E CORSI SULLA GESTIONE DEI DATI

- CESSDA Data management expert guide (corso free in 7 moduli)
- Essentials4data (corso free in 6 moduli)
- FOSTER pagina dei corsi (scorrere i singoli moduli su Data protection, Data sharing...)

#### COME SCRIVERE UN FILE README

- Guida MIT Boston
- Guida TU Delft

#### COME CALCOLARE I COSTI

- Data Wizard cost evaluator
- TU Delft costing tool

#### FILE NAMING E VERSIONING

- File naming conventions
- File naming and folder structure
- Data versioning ANDS
- Data versioning RDA

#### BACKUP E STORAGE

- Storage pro e contro
- Appraisal (cosa conservare)

#### ASPETTI LEGALI

- Information Guide: Introduction to Ownership of Rights in Research Data 2018
- Legal Guide OpenAIRE (diverse sezioni su GDPR, direttiva sui generis, protezione dei dati...)
- How do I license research data OpenAIRE
- webinar Aspetti legali (ITA) 2019
- webinar Legal aspects (ENG) 2020
- Personal data FOSTER project
- Data ethics FOSTER project

#### VIDEO

- Incubo del data steward (orsetti)
- Data management dai ricercatori per i ricercatori (3 video)

[e un

OSF  
«OPEN SCIENCE IN PRATICA»  
TROVATE RACCOLTI TUTTI I  
LINK CHE VEDRETE NELLE SLIDE  
[REGISTRAZIONE]

[DMP]

# 4 pilastri

ands AUSTRALIAN NATIONAL DATA SERVICE

About ANDS | Events | Contact Us

Australia data service

About us | News and Events | Partners and Communities | Working with data | Online Services | Guides and resources

Findable | Persistent Identifiers (PIDs) | Rich metadata | Indexed data repositories | PIDs in metadata

Accessible | Standard communications protocol | Open, free protocol | Authentication, where necessary | Metadata is always available

FAIR data training

If you run workshops on FAIR data, or include FAIR in training that you are already running check out these ideas and resources.

Search the ANDS Site

Enter Keywords



## Digital Curation Center UK

Because good research needs good data

- About
- News
- Events
- Services
- Guidance
- Research



The Digital Curation Centre in collaboration with Research Data Netherlands have developed an online course on **Delivering Research Data Management Services (DRDMS)**.

After two enrolmen the 5th of



## Dutch data service

Nederlands | Contact | Search this website

- HOME
- FOR RESEARCHERS
- FOR DATA PROFESSIONALS
- FOR PARTNERS
- PROJECTS
- ABOUT DANS
- NEWS AND EVENTS

Welcome at DANS: the Netherlands institute for permanent access to digital research resources.

What can we do for you?



## Dutch consortium

Zoeken in de site

- Home
- Experts/Mailing List
- Task groups
- RDM Advice & Tips
- RDM in NL
- About us/Meetings

### The data support collective



DANS for researchers

DANS for data professionals

DANS for partners



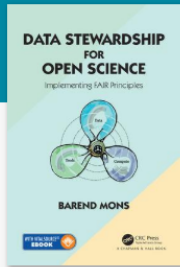
# ... e un maestro



Taylor & Francis Group  
an informa business

2019

Search for keywords, authors, titles, ISBN



## Data Stewardship for Open Science Implementing FAIR Principles

the worst way imaginable to communicate the outcome of the scientific process. If science has become indeed data driven and *data is the oil of the 21st century*, we better put data centre stage and publish data as first-class research objects, obviously with supplementary narrative where needed, steward them throughout their life cycle, and make them available in easily reusable format.

Yet another recent study claimed that only about 12% of NIH funded data finds its way to a trusted and findable repository. Philip Bourne, when associate director for data science at the U.S.A. National Institutes of Health coined the term dark data for the 88% that is lost in amateur repositories or on laptops. When we combine the results of the general reproducibility related papers and the findability studies,

GET ACCESS

PREVIEW PDF

PASSARE DA ARTICLE+  
A DATA +  
[CAPITOLI ACCESSIBILI DA  
DATA WIZARD]



In conclusion to this paragraph, my statement in 2005: Text-mining? Why bury it first and then mine it again? [Mons, 2005] is still frighteningly relevant.

*A good data steward publishes data with a supplementary article(Data(+)).*

11 5%

nature

Feb. 25, 2020

Subscribe

WORLD VIEW · 25 FEBRUARY 2020

## Invest 5% of research funds in ensuring data are reusable



It is irresponsible to support research but not data stewardship, says Barend Mons.

Barend Mons

I tell research institutions that, on average, 5% of overall research costs should go towards data stewardship. With €300 billion (US\$325 billion) of public money spent on research in the European Union, we should expect to spend €15 billion on data stewardship. Scientists, especially more experienced ones, are often upset when I say this. They see it as 5% less funding for research.

Bunk. First, taking care of data is an ethical duty, and should be part of good research practice. Second, if data are treated properly, researchers will have significantly more time to do research. Consider the losses incurred under the current system. Students in PhD programmes spend up to 80% of their time on 'data munging', fixing formatting and minor mistakes to make data suitable for analysis – wasting time and talent. With 400 such students, that would amount to a monetary waste equivalent to the salaries of 200 full-time employees, at minimum. So, hiring 20 professional data stewards to cut time lost to data wrangling would boost effective research capacity.

Many top universities are starting to see that the costs of not sharing data are significant and greater than the associated risks. Data stewardship offers excellent returns on investment.

- PRENDERSI CURA DEI DATI È ETICO
- ASSUMERE DATA STEWARDS FA RISPARMIARE TEMPO
  - FAIR=FULLY ARTIFICIAL INTELLIGENCE READY

Funders hold the stick: they should disburse no further funding without a properly reviewed and budgeted data-stewardship plan. The carrot is that FAIR data allow much more effective artificial intelligence (FAIR can also mean 'fully AI ready'), which will open up unprecedented research opportunities and increase reproducibility.

# Costi

CILIEGINO  
CATI SICILIA

€ 3,60

## The problem

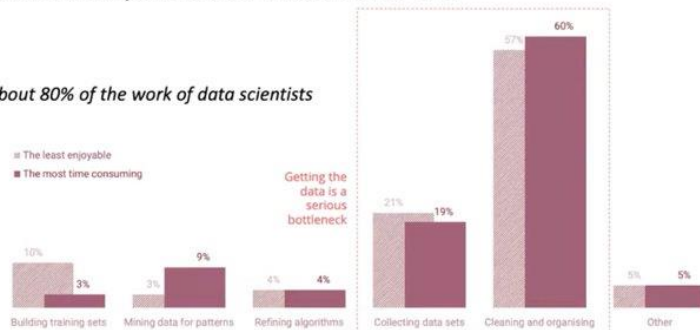
Data science report, 2016, cit. by Susanna Sansone [Apr. 27, 2021](#)

Discoveries are made using shared data and this requires data that are:

- Retrievable and structured in standard format(s)
- Self-described so that third parties can make sense of it

Data preparation accounts for about 80% of the work of data scientists

[Forbes article](https://www.forbes.com/sites/glpres/2016/03/23/data-preparation-most-time-consuming-least-enjoyable-data-science-task-survey-says/#276a35e6637) on 2016 Data Scientist Report




10,2 bn + 16 bn

CI SONO COSTI PER GESTIRE, RENDERE FAIR E CONSERVARE I DATI...  
MA PENSIAMO

- A QUANTO COSTEREBBE NON CONSERVARLI E NON GESTIRLI
- A QUANTO TEMPO PERDETE PER «PULIRLI» PRIMA DI POTERLI USARE (79% DEL TEMPO PER PREPARARLI]


# Costi



OpenAIRE

## What will it cost to manage and share my data?

What to cost in?




OpenAIRE

OpenAIRE How to

SERVICES SUPPORT OPEN SCIENCE IN EUROPE

Guides for Researchers

### How to identify and assess Research Data Management (RDM) costs



Based on these examples, we have performed some comparative calculations. The cheapest repository changes at different points so shop around!

- 20GB** of data could cost you between €0 - €109 to deposit
- 75GB** of data could cost you between €245 - €340 to deposit
- 200GB** of data could cost you between €790 - €906 to deposit

DSW Storage Costs Evaluator [Data wizard](#)

Total costs: 2 261 € TB costs per year: 452 € [Result details](#)

Volume: 500 GB

Lifetime:



DM costing tool

## Data Management costing tool

4TU. CENTRE FOR RESEARCH DATA TU Delft

DSW Storage Costs Evaluator

Total costs: 2 261 € TB costs per year: 452 € [Result details](#)

Storage drives	158 €
Storage servers	421 €
Networking	183 €
Tape backup	68 €
Setup	575 €
Incident response	362 €
Uninterruptible power supplies	495 €

# 1. GESTIRE I DATI

*Good data management facilitates the reuse of data, which helps avoid duplication of effort, and mitigates against data loss. It also supports collaboration, facilitates continuity across projects, and improves the visibility and impact of research outputs.*

2020

Jisc

How and why you should manage your research data: a guide for researchers

An introduction to engaging with research data management processes.

[JISC Guide](#)



D | C | C

because good research needs good data

# Perché gestire i dati

RISPARMIO DI TEMPO  
MAGGIORE EFFICIENZA  
CONSERVA E PROTEGGE  
I DATI SONO UN «PRODOTTO»  
TRASPARENZA/INTEGRITÀ



- **Save Time** – By spending a little bit of up-front time and planning and organising the data you produce you will save time and resources in the long run.
- **Increase your efficiency** – If you document your data properly whenever you or someone else comes to it they will be able to understand it quickly and without difficulty. Thus saving time and increasing efficiency.
- **Preserve and protect your data** – It is relatively easy to produce data that will be useful only the once and for a very specific purpose. Learn how to ensure that the data can be useful again and again, and how to make sure that it is never lost.
- **Data is an output in its own right** – that's right; data itself is increasingly being seen as an important output of research. If shared, it can better enable researchers. The REF (Research Excellence Framework) now takes note of it.
- **Meet grant requirements** – Many funding bodies now require that researchers archive data as well as the resulting publications as part of their project. Good data management will make this easy rather than a last minute chore.
- **Open Access** – In the UK government policy has moved to an open access framework. Producing and making available data is a vital part of this process. Journals are increasingly making room for data alongside articles, for example.
- **Transparency/research integrity** – If required you have all the documents and materials easily available making your research more transparent if questioned.

[Why data management](#)



# Main Points for Good Data Management

## Data acquisition

- Check the type, source of the data and how to gather/collect it
  - Data types (to help define sensitivity of data)
  - Data format (to help define the tools and software)
  - Data size (to help define storage and infrastructure)
- Check the ownership of the collected and processed data
  - Check with the data source about conditions (e.g. licence)
  - Check the need to make a data process on the ownership / access control
  - Are there (own) institutional policies that apply?
  - Can the data be shared with other parties?
- Confidentiality of the data (if applicable):
  - Register crucial information regarding data
  - Ensure security of confidential data (personal data, or data that would harm society with disclosure)
  - Ensure compliance with General Data Protection Regulation (GDPR) / Verordening gegevensbescherming when applicable
  - Ensure there are procedures in place to consult of a privacy advisor/data protection officer

## Data collection

- Establish a workflow for data collection
  - How will the data be collected?
  - Who has access to which data in short / long term?
  - What resources are needed for data analysis?
  - How will the data be exchanged / transferred among relevant stakeholders?
- Storage arrangement
  - Check available storage capacity and backup strategy

## Data storing / backup

- Create a clear folder structure and consistent file naming convention
- Make a backup strategy where data is stored at least two different physical locations and preferably automatically backed up
- Access control to confidential data
- Apply encryption at disk or folder level if needed
- Create a consistent and standard versioning of the data files
- Determine the minimal documentation of the data that is required to find it, understand it and use it

## Data sharing

- Create proper data sharing procedures
  - Consider agreements established in the Data acquisition phase, and evaluate/assess data sharing with other parties
  - Be aware of the permission and consequence of sharing confidential data
- Copyright / Licensing
  - How should others use the data
  - Who should be attributed for creating/gathering the data

## Organizational Implications

In addition to the above mentioned actions, there are also a few things to consider to make data management a standard practice in daily operations.

PER FARSI LE DOMANDE  
GIUSTE



[prepararsi]

FILE

...PERCHÉ IL PROBLEMA NON È SOLO DATI APERTI/CHIUSI A FINE RICERCA...  
MA, BEN PIÙ IMPORTANTE,  
DOVE LI CONSERVO MENTRE CI LAVORO?  
CHI HA ACCESSO?  
CHE SISTEMA DI SICUREZZA È PREVISTO?

Level	Data Classification and Examples (abridged version)
5	<p><b>Information that would cause severe harm to individuals or the University if disclosed.</b></p> <ul style="list-style-type: none"><li>• Research information classified as Level 5 by an IRB or otherwise required to be stored or processed in a high security environment and on a computer not connected to the Harvard data networks</li><li>• Certain individually identifiable medical records and genetic information, categorized as extremely sensitive</li></ul>
4	<p><b>Information that would likely cause serious harm to individuals or the University if disclosed.</b></p> <ul style="list-style-type: none"><li>• High Risk Confidential Information (HRCI) and research information classified as Level 4 by an IRB</li><li>• Personally identifiable financial or medical information</li><li>• Information commonly used to establish identity that is protected by state, federal, or foreign privacy laws and regulations</li><li>• Individually identifiable genetic information that is not Level 5</li><li>• National security information (subject to specific government requirements)</li><li>• Passwords and Harvard PINs that can be used to access confidential information</li></ul>
3	<p><b>Information that could cause risk of material harm to individuals or the University if disclosed.</b></p> <ul style="list-style-type: none"><li>• Research information classified as Level 3 by an IRB</li><li>• Information protected by the Family Educational Rights and Privacy Act (FERPA) to the extent it is not covered under Level 4 including non-directory student information and directory information about students who have requested a FERPA block</li><li>• HUIDs associated with names or any other information that could identify individuals</li></ul> <p>employees may discuss terms and conditions of employment with each other and third</p> <p>any student information and directory information about students who have requested a</p> <p>or any other information that could identify individuals</p> <p>employees may discuss terms and conditions of employment with each other and third</p> <p>protected under state, federal and foreign privacy laws not classified as Level 4 or 5</p> <p><b>which would not cause material harm, but which the University has chosen to</b></p> <p>and intellectual property not in Level 3 or 4</p> <p>ed as Level 2 by an IRB</p> <p>papers, drafts of research papers</p> <p>n about the University physical plant</p>
1	<p><b>Public information.</b></p> <ul style="list-style-type: none"><li>• Research data that has been de-identified in accordance with <a href="#">applicable rules</a></li><li>• Published research</li><li>• Published information about the University</li><li>• Course catalogs</li><li>• Directory information about students who have not requested a FERPA block</li><li>• Faculty and staff directory information</li></ul>

# Serve formazione?

[DMP]



Data Management Expert Guide

- 1. Plan
- 2. Organise & Document
- 3. Process
- 4. Store
- 5. Protect
- 6. Archive & Publish
- 7. Discover



## Plan

In this introductory tour, you will become aware of what data management and a data management plan (DMP) are and why they are important. General concepts such as social science data and FAIR data will be explained. Based on our recommendations and good practice examples, you will be able to start writing your DMP.

## Organise & Document

If you are looking for good practices in designing an appropriate data file structure, naming, documenting and organising your data files within suitable folder structures, this chapter is for you.

## Process

## Store



To be able to plan a storage and backup strategy, you will learn about different storage and backup solutions and their advantages and disadvantages. Also, measures to protect your data from unauthorised access with strong passwords and encryption will be explained.

## Protect



This chapter highlights your legal and ethical obligations and shows how a combination of gaining consent, anonymising data, gaining clarity over who owns the copyright to your data and controlling access can enable the ethical and legal sharing of data.

## Archive & Publish



When you arrive at this chapter you will have learnt to differentiate between currently available data publication services. You will also find a number of stepping stones on how to promote your data.

## Discover



How can you discover and reuse existing or previously collected datasets?

# Con un supporto pratico

ALLA FINE D OGGNI MODULO  
TROVATE «ADAPT YOUR  
DMP» PER APPLICARE I  
CONCETTI CHE AVETE  
APPENA IMPARATO



## ⊕ Versioning

## ⊖ Interoperability

In order to be able to link your work to other research, it might be useful to build on established terminologies as well as commonly uses coding and soft- and hardware wherever this is possible.

- Which *software and hardware* will you use? How does this relate to other research?

If applicable:

- Will established *terminologies/ontologies* (i.e. structured controlled vocabularies) be used in the project? If not, how does yours relate to established ones?
- Which *coding* is used (if any)? How does this relate to other research?

## Adapt your DMP: part 6

This is the sixth 'Adapt your DMP' section in this tour guide. To adapt your DMP, consider the following elements and corresponding questions:

## ⊖ Deposit your data

- Will the data you produce and/or used in the project be useable by third parties, in particular after the end of the project?
- Which data and associated metadata, documentation and code will be deposited?
- What methods or software tools are needed to access the data?
- Is documentation about the software needed to access the data included?
- Is it possible to include the relevant software (e.g. in open source code)?
- What data quality assurance processes will you apply?

[DMP]

# Formazione



research  
data  
netherlands

*Essentials 4  
Data Support*

[Essentials4data](#)

ABOUT THE COURSE >

START THE COURSE >

LOGIN >

I - A bird's-eye view

Data jargon

DOI

FAIR data

GDPR

Integrity

Linked data

Metadata

Open data

Open science

Persistent identifier (PID)

Preferred format

I - A bird's-eye view >

II - Planning phase >

III - Research phase >

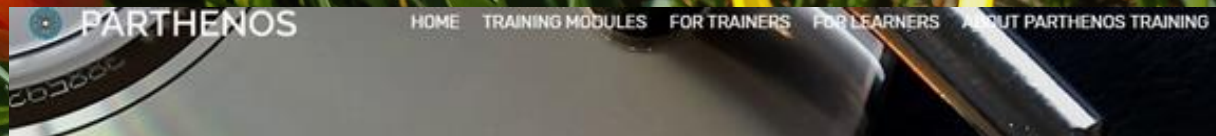
IV - Harvest phase >

V - Legislation and policy >

VI - Data support >

Closing remarks

# Formazione (scienze umane)



Parthenos

## MANAGE, IMPROVE AND OPEN UP YOUR RESEARCH AND DATA

SHARE

### About the module

This module will look at emerging trends and best practice in data management, quality assessment and IPR issues

We will look at policies regarding data management and their implementation, particularly in the framework of a Research Infrastructure

#### Learning Outcomes

By the end of this module, you should be able to:

- Understand and describe the FAIR Principles and what they are used for

#### BROWSE

- Introduction to Research Infrastructures
- Management Challenges in Research Infrastructures
- Introduction to Collaboration in Research Infrastructures
- Manage, Improve and Open up your Research and Data**
- Introduction to Research Data Management

### How does humanities data tend to be different?

There are problems with sharing and managing the humanistic data, however. First of all, much of it is not digital. Humanists still tend to gravitate toward multimodal knowledge creation systems, hybrid digital and technical worlds that resist norms of deposit and reuse. Second, the semiotic systems of humanities data can be quite personal and individual: we prepare our sources to be useful for us, and what works for our research questions and personal epistemic instruments may not work at all for anyone else. Finally, and perhaps most importantly, cultural data is seldom if ever 'raw,' and seldom, if ever, under the sole ownership of the researcher him or herself. The records of human activity and creativity belong to everyone and no one, they are often preserved and curated by dedicated public institutions or private publishers. Whatever humanities data is, it is

### TRAINING MODULES

- Introduction to Research Infrastructures
- Management Challenges in Research Infrastructures
- Introduction to Collaborations in Research Infrastructures
- Manage, Improve and Open Up Your Research Data
- Formal Ontologies: A Complete Novice's Guide
- Digital Humanities Research Questions and Methods

# ...formazione per una ricerca

ri



## Welcome

### The Turing way

The Turing Way is an open source community-driven guide to reproducible, ethical, inclusive and collaborative data science.

Our goal is to provide all the information that data scientists in academia, industry, government and the third sector need at the start of their projects to ensure that they are easy to reproduce and reuse at the end.

The book started as a guide for reproducibility, covering version control, testing, and continuous integration. However, technical skills are just one aspect of making data science research "open for all".

In February 2020, *The Turing Way* expanded to a series of books covering reproducible research, project design, communication, collaboration, and ethical research.

### 3. Reproducible and open research

**What does this mean?** Work with computational methods occurs across a wide spectrum in the humanities, from rigorously empirical to experimental and exploratory or creative approaches (Smithies, 2017). Accordingly, different degrees of reproducibility — intended as the ability to reproduce comparable results with the same data and same analysis methods — apply to the differing research processes.

**What is the issue/context?** Reproducible and open research depends upon access to data and appropriate computational infrastructure. New computational methods and approaches are being undertaken in different computational environments, some of which are in closed and/or proprietary infrastructure and others in open and/or public infrastructure. Very often new and exciting computational methods are time-consuming and hard to implement, or have dependencies that are difficult to establish, raising additional barriers to making research reproducible. The desire for reproducibility follows a broader trend in scientific disciplines and data science in particular<sup>67</sup>, and more investment is needed in regard to sustainability of code and open source frameworks. Journals such as the *Journal of Open Humanities Data*<sup>68</sup> and *Research Data in the Humanities*<sup>69</sup>, which focus on the publication of digital research objects and their critical description, do exist, but they are niche venues and their awareness among humanities researchers is still low.

**What do we propose and for whom?** We promote transparent and reproducible research in the humanities, covering data, code, workflows (Liu, 2017), computational environments, methods and documentation. We encourage partnerships and initiatives involving humanities



PRATICHE DI  
RIPRODUCIBILITÀ

The 2020  
Alan Turing  
Institute

Humanities and data science  
special interest group

**The challenges and prospects  
of the intersection of humanities  
and data science:**  
A white paper from  
The Alan Turing Institute

Welcome

Guide for Reproducible Research

Overview

Open Research

Version Control

Licensing

Research Data Management

Reproducible Environments

BinderHub

Code quality

Code Testing

Code Reviewing Process

Continuous Integration

Reproducible Research with  
Make

Research Compendia

Risk Assessment

Case Studies

Guide for Project Design

Guide for Communication

Guide for Collaboration

Guide for Ethical Research

Community Handbook

Afterword

# Formazione SSHOC

The SSH Training Discovery Toolkit provides an inventory of training materials relevant for the Social Sciences and Humanities.

Use the search bar to discover materials or browse through

## Curated topic

- Research data management/FAIR data (40)
- Didactics (28)
- Open Science (19)
- Quantitative analysis (14)
- Copyright (13)
- Data visualization (9)
- Survey data (8)
- Citizen science (7)
- Digital edition (7)
- Programming with R (6)
- Python/Jupyter (6)
- Text encoding and TEI (5)
- GIT (4)
- Spatial data (4)

## Library Carpentry

Library Carpentry workshops teach people working in library- and information-related roles how to: [The Carpentries](#)

## MOOC Delivering Research Data Management Services

Collaboratively developed MOOC by RDNL, DCC and the University of Edinburgh, published on FutureLearn [Digital Curation Centre Edinburgh](#)

## MOOC on SSHOC services



## SSH Training Community

[Home](#) [Training](#) [SSHOC In Action](#) [Resources](#) [News & Events](#)

## National Coordination Point Research Data Management: RDM Advice and Tips

[National Coordination Point Research Data Management \(LCRDM\)](#) is a network of experts in the field of research data management. We're a worldwide community of trainers who collaborate to improve our professional capabilities by sharing our expertise and resources. Our focus is on trainers serving the Social Sciences and Humanities communities. Our activities are centred around tools and services which are offered through the European Open Science Cloud (EOSC) and are valuable for SSH trainers. By bringing together experts from multiple disciplines, we are looking to create synergies and exploit our collective knowledge and skills for mutual benefit.

## OpenAIRE Training and Support

[OpenAIRE Training and Support](#) provides training and support for scholarly communication towards open and transparent research and innovative research. **Member Benefits**

- Monthly community calls where we exchange and discuss various topics of interest
- Opportunities to create new training materials to support and enhance the use of SSH tools and services (We have budget!)
- A guaranteed insert in the international SSH Trainer Directory which will increase individual visibility (Directory in development)
- Preferential access to [Train-the-Trainer Bootcamps](#), and to [workshops, webinars and conferences](#)
- A standing invitation to contribute personal resources to the [SSHOC Training Discovery Toolkit](#) and advise on improvements
- Exclusive access to a communal mailing list and Google drive to facilitate networking and information sharing with your peers

[SSHOC Workshop Notes: Citizen Science & Cultural Heritage. Planning for Success](#)  
If you ever played with the idea of crowdsourcing your project, but you took a...

[New Report Published: SSHOC Speech-to-text Workshop – Linking Social Survey and Linguistic Infrastructures through speech interviews](#)  
Idea of collaboration between research infrastructures and disciplines is...

[Workshop notes: Code of Conduct for the Social Sciences and Humanities](#)  
Written by Veronika Keck, GESIS – Leibniz Institute for the Social Sciences,...

[New Report Published: Delivery of user-validated Knowledge Graph, and Election Studies Analytics dashboard](#)  
A new SSHOC report titled

# Formazione

## Managing and Sharing Research Data: A Guide to Good Practice

by Louise Corti, Veerle Van den Eynden, Libby Bishop and Matthew Woollard

Second Edition

### Student Resources

#### 1. Discovery & Planning

##### Videos

Case Studies

Weblinks

Tools and Templates

Checklists

Answers To In-chapter Exercises

#### 2. Data Collection

#### 3. Data Processing & Analysis

#### 4. Publishing & Sharing

#### 5. Preserving Data

#### 6. Reusing Data

### Videos

#### Research data lifecycle

Video visualizing the data-related activities typically undertaken in the research data lifecycle. The data lifecycle covers the stages in the existence of digital data: discovery and planning, collection, processing and analysis, publishing and sharing, preserving and reusing.

#### Write a data management plan

Video tutorial on how to write a data management plan, for example for a research grant application.

#### Data skills: providers of international data

An overview of international governmental organizations such as the International Monetary Fund, the Organisation for Economic Co-operation and Development and the International Energy Agency (IEA) that provide aggregate social and economic data between countries.

- Which topics are covered by the World Development Indicators
- Which organization publishes international comparable data

#### The what, why and how of data management planning

Video explaining what data management planning is, how you go about it, and how it can help you. The video illustrates how, when designing design, you can plan which data to collect, how to store them, how to describe them, and how to share them with your colleagues.

### Handbook online resources

### Student Resources

#### 1. Discovery & Planning

Videos

Case Studies

Weblinks

Tools and Templates

Checklists

Answers To In-chapter Exercises

#### 2. Data Collection

#### 3. Data Processing & Analysis

#### 4. Publishing & Sharing

#### 5. Preserving Data

#### 6. Reusing Data

### Tools and Templates

#### DMP online

Web-based tool developed by the Digital Curation Centre, designed to help researchers develop data management plans according to the requirements of major research funders, publishers or institutional requirements. Using the tool, one can create, store, update and share multiple versions of a data management plan at the grant application stage and during the research lifecycle. Plans can be customized according to funder or institution, and exported in a variety of formats. Funder- and institution-specific best practice guidance is provided to users via a range of tailored templates.

#### DMPTool

Online tool developed by the California Digital Library to help researchers generate data management plans required by funders. The tool allows researchers to select their institution and research funder and presents a plan template according to that funder's requirements. Funder-specific and institution-specific guidance and resources for each topic are included. Plans can be exported or shared online.

#### Data Stewardship Wizard

Online tool to develop Data Management Plans for FAIR Open Science, through questions, hints, external resources and community help.

#### FAIR self-assessment tool

Online tool to assess the FAIRness of a dataset, i.e. how Findable, Accessible, Interoperable and Reusable an existing dataset is, and to determine how to enhance its FAIRness.

- How can a data management plan contribute to research transparency?



# Imparare a gestire

**FOSTER** [About](#) [Resources](#) [Events](#) [Courses](#) [News](#)

## Managing and Sharing Research Data

Data-driven research is becoming increasingly common in a wide range of academic disciplines, from Archaeology to Zoology, and spanning Arts and Science subject areas alike. To support good research, we need to ensure that researchers have access to good data. Upon completing this course, you will:

- understand which data you can make open and which need to be protected
- know how to go about writing a data management plan
- understand the FAIR principles
- be able to select which data to keep and find an appropriate repository for them
- learn tips on how to get maximum impact from your research data

[Start the Free Course](#)

**Full details**  
**Level of knowledge:** Introductory: no previous knowledge is required

**Topics**

# ...una via veloce

Au Loup Garou Gourmand  
La Maison des  
100 Bières Bretonnes

eScience vidensportal Video 2019

eScience [Få styr på data](#) [Supercomputing](#) [Træningskurser](#) [Om os](#) [Podcasts](#)

[Få styr på data](#) » eLearning course about the importance of good research data management (RDM)

### eLearning course about the importance of good research data management (RDM)

Within the framework of the Danish National Forum for Data Management, the Danish Universities have developed the eLearning course "Research Data Management".

## 90%

of the world's data was created within the last two years

## Take the course

### Module 1: Introduction



**Reference:** Vlachos, E., Larsen, A.V., Zürcher, S., Hansen, A.F. (2019). 'Introduction'. In: Holmstrand, K.F., den Boer, S.P.A., Vlachos, E., Martínez-Lavanchy, P.M., Hansen, K.K. (Eds.), *Research Data Management* (eLearning course). doi: 10.11581/dtu.00000048

### Module 2: FAIR principles



**Reference:** Martínez-Lavanchy, P.M., Hüser, F.J., Buss, M.C.H., Andersen, J.J., Begtrup, J.W. (2019). 'FAIR Principles'. In: Holmstrand, K.F., den Boer, S.P.A., Vlachos, E., Martínez-Lavanchy, P.M., Hansen, K.K. (Eds.), *Research Data Management* (eLearning course). doi: 10.11581/dtu.00000049

### Module 3: Data Management Plans




**Reference:** den Boer, S.P.A., Buss, M.C.H., Hüser, F.J., Smed, U. (2019). 'Data Management Plans'. In: Holmstrand, K.F., den Boer, S.P.A., Vlachos, E., Martínez-Lavanchy, P.M., Hansen, K.K. (Eds.), *Research Data Management* (eLearning course). doi: 10.11581/dtu.00000050

# ...e le scienze umane?

DARIAH-CAMPUS Resources Topics Sources Course Registry About **May 2019**

## DARIAH Pathfinder to Data Management Best Practices in the Humanities

Written by Erzsébet Tóth-Czifra May, 03 2019 Source: DARIAH Pathfinders, DARIAH Topics: Data management



**1. Why research data management?**

Systematically planning how you will collect, document, organize, manage, share and preserve your data has many benefits. It helps to build a common framework of understanding with your

### TABLE OF CONTENTS

1. Why research data management?
2. Data in the Humanities
3. The devil is in the context: a processual view on data curatic
4. Sharing your data
  - 4.1. Cite to be cited!
  - 4.2. Be aware of your licensing options
  - 4.3. A case study: different levels of being an open scholar
5. A recipe for your research project: the Data Management Plan
6. Data in publications and data as publications
  - 6.1. The networked publication: interlinking the underlying data with your papers
  - 6.2. Data journals in humanities



<b>10. THE RISK OF LOSING THE THICK DESCRIPTION: DATA MANAGEMENT CHALLENGES FACED BY THE ARTS AND HUMANITIES IN THE EVOLVING FAIR DATA ECOSYSTEM</b>	<b>235</b>
<i>Erzsébet Tóth-Czifra</i>	
Realising the Promises of FAIR within Discipline-Specific Scholarly Practices	235
A Cultural Knowledge Iceberg, Submerged in an Analogue World	237
Legal Problems that Are Not Solely Legal Problems	239
The Risk of Losing the Thick Description upon the Remediation of Cultural Heritage	242
The Scholarly Data Continuum	247
Data in Arts and Humanities — Still a Dirty Word?	250
The Critical Mass Challenge and the Social Life of Data	251
The Risk of Losing the <i>Thick Description</i> — Again	255
Conclusions: On our Way towards a Truly FAIR Ecosystem for the Arts and Humanities	258
Bibliography	263

Edmond, 2020



# Dati e scienze umane

## 1. Methodological frameworks and epistemic cultures.

We call for the use of a common methodological terminology in research at the intersection between humanities and data science, and for a wider use of shared research protocols across these domains. We recommend that authors make the methodological framework that they are using explicit in their publications, and we call for inclusive research practices to be fostered across research projects.

## 2. Best practices in the use and evaluation of computational tools.

We encourage practices that ensure transparency and openness in research, and training programmes that help to choose the most suitable computational tools and processes in humanities research. We also call for computational tools to be evaluated in a dialogue between data scientists and digital humanists.

## 3. Reproducible and open research.

We promote transparent and reproducible research in the humanities workflows, computational environments, methods, and documentation and academic institutions should put in place further incentives for hum publish the digital resources, code, workflows and pipelines they create outputs, e.g. in the form of publications in data journals.

## The Alan Turing Institute

Humanities and data science  
special interest group

### The challenges and prospects of the intersection of humanities and data science:

A white paper from  
The Alan Turing Institute

– **History and critique of data science.** This area analyses the characteristics of data science work, sometimes with a focus on the historicity of datasets, and tackles ethical and methodological questions aimed at improving current practices, for example, on issues such as diversity (D'Ignazio and Klein, 2020) and privacy. It also problematizes the very definition of data, considering their complexities, their inherent biases, their contextual and historical natures, in a critical and nuanced way (e.g. Drucker, 2011). Exemplary research in this area includes, among many others, MacKenzie (2017)'s study of the interface between machine learning and critical thought, Kaltenbrunner (2014; 2015)'s studies of infrastructure as a relational and emergent phenomenon that shapes data-driven humanities research and researchers, and the questions they can ask, in complex ways, and Noble (2018)'s study of how white patriarchy and algorithmic bias has resulted in the misrepresentation of women of colour and minorities in search engine results.

– **Algorithmic creativity and cultural innovation in the arts and humanities.** This area focuses on computational creativity, aiming to perform creative tasks with the aid of machines and to explore the plasticity of digital forms for delivering new radical ways of representation and mediation of the arts and humanities. This is the focus of various initiatives such as the Turing AI & Arts group<sup>61</sup> and the Creative Informatics programme in Edinburgh<sup>62</sup>. Individual artists are also increasingly including computational methods in their practice<sup>63</sup>. Examples of projects undertaken in collaboration with academic researchers, creative industries and the GLAM sector include the King's Digital Lab Digital Ghost Hunt<sup>64</sup> experience and the AI and Storytelling project<sup>65</sup>.

# Data Management - caveat



[DMP]

Therefore, it doesn't necessarily matter if you plan to share your data with other scholars, what matters is considering this prospect as you work out how you are going to go about your research. It will help you to understand what it is you are doing more clearly and give you the basis to share that data later on if you so wish.

**PORT**  
postgraduate online  
research training  
PORT DMP

SCHOOL OF  
ADVANCED STUDY  
UNIVERSITY  
OF LONDON

**NON IMPORTA SE ALLA FINE CONDIVIDERETE I DATI O NO.  
QUI SI DOCUMENTANO IL PROCESSO DI RICERCA E LE SCELTE DI METODO**

# Data management ABC – Per partire

Ask yourself this:

[DMP]

What is needed to validate the results of your research?

If you were to produce an article researching, for example, the criminal underclass in early-twentieth century New York, what data would you need to include for someone else to replicate your results? Think about it in terms of your own research.

A bibliography would be the most immediate and obvious starting point, revealing to the reader all the sources that you have used to base your research. But what of the gathering mechanisms you used? Did you create a database or undertake statistical analysis? If so you need to make the database and statistics available. This doesn't just mean providing the files in a readable format, but to provide documentation and to make sure that the data is clearly identified with explicit headings, well-structured, and easily identified.

Focusing on what is needed for validation and re-use, rather than the obvious attributes of research data, is useful. It helps you to think through the process of research from a different perspective and what it is you have actually done to come to your conclusions. It also allows you to show the process you have undertaken; revealing how valuable your approach might be and making the

COSA SERVE A VALIDARE  
LA MIA RICERCA?  
TUTTO QUESTO VA  
INSERITO NEL DMP.  
PROSPETTIVA DIVERSA  
SULLA VOSTRA RICERCA



# Data management ABC – File naming

## EXERCISE ONE

### FILE NAMING

1. Read through the following file names.
2. If you returned to this data folder in a year's time do you think you would be able to recognise what each of these files contains?
3. What information do you think you need in a file name in order to identify what is in the file's contents?

FRA UN ANNO  
SAPRESTE DIRE  
COSA  
CONTENGONO?

 Doc. 1	 My data
 IMPORTANT	 My Passwords
 Thesis Final final	 Thesis version 12
 My study	 Data chart for interviews
 Interview with Jane	 Int 1 (2)

# Data management ABC – File naming

[DMP]


SCEGLIERNE UNA... ED ESSERE  
CONSISTENTI!

## File naming conventions

The conventions comprise the following 13 rules. Follow the links for examples and explanations of the rules.

1. Keep file names short, but meaningful
2. Avoid unnecessary repetition and redundancy in file names and file paths.
3. Use capital letters to delimit words, not spaces or underscores
4. When including a number in a file name always give it as a two-digit number, i.e. 01-99, unless it is a year or another number with more than two digits.
5. If using a date in the file name always state the date 'back to front', and use four digit years, two digit months and two digit days: YYYYMMDD or YYYYMM or YYYY or YYYY-YYYY.
6. When including a personal name in a file name give the family name first followed by the initials.
7. Avoid using common words such as 'draft' or 'letter' at the start of file names, unless doing so will make it easier to retrieve the record.
8. Order the elements in a file name in the most appropriate way to retrieve the record.
9. The file names of records relating to recurring events should include the date and a description of the event, except where the inclusion of any of either of these elements would be incompatible with rule 2.
10. The file names of correspondence should include the name of the correspondent, an indication of the subject, the date of the correspondence and whether it is incoming or outgoing correspondence, except where the inclusion of any of these elements would be incompatible with rule 2.
11. The file name of an email attachment should include the name of the correspondent, an indication of the subject, the date of the correspondence, 'attach', and an indication of the number of attachments sent with the covering email, except where the inclusion of any of these elements would be incompatible with rule 2.
12. The version number of a record should be indicated in its file name by the inclusion of 'V' followed by the version number and, where applicable, 'Draft'.
13. Avoid using non-alphanumeric characters in file names.

File naming



THE UNIVERSITY of EDINBURGH

RECORDS MANAGEMENT

Records Management home

- Guidance
- Managing records
- Practical guidance
- Naming conventions**

Home > Records Management > Guidance > Managing records

### Naming conventions

Make finding electronic records easier.



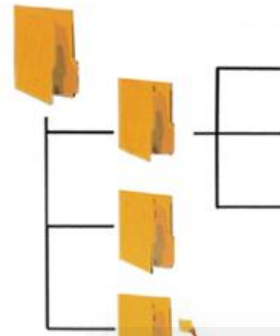
# [DMP]

## Data management ABC – File naming

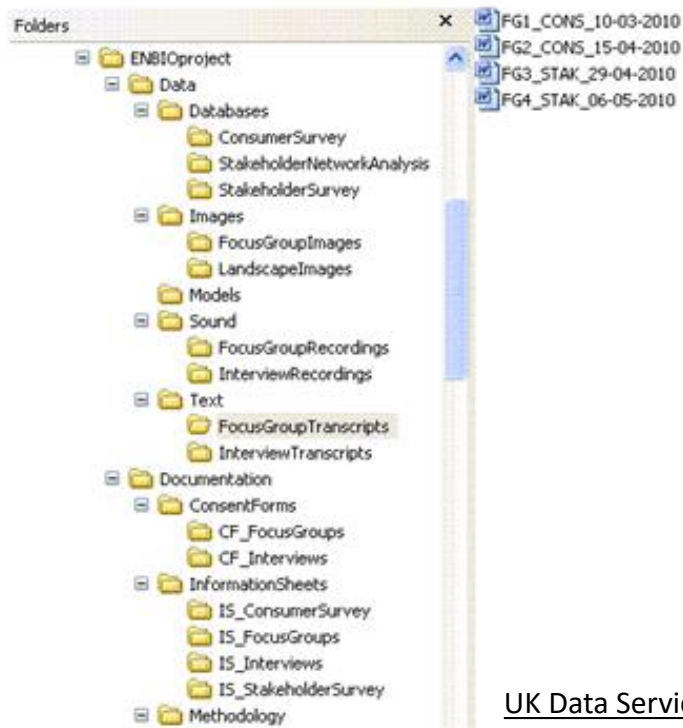
### Folder structure

Structuring your data files in folders is important for making it easier to locate and organise files and versions. A proper folder structure is especially needed when collaborating with others.

### CESSDA training



It helps to restrict the level of folders to three or four deep and not to have more than ten items in each list.



UK Data Service

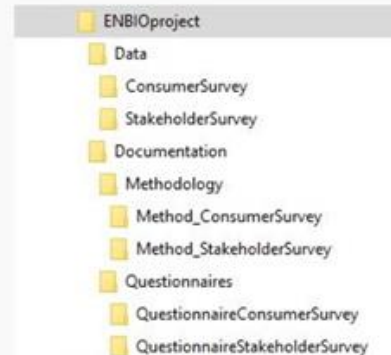
to organise your data plan and organisation of al relevant to the data to the data folders, information on the data processing procedures.

erarchy of your files and ep or shallow hierarchy is ve several independent advisable to create a separate data folder look at the examples in the accordion below



### Survey data

For this survey, data and documentation files are held in separate folders. Data files are to data type and then according to research activity. Documentation files are organised documentation file and research activity. It helps to restrict the level of folders to three more than ten items on each list.



# Data management A Readme file

A readme file provides information about a dataset and is intended to help ensure that the data can be correctly interpreted, by yourself at a later date or by others when sharing or publishing data.

A readme file must be submitted along with the dataset file(s).

The outline below should be completed with information relevant to the submitted dataset.

#### Best practices

- **Create one readme file for each dataset**
- **Name the file README;** not readme, read\_me, ABOUT, etc.
- **Write your readme document as a plain text file;** save as README.txt or README.md when writing in [Markdown](#). Or use README.pdf when text formatting is important for your file.

[DMP]

## 1. Introductory information

- **Title of the dataset**
- **For each file or group of similar files, a short description of what data it contains**
- Explain the file naming convention, if applicable
- Format of the file if not obvious from the file name
- If the data set includes multiple files that relate to each other, the relationship between the files or a description of the file structure that holds them
- Contact information; in case users have questions regarding the data files

## 2. Methodological information

- **Method description for collecting or generating the data, as well as the methods for processing data, if data other than raw data are being contributed**
- Any instrument-specific information needed to understand or interpret the data
- Software (including version number) used to produce, prepare, render, compress, analyze and/or needed to read the dataset, if applicable
- Standards and calibration information, if appropriate

## 3. Data specific information

- **Full names and definitions (spell out abbreviated words) of column headings for tabular data**
- **Units of measurement**
- **Definitions for codes or symbols used to record missing data**
- **Specialized formats or abbreviations used**

## 4. Sharing and Access information

- Licenses or restrictions placed on the data; Licenses allow you to specify the 'terms-of-use' for your data. The archive provides a license that is explained in its [terms of use](#) and applies this license as default selection. You can use this [licensing wizard](#) to help you to pick a more appropriate license for the use of your data. This license will then be displayed in the metadata.

# [es. di cosa documentare]

## Structured tabular data should have as documentation (where applicable):

- variable names, labels and descriptions (maximum 80 characters)
- units of measurement for variables
- reference to the question number of a survey or questionnaire

Example: variable 'q11hexw' with label 'Q11: hours spent taking physical exercise in a typical week' — the label gives the unit of measurement and a reference to the question number (Q11)

- value code labels

Example: variable 'p1sex' = 'sex of respondent' with codes '1=female', '2=male', '8=don't know', '9=not answered'

- coding and classification schemes explained, with a bibliographic and dated reference (some standards change over time)

Examples: Standard Occupational Classification, 2000 — a series of codes to classify respondents' jobs; ISO 3166 alpha-2 country codes — an international standard of 2-letter country codes

- codes for missing data, with reason data are missing (blanks, system-missing or '0' values are best avoided)

Example: '99=not recorded', '98=not provided (no answer)', '97=not applicable', '96=not known', '95=error'

- deviating universe information for variables in case of skipped cases or questions
- derived or constructed variables created after collection, giving code, algorithm or command files used to create them — simple derivations, such as grouping age data into age intervals, can be explained in the variable and value labels; complex derivations can be described by providing the algorithms, logical statements or functions used to create derived variables, such as the SPSS or Stata command

hse09ai.sav [DataSet2] - PASW Statistics Data Editor

	Name	Type	Width	Decimals	
175	quala10	Numeric	2	0	Which of the
176	activb	Numeric	2	0	Activity status
177	empstat	Numeric	2	0	Manager/Fore
178	everjob	Numeric	2	0	Ever had paid
179	ftptime	Numeric	2	0	Full-time or pa
180	howlong	Numeric	2	0	How long have
181	wkstr12	Numeric	2	0	Able to start w
182	wklook4	Numeric	2	0	Looking paid
183	nemplee	Numeric	2	0	Number empk
184	nssec	Numeric	5	1	NS-SEC - lon
185	othpaid	Numeric	2	0	Ever had other employment (waiting to start work)
186	payage	Numeric	3	0	Age when last had a paid job
187	paylast	Numeric	4	0	Year left last paid job
188	paymon	Numeric	2	0	Month last left paid job
189	sclass	Numeric	2	0	Social Class
190	seg	Numeric	2	0	Socio-Economic Group
191	snemplee	Numeric	2	0	Self employed, how many employees
192	age	Numeric	3	0	Age last birthday

Field Name | Data Type | Description

Field Name	Data Type	Description
Farmercode	Text	Farmer identification code
Scheme	Text	Drainage scheme
Transcript	Hyperlink	Transcript of interview
Date	Date/Time	Date of interview
Interviewer	Text	Name of interviewer
Incomplete data	Yes/No	Observation has data missing
Nb holdings	Number	Number of holdings farmed by farmer
Tot Farmsize	Number	Total size business unit (ha)
% area owned	Number	Area owned of business unit, in percentages (%)
Farmtype	Text	Main enterprise of business unit
Organic	Yes/No	Farm is organic
Winter wheat	Number	Winter wheat (ha)
Spring wheat	Number	Spring wheat (ha)
Winter barley	Number	Winter barley (ha)
Spring barley	Number	Spring barley (ha)

	Site	Location	Type	Instrument	From	
2	Beckingham	Beckingham & Idle Baro	Barometer	73937	7/2/2007	18/10/07
3	Beckingham	Beckingham Ditch	Diver	80137	7/2/2007	16/1/07
4	Beckingham	Beckingham Fld Centre	Diver	80136	7/2/2007	16/1/07
5	Beckingham	Beckingham Fld Edge	Diver	80129	7/2/2007	16/1/07
6	Bushley	Bushley Barometer	Barometer	77599	14/2/2007	4/11/07
7	Bushley	Bushley Ditch	Diver	63017	14/2/2007	23/1/07
8	Bushley	Bushley Fld Centre	Diver	53632	14/2/2007	23/1/07
9	Bushley	Bushley Fld Edge	Diver	53194	14/2/2007	12/4/07
10	Cuddyarch Sough	Cuddyarch Sough Baro	Barometer	62943	10/5/2007	30/1/07
11	Cuddyarch Sough	Cuddyarch Sough Fld Centre	Barometer	62963	10/5/2007	30/1/07
12	Cuddyarch Sough	Cuddyarch Sough Fld Edge	Barometer	62969	10/5/2007	30/1/07
13	Cuddyarch Sough	Wedholme Sough (River)	Diver	48432	10/5/2007	30/1/07
14	Idle	Idle Ditch	Diver	80133	7/2/2007	7/11/07
15	Idle	Idle Fld Centre	Diver	80131	7/2/2007	16/1/07
16	Idle	Idle Fld Edge	Diver	80132	7/2/2007	16/1/07
17	Idle	Idle Upland	Barometer	77531	8/2/2007	18/10/07
18	Morda	Morda Baro	Barometer	62975	31/5/2007	29/1/07
19	Morda	Morda Ditch	Barometer	62970	31/5/2007	29/1/07

# tare

This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike International licence (CC BY-NC-SA 4.0). To view a copy of this licence, visit <https://creativecommons.org/licenses/by-nc-sa/4.0/>



Data collection number 0000

**Title**  
**Depositor, A.**

Interview ID	Date of birth /Birth year /Age	Gender	Occupation	Organisation	Marital status	Household ID	Relationship	Country of origin	Interview topics	Notes	Place of interview	Date of interview	No of pages	Text file name	Audio file name

- Notes** (delete these from the final list)
- The nature of the data collection and the chosen anonymisation strategy will affect which fields are to be included in the data list.
  - Fields and columns should be filled in in a consistent format throughout the data list.
  - Bold fields should be seen as a minimum for effective reusability of the data.
  - Italic fields should be used as appropriate, and ideally in the order they appear here.
  - Fields that are relevant for your specific data collection should be added to the table.
  - When the table is completed, remove italics, make all headers bold, align fields, and delete any blank columns.

**DATI QUALITATIVI**

**Study Number 6377**  
**Integrated Floodplain Management, 2006-2008**  
**Morris, J.**

**Floodplain farm survey**

Interview ID	Farmer code	Age	Farm scheme	Farm type	Size of farm (hectare)	Number of holdings	Date of interview	Interviewer name	No of pages	Text file name	Audio file name
1	Be1	35-45	Beckingham	Beef	360	1	04.12.2006	Helena	28	6377int001	6377int001
2	Be2	45-55	Beckingham	Arable	364	1	05.12.2006	Helena	21	6377int002	6377int002
3	Be3	45-55	Beckingham	Arable	372	2	06.12.2006	Helena	22	6377int003	6377int003
4	Be4	45-55	Beckingham	Arable	194	3	06.12.2006	Helena	18	6377int004	6377int004
5	Be5	55-65	Beckingham	Arable	108	1	07.12.2007	Helena	21	6377int005	6377int005
6	Be6	45-55	Beckingham	Arable	1254	2	01.02.2008	Helena	19	6377int006	
7	Bu1	55-65	Bushley	Mixed	101	2	13.02.2007	Quentin	29	6377int007	6377int007
8	Bu2	>65	Bushley	Mixed	97	1	15.02.2007	Quentin	15	6377int008	6377int008
9	Bu3	>65	Bushley	Arable	194	4	13.02.2007	Quentin	21	6377int009	6377int009
10	Bu4	55-65	Bushley	Mixed	202	1	15.03.2007	Helena	19	6377int010	6377int010
11	Cu1	35-45	Cuddyarch	Dairy	64	1	08.05.2007	Helena	19	6377int011	6377int011
12	Cu2	55-65	Cuddyarch	Dairy	189	2	08.05.2007	Helena	18	6377int012	6377int012
13	Cu3	55-65	Cuddyarch	Mixed livestock	76	1	08.05.2007	Helena	13	6377int013	6377int013
14	Cu5	45-55	Cuddyarch	Mixed livestock	198	1	09.05.2007	Helena	24	6377int014	6377int014
15	Cu6	55-65	Cuddyarch	Dairy	89	1	09.05.2007	Helena	14	6377int015	6377int015
16	Cu7	>65	Cuddyarch	Mixed livestock	190	4	11.05.2007	Helena	20	6377int016	6377int016
17	Cu8	55-65	Cuddyarch	Mixed livestock	109	2	11.05.2007	Helena	22	6377int017	6377int017
18	ld1	55-65	Idle	Arable	158	3	07.02.2007	Quentin	17	6377int018	6377int018a
18	ld1	55-65	Idle	Arable	158	3	07.02.2007	Quentin	17	6377int018	6377int018b
19	ld1b	55-65	Idle	Arable	158	3		Quentin	22	6377int019	
20	ld2	45-55	Idle	Dairy	150	1	08.02.2007	Quentin	17	6377int020	6377int020

# [es. di cosa documentare]

## Variable Information Log

UK data service - data documentation

### Introduction

8 For datasets being deposited that include secondary data resources, researchers are advised to prepare a descriptive Variable Information Log describing these resources.  
9 The Variable Information Log should include the variable name, its source, how it was collected, a brief description, and any restrictions noted on its further use. (See the notes below)

### Notes

12 These fields should be completed for the original data sources for each variable:

13	<b>Variable name:</b>	Provide a list of all the variables (name/number) used in the dataset.
14	<b>Variable label:</b>	A brief description necessary to identify the variable.
15	<b>Source:</b>	Source of the dataset/data owner or producer (e.g. World Bank data, IMF data, Penn World Tables data).
16	<b>Dataset version:</b>	Datasets keep evolving, so best practice is to indicate which version has been used.
17	<b>URL/DOI:</b>	Provide a persistent identifier or link of the source dataset used. Alternatively, if the data are not available online, provide a brief description of how they were obtained.
18	<b>License information:</b>	Please indicate the licensing information (type of data), as it is important to ensure that the researchers have permission from the data owners. For example, Open data, Data owned by the researcher (you), Data owned by another researcher or Third party licensed data.
19	<b>Unit of analysis</b>	Indicate the unit of analysis used in the primary dataset (individuals, cases, addresses).
20	<b>Date data downloaded/obtained</b>	It is important to state the date when the dataset was downloaded or obtained and used for analysis. The data source may have been updated since that time.
21	<b>Brief description of the data:</b>	Provide a brief description of the dataset, including what was the aim of the study. If a codebook is publicly available for the data used, provide a link.
22	<b>Data collection method:</b>	Where the data collection procedure for the dataset is well documented, provide a link to that information. If there is little information available, provide a brief description on how data were gathered.

23  
24  
25  
26  
27  
28  
29 This work is licensed under a Creative Commons Attribution-Non-commercial-Share Alike  
30 International licence (CC BY-NC-SA 4.0). To view a copy of this licence, visit  
31 <https://creativecommons.org/licenses/by-nc-sa/4.0/>



# Data management ABC – Versioning



[DMP]

## Data versioning

Unlike the software domain, the data community doesn't yet have a standard numbering system. Three representative data version numbering patterns in use include:

Numbering system 1

Numbering system 2

Numbering system 3

### What tools are available for data versioning?

There is no one-size-fit-all solution for data versioning and tracking changes. Data come in different forms and are managed by different tools and methods. In principle, data managers should take advantage of data management tools that support versioning and track changes.

Example approaches include:

[Git \(and Github\) for Data](#) (with size <10Mb or 100k rows) which allows:

- effective distributed collaboration – you can take my dataset, make changes, and share those back with me (and different people can do this at once)
- provenance tracking (i.e. what changes came from where)
- sharing of updates and synchronizing datasets in a simple, effective, way.

[Data versioning at ArcGIS](#)

- Users of ArcGIS can create a geodatabase version, derived from an existing version. When you create a version, you specify its name, an optional description, and the level of access other users have to the version. As the owner of the version, you can change these properties or delete a version at any time.

## What do we mean by the term 'data versioning'?

A version is "a particular form of something differing in certain respects from an earlier form or other forms of the same type of thing". In the research environment, we often think of versions as they pertain to resources such as manuscripts, software or data. We may regard a new version to be created when there is a change in the structure, contents, or condition of the resource.

In the case of research data, a new version of a dataset may be created when an existing dataset is reprocessed, corrected or appended with additional data. Versioning is one means by which to track changes associated with 'dynamic' data that is not static over time.

## Why is data versioning important?

Increasingly, researchers are required to cite and identify to support research reproducibility and trustworthiness: accurately indicate exactly which version of a dataset particularly challenging where the data to be cited are accessed via a web service.

### [Numbering system 1](#)

Data versioning follows a similar path to software versioning, usually applying a two-part numbering rule: Major.Minor (e.g. V2.1). Major data revision indicates a change in the formation and/or content of the dataset that may bring changes in scope, context or intended use. For example, a major revision may increase or decrease the statistical power of a collection, require change of data access interfaces, or enable or disable answering of more or less research questions. A Major revision may incorporate:

- substantial new data items added to /deleted from a collection
- data values changed because temporal and/or spatial baseline changes
- additional data attributes introduced
- changes in a data generation model
- format of data items a changed
- major changes in upstream datasets.

Minor revisions often involve quality improvement over existing data items. These changes may not affect the scope or intended use of initial collection. A Minor revision may include:

- renaming of data attribute
- correction of errors in existing data
- re-running a data generation model with adjustment of some parameters
- minor changes in upstream datasets.

# Data management ABC – Versioning

University of Leicester

[Version chart](#)

## Good Practice and Guidance – Document Version Control Chart (Draft)

### 1. Create Document/File

- Save the document according to file naming guidance/good practice.

### 2. Document Identification

- Identify on the document e.g. in header or footer, the author, filename, page number and date the document is created/revised.

### 3. Version Control Table

- Versions and changes documented with Version Control Table where significant/formal/project based.

### 4. Version Number

- Current version number identified on the first page and where appropriate, incorporated into the header or footer of the document.
- Version number is included as part of the file name.

### 5. First Draft Version

- Named as version "0-1" (no full stops in electronic file names).
- Subsequent draft versions 0-2, 0-3, 0-4 ...

### 6. First Final/Approved Version

- When document is final/approved it becomes version 1-0.

### 7. Changes to Final Version

- Changed/revised final version becomes x-1.
- Subsequent drafts to Final version become e.g. 1-1, 1-2, 1-3 etc.

### 8. Further Final/Approved Documents

- Version number increased by "1-0" e.g. 1-0, 2-0, 3-0 etc.
- e.g. Amendments to Final 1-0 are 1-1, 1-2, 1-3 and as approved becomes 2-0.

[DMP]

Example version control table:

UK Data Service

Title:	Vision screening tests in Essex nurseries		
File Name:	VisionScreenResults_00_05		
Description:	Results data of 120 Vision Screen Tests carried out in 5 nurseries in Essex during June 2007		
Created By:	Chris Wilkinson		
Maintained By:	Sally Watsley		
Created:	04/07/2007		
Last Modified:	25/11/2007		
Based on:	VisionScreenDatabaseDesign_02_00		
Version	Responsible	Notes	Last amended
00_05	Sally Watsley	Version 00_03 and 00_04 compared and merged by SW	25/11/2007
00_04	Vani Yussu	Entries checked by VY, independent from SK	17/10/2007
00_03	Steve Knight	Entries checked by SK	29/07/2007
00_02	Karin Mills	Test results 81-120 entered	05/07/2007
00_01	Karin Mills	Test results 1-80 entered	04/07/2007

# Data management ABC – Versioning

[DMP]



## Version control

Version control can be done through:

- Uniquely identifying different versions of files using a systematic naming convention, such as using version numbers or dates (date format should be YYYY-MM-DD, see '[File naming](#)');
  - Record the date within the file, for example, 20010911\_Video\_Twintowers;
  - Process the version numbering into the file name, for example, HealthTest-00-02 or HealthTest\_v2;
  - Don't use ambiguous descriptions for the version you are working on. Who will know whether MyThesisFinal.doc, MyThesisLastOne.doc or another file is really the final version?
- Using version control facilities within the software you use;
- Using versioning software like [Subversion](#) (2017);
- Using file-sharing services with incorporated version control (but remember that using commercial cloud services as the Google cloud platform, Dropbox or iCloud comes with specific rules set by the provider of these services. Private companies have their own terms of use which applies for example to copyrights);
- Designing and using a version control table. In all cases, a file history table should be included within a file. In this file, you can keep track of versions and details of the changes which were made. Click on the tab to have a look at [an example which was taken from the UK Data Service](#) (2017c).

[CESSDA training](#)



# Data management ABC – Versioning

[DMP]



The screenshot shows the Git website homepage. At the top left is the Git logo (a red diamond with a white 'G' and a red 'i') followed by the text 'git --distributed-even-if-your-workflow-isnt' and the word 'Git' in a serif font. To the right is a search bar with the placeholder text 'Search entire site...'. Below the header, there are two paragraphs of text describing Git as a free and open source distributed version control system. To the right of the text is a diagram showing a branching model with several stacks of books representing code repositories connected by colored lines. At the bottom, there are four navigation links: 'About', 'Documentation', 'Downloads', and 'Community', each with an icon and a brief description. On the right side, there is a monitor displaying the latest source release '2.31.1' and a 'Download 2.31.1 for Windows' button.

 **git** --distributed-even-if-your-workflow-isnt Git

Git is a [free and open source](#) distributed version control system designed to handle everything from small to very large projects with speed and efficiency.

Git is easy to learn and has a [tiny footprint](#) with lightning fast performance. It outclasses SCM tools like Subversion, CVS, Perforce, and ClearCase with features like [cheap local branching](#), convenient staging areas, and [multiple workflows](#).



 **About**  
The advantages of Git compared to other source control systems.

 **Documentation**  
Command reference pages, Pro Git book content, videos and other material.

 **Downloads**  
GUI clients and binary releases for all major platforms.

 **Community**  
Get involved! Bug reporting, mailing list, chat, development and more.

  
Latest source Release  
**2.31.1**  
Release Notes (2021-03-26)  
[Download 2.31.1 for Windows](#)

# Data management ABC – Data entry

[DMP]



## Data Management Expert Guide

- 1. Plan >
- 2. Organise & Document >
- 3. Process ▼
  - Data entry and integrity
  - Quantitative coding
  - Qualitative coding
  - Weights of survey data
  - File formats and data conversion
  - Data authenticity
  - Wrap up: Data quality
  - Adapt your DMP: part 3
  - Sources and further reading
- 4. Store >
- 5. Protect >
- 6. Archive & Publish >

- ⊕ Check the completeness of records
- ⊕ Reduce burden at manual data entry
- ⊕ Minimise the number of steps
- ⊕ Conduct data entry twice
- ⊕ Perform in-depth checks for selected records
- ⊕ Perform logical and consistency checks
- ⊕ Automate checks whenever possible

# Data management ABC –

The UK Data Service developed a free easy-to-use open source tool known as **QAMyData** that provides a **health check for numeric data**. The tool uses automated methods to detect and report on some of the most common problems in survey or numeric data, such as missingness, duplication, outliers and direct identifiers. Requirements were scoped through a series of engagement exercises with the Service's own data curation team, other data publishers, managers and quantitative researchers to create a comprehensive list of 'tests' that are typically used when quality assessing numeric data files.

[QAMydata](#)

The tool offers a number of configurable tests that have been categorised into four types: file, metadata, data integrity, and identifiers, which can be run on popular file formats, including SPSS, Stata, SAS and CSV. A standard *config* file has default settings for each test, such as a threshold for pass or fail on various tests (e.g. detect value label that are truncated, email addresses identified as a string, or undefined missing values) which can be easily adapted to meet the user's own desired thresholds. The configuration feature allows the creation of a unique **Data Quality Profile**. The software creates a '**data health check**' that details errors and issues as both a summary and detailed report, providing a location of the failed test. New tests can easily be added. Data depositors and publishers can act on the results and resubmit the file until a clean bill of health is produced.

«STATO DI SALUTE»  
DATI CON CHECK  
PERSONALIZZABILI

QAMyData: Table of QA tests included (V1.0)

Type of check available	Specific test	User note
Basic file checks	File opens	Checks whether acceptable format
	Bad filename check, regular expression via RegEx pattern	Regex requires quotes "[a-z]". To use a special characters, e.g. a backslash (\) a backslash before is required e.g. \\
Metadata checks	Report on number of cases and variables	Always run
	Count of grouping variables	
	Missing variable labels	Must be set to true, or the test will not run
	No label for user defined missing values e.g. -9 not labelled	SPSS only
	'Odd' characters in variable names and labels	User specifies the characters
	'Odd' characters in value labels	User specifies the characters
	Maximum length of variable labels, e.g. >79 characters	User specifies the length
Maximum length of value labels, e.g. >39 characters	User specifies the length	

Data integrity checks	Report number of numeric and string variables	
	Check for duplicate IDs	User specifies the variables. Multiple variables can be added on new lines e.g. Caseno or AnotherVariable
	'Odd' characters in string data	User specifies the characters
	Spelling mistakes (non-dictionary words) in string data using a dictionary file	User specifies a dictionary file
	Percentage of values missing ('Sys miss' and undefined missing)	User sets the threshold, e.g. more than 25%
Disclosure risk checks	Identifying disclosure risk from unique values or low thresholds (frequencies of categorical variables or minimum values)	User sets the threshold value, e.g. 5
	Direct identifiers using a RegEx pattern search	User runs separately for postcodes, telephone numbers etc. Advise tests run separately as resource intensive

Useful checks to add	Specific test	User note
Metadata checks	Export a Code book DDI	
Data integrity checks	Expected format for a variable for coded data	User specifies field type/format e.g. ICD code
	Values outlying the listed code values	

## Basic File Checks

Name	Status (N)	Description
Bad file name	Failed (1)	File name should match the user specified pattern

## Metadata Checks

Name	Status (N)	Description
Missing variable labels	Failed (3)	Variables should have a label
Variable odd characters	Failed (2)	Variable names and labels should not contain the specified characters ["@", "#", "!", "%", "&", ":", ";", "<", ">"]
Variable label max length	Failed (3)	Variable labels should not exceed the defined number of characters (79 characters)

# Data Management ABC - conservazione

[DMP]

LUNGO O BREVE  
TERMINE?

## Checksum Checker

Software for Digital Preservation

Download version 3.0.1, released 25 March 2014 AEST

Checksum Checker is free and open source software developed by the National Archives of Australia. Checksum Checker is a piece of software that is used to monitor the contents of a digital archive for data loss or corruption.

Checksum Checker is a component of the Digital Preservation Software Platform (DPSP).

### Features

As part of the Digital Preservation Recorder (DPR) workflow, checksums are generated for each Archival Information Package (AIP). Checksum Checker generates a new checksum for each AIP and compares it against the stored checksum. If the checksums do not match, then the AIP is flagged as being corrupt.

Checksum Checker incorporates the following features:

- Checksum Checker functions as a service.
- Checksum Checker sends automated emails to a nominated administrator email address, coinciding with certain events (such as the start of a checking run or when an error is encountered).

Checksum Checker is released under the GPLv3, and is available for download. <http://checksumchecker.sourceforge.net/>

Search



Home



Download



Docs



F.A.Q



Licensing



External Links



Contact Us

Storage Solutions	Advantages	Disadvantages	Suitable for
<b>Personal Computer &amp; Laptop</b>	<i>Always available</i>  <i>Portable</i>	<i>Drive may fail</i>  <i>Laptop may be stolen</i>	<i>Temporary storage</i>
<b>Networked drives</b>  File servers managed by your university, research group or facilities like a NAS-server	<i>Regularly backed up</i>  <i>Stored securely in a single place</i>	<i>Costs</i>	<i>Master copy of your data</i>  <i>(if enough storage space is provided ..)</i>
<b>External storage devices</b>  USB flash drive, DVD/CD, external hard drive	<i>Low cost</i>  <i>Portability</i>	<i>Easily damaged or lost</i>	<i>Temporary storage</i>
<b>Cloud services</b>	<i>Automatic synchronization between folders and files</i>  <i>Easy to access and use</i>	<i>It's not sure whether data security is taken care of</i>  <i>You don't have direct influence on how often backups take place and by whom</i>	<i>Data sharing</i>

1

2

3

4

5

6

Organize and document research data. Make digital versions of paper data documentation in a PDF/A format (suitable for long-term storage).

# Data Management ABC- backup and storage

[DMP]

## Portable devices

## Cloud storage

## Local storage

## Networked drive



Laptops, tablets, external hard-drives, flash drives and Compact Discs

### Advantages

- Allow easy transport of data and files without transmitting them over the Internet. This can be especially helpful when working in the field.
- Low-cost solution.

### Disadvantages/Risks

- Easily lost, damaged, or stolen and may, therefore, offer an unnecessary security risk.
- Not robust for long-term storage or master copies of your data and files.
- Possible quality control issues due to version confusion.

### Precautions (sensitive data)

Use in encrypted password

### Advantages

- Automatic backups.
- Often automatic version control.

### Disadvantages/Risks

- Not all cloud services are secure. May not be suitable for sensitive data containing personal information about EU citizens.
- Insufficient control over where the data is stored and how often it is backed up.
- Free services by commercial providers (e.g. Google Drive, Dropbox) may claim rights to use content you manage and share them for their own purposes.
- Data can be lost if your account is suspended or accidentally deleted, or if the provider goes out of business.

### Precautions for (sensitive) personal data

- Encrypt all (sensitive) personal data before uploading it to the cloud. This is particularly important to avoid conflict with European data protection regulations if you do not know in which countries servers used for storage and backup are located (see 'Security' for more information on encryption; also see 'Protecting data').

### Recommendations

- Do: use cloud services for granting shared, remote and easy access to data and other files to all involved in the project.
- Do: Read the terms of service. Especially focus on rights to use content given to the service provider.
- Do: Opt for European, national, or institutional cloud services which store data in Europe if possible.
  - B2drop (EUDat, n.d.) is an example of a European cloud storage solution.
  - SWITCHdrive (SWITCH, 2017) is a Swiss solution.
  - DataverseNL (Data Archiving and Networked Services, 2017) is an example of a service for Dutch researchers that allows the storage and sharing of data both during and after the research period.
- Don't: make this your only storage and backup solution.
- Don't: use for unencrypted (sensitive) personal data.

CESSDA Guide

CI SONO STRUMENTI DIVERSI PER ESIGENZE DIVERSE (DURANTE/AL TERMINE). DURANTE, DOVETE ANCHE POTERCI LAVORARE CON IL TEAM

# Cosa conservare?



## Establishing criteria for selection decisions

You should establish criteria to guide selection decisions. The DCC's How to Select and Appraise Research Data for Curation[56] proposes seven criteria as outlined below:

1. **Relevance to mission:** the resource content fits any priorities stated in the institution's mission, or funding body policy including any legal requirement to retain the data beyond its immediate use.
2. **Scientific or historical value:** is the data scientifically, socially, or culturally significant? Assessing this involves inferring anticipated future use, from evidence of current research and educational value.
3. **Uniqueness:** the extent to which the resource is the only or most complete source of the information that can be derived from it, and whether it is at risk of loss if not accepted, or may be preserved elsewhere.
4. **Potential for redistribution:** the reliability, integrity, and usability of the data files may be determined; these are received in formats that meet designated technical criteria; and Intellectual Property or human subjects issues are addressed.
5. **Non-replicability:** it would not be feasible to replicate the data/resource or doing so would not be financially viable.
6. **Economic case:** costs may be estimated for managing and preserving the resource, and are justifiable when assessed against evidence of potential future benefits; funding has been secured where appropriate.
7. **Full documentation:** the information necessary to facilitate future discovery, access, and reuse is comprehensive and correct; including metadata on the resource's provenance and the context of its creation

[DMP]

- RILEVANTI PER LA MISSIONE DELL'ENTE
- VALORE STORICO
  - UNICITÀ
- POTENZIALE DI RIUSO
  - NON REPLICABILI
  - COSTO/BENEFICI
- DOCUMENTAZIONE COMPLETA

# Comparare a proteggere

## What are personal data?

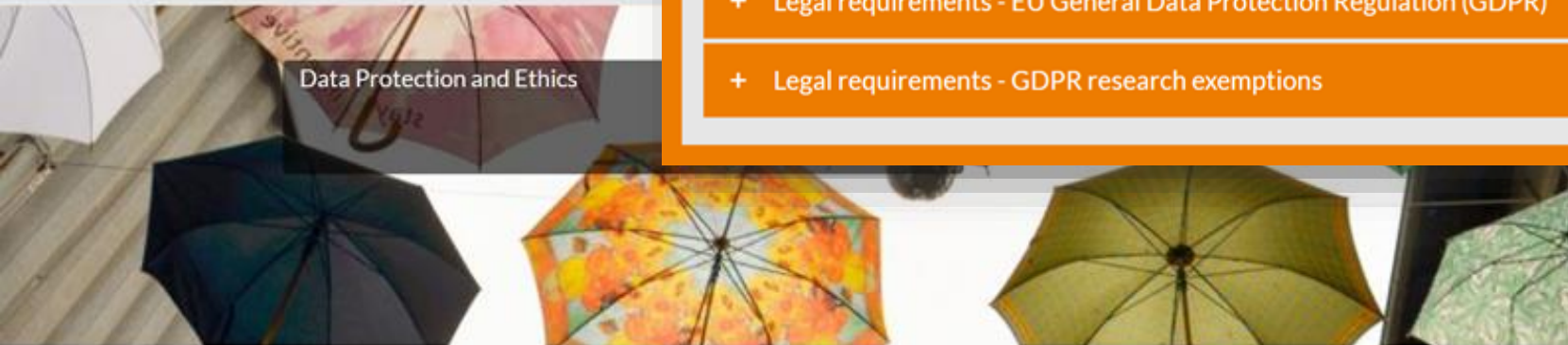
Click the plus sign to expand the text box

- + What are personal data?
- + Protecting personal data
- + Legal requirements - EU General Data Protection Regulation (GDPR)
- + Legal requirements - GDPR research exemptions

## FOSTER data protection



## Data Protection and Ethics



This course covers data protection in particular and ethics more generally. It will help you understand the basic principles of data protection and introduces techniques for implementing data protection in your research processes. Upon completing this course, you will know:

- what personal data are and how you can protect them
- what to consider when developing consent forms
- how to store your data securely
- how to anonymise your data

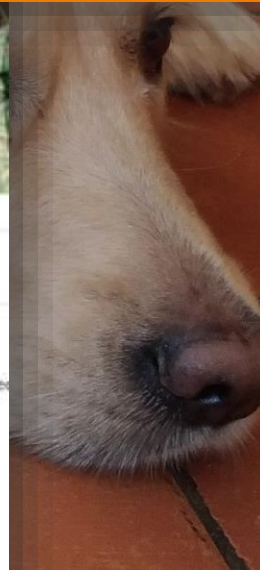
Start the Free Course



## Full details

Level of knowledge: Introductory: no previous knowledge is required

## Topics



# [dati personali]

## ⊖ Legal Basis

Personal data can only be processed when there is a valid legal basis to do so. The GDPR recognises six bases (grounds):

- consent of the data subject
- necessary for the performance of a contract
- legal obligation placed upon the data controller
- necessary to protect the vital interests of the data subject
- carried out in the public interest or in the exercise of official authority (public task)
- legitimate interest pursued by the data controller

### The research exemption

The GDPR contains an exemption which entails that some of the principles above are slightly different when you collect and process personal data for research purposes. This is called the 'research exemption'.

*Processing for archiving purposes in the public interest, scientific or historical research purposes or statistical purposes, shall be subjected to appropriate safeguards, in accordance with this Regulation, for the rights and freedoms of the data subject. Those safeguards shall ensure that technical and organisational measures are in place in particular in order to ensure respect for the principle of data minimisation. Those measures may include pseudonymisation provided that those purposes can be fulfilled in that manner. Where those purposes can be fulfilled by further processing which does not permit or no longer permits the identification of data subjects, those purposes shall be fulfilled in that manner | General Data Protection Regulation, [Article 89](#).*

In practice, this means that Principle II. and V. are less strict. Further processing of personal data for the purposes of archiving, scientific or historical research purposes and statistical purposes is not



ART. 89  
ECCEZIONI PER LA  
RICERCA MA SEMPRE  
SU UNA BASE LEGALE  
(CHE VA ESPLICITATA)



CESSDA guide  
Data Management Expert Guide



# [dati personali]

## I. Process lawfully, fair and transparent



The participant is informed of what will be done with the data and data processing should be done accordingly.

## II. Keep to the original purpose



Data should be collected for specified, explicit and legitimate purposes and not further processed in a manner that is incompatible with those purposes.

## III. Minimise data size



Personal data that are collected should be adequate, relevant and limited to what is necessary.

## IV. Uphold accuracy



Personal data should be accurate and, where necessary kept up to date. Every reasonable step must be taken to ensure that personal data that are inaccurate are erased or rectified without delay.

## V. Remove data which are not used



Personal data should be kept in a form which permits identification of data subjects for no longer than is necessary for the purposes for which the personal data are processed.

## VI. Ensure data integrity and confidentiality



Personal data are processed in a manner that ensures appropriate security of the personal data, including protection against unauthorised or unlawful processing and against accidental loss,

# GDPR e ricerca

## Introduction

The GDPR in research, a.o. special categories of personal data, processing in/outside the European Economic Area (EEA), and privacy by design/default.

- > [GDPR in research: introduction](#)
- > [FAQ GDPR in research](#)

## Data minimisation

The data minimisation principle comprises that data has to be adequate, relevant and limited to what is necessary for the purposes for which they are processed.

- > [GDPR in research: data minimisation](#)
- > [FAQ data minimisation](#)

## Data quality

The data quality principle comprises that data has to be of good quality, i.e. the data has to be accurate and up-to-date.

- > [GDPR in research: data quality](#)
- > [FAQ data quality](#)

## Goal setting

In the goal setting, you describe what personal data you process, with which legitimate purpose and for how long.

- > [GDPR in research: goal setting](#)
- > [FAQ goal setting](#)

## Minimisation of use

Minimise the processing of and access to personal data, for a pre-defined purpose and period of time, and only by authorised persons.

- > [GDPR in research: minimisation of use](#)
- > [FAQ minimisation of use](#)

## Security measures

Make sure that the personal data you collect is well secured. When working with personal data, make use of privacy protection techniques.

- > [GDPR in research: security measures](#)
- > [FAQ security measures](#)

## Transparency

The GDPR requires the controller to be transparent to data subjects about the processing of their personal data.

- > [GDPR in research: transparency](#)
- > [FAQ transparency](#)

## Rights of data subjects

Fundamental of the GDPR are the right of data subjects concerning the processing of their personal data.

- > [GDPR in research: rights of data subjects](#)
- > [FAQ rights of data subjects](#)

## Research Data Management

### GDPR in research

HOME PLANNING RESEARCH COLLECTING DATA PROCESSING DATA ARCHIVING DATA **GDPR IN RESEARCH** SUPPORT & TRAINING

Research Data Management > GDPR in research

#### GDPR in research

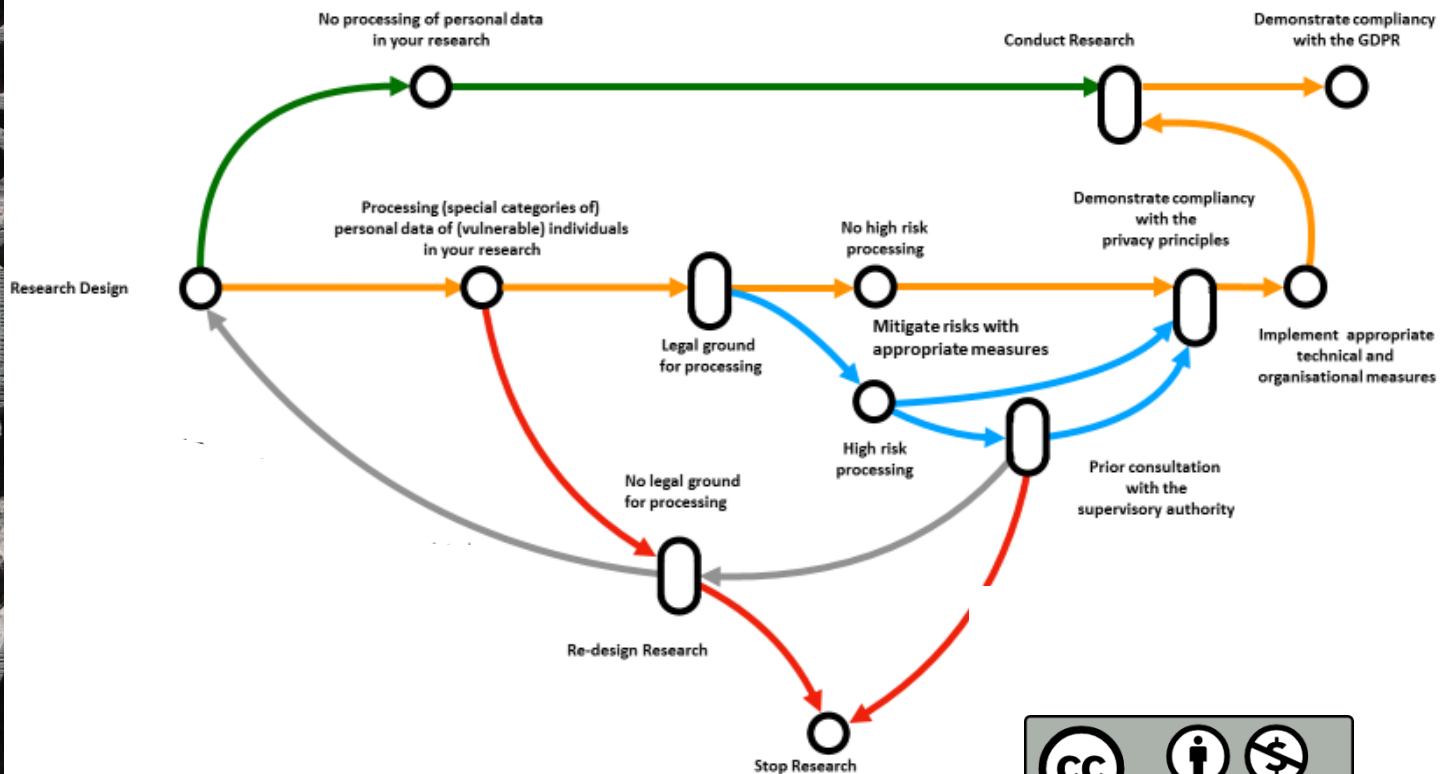
As of May 25 2018, the GDPR (General Data Protection Regulation), or AVG (Algemene Verordening Gegevensbescherming) in Dutch, will apply to the entire European Union. The GDPR has its implications for research. Anyone who collects personal data within Radboud University during their research, must follow 8 guidelines following the Privacy by design principle.

The guidelines are only applicable for research with **personal data**. Personal is any data that can lead to the identification of an individual. For example name, birth date, email-address and IP address are direct personal data. But also a combination of data can lead to the identification of an individual and should therefore be treated as personal data. If you **don't process personal data** in your research, then the GDPR is not applicable. This is for instance the case when your research only includes anonymised data (but be aware that pseudonymised data is personal data).



# [Data and GDPR]

## The Privacy Impact Assessment (PIA) Route Planner for Academic Research Inspired by Harry Beck's London Metro Map



Erasmus University Rotterdam  
marlon.domingus@eur.nl  
February 2018

# The Logic of a Privacy Impact Assessment (PIA) for Academic Research

Q1. Do you process (special categories of) personal data of (vulnerable) individuals in your research?



**YES**

**NO**  
Proceed - no measures required for safeguarding privacy.

## "Personal Data" (GDPR\*, Article 4):

Any information relating to an identified or identifiable natural person: a name, an identification number, location data, an online identifier, one or more factors specific to the physical, physiological, genetic, mental, economic, cultural or social identity of that natural person.

## "Special Categories of Personal Data (Sensitive Data)" (GDPR, Article 9):

Data revealing racial or ethnic origin, political opinions, religious or philosophical beliefs, or trade union membership, the processing of genetic data, biometric data for the purpose of uniquely identifying a natural person, data concerning health or data concerning a natural person's sex life or sexual orientation.

**Action**

## Records of processing activities (GDPR\*, Article 30):

The university shall maintain a digital record of the processing activities in your research to demonstrate compliance to the GDPR.

This register contains:

1. The name and contact details of the researcher, the research partners and service providers;
2. The purposes of the processing;
3. A description of the categories of data subjects and of the categories of personal data;
4. The categories of recipients to whom the personal data have been or will be disclosed.

Q2. What is the legal ground for this processing?

## Lawfulness of Processing (GDPR\*, Article 6, 89):

1. The individuals participating in your research have freely given their explicit consent for one or more specific purposes.
2. Your research contributes to a legitimate interest, yet results in no high risks for the individuals participating in the research.
3. Your research has a scientific, historical or statistical purpose, yet results in no high risks for the individuals participating in the research.

**Action**

## Data protection by design and by default (GDPR\*, Article 25):

Implement appropriate technical and organisational measures:

1. **Individual participating in your research (data subject).** Is the participant well informed, aware of possible risks for her/him and aware of the purpose of the research?
2. **Data.** Is the data de-identified and encrypted?
3. **Access Management.** How is access managed and controlled for the PI / team (expanded) / public?
4. **Software / Platform.** Are the *Terms of Service* for used software / platform checked (where is the data and who has access and has which usage rights)?
5. **Devices.** Are devices used safe? Encrypted drive, encrypted communication, strong password / two factor authentication.
6. **Partners.** Are the research partners / service partners trusted and are appropriate legal agreements made, with regards to roles, rights and responsibilities?
7. **Safe and secure collaboration.** Is the ((cross border) communication to, in and from the) collaboration platform end to end encrypted, are roles and permissions defined and implemented, is logging and monitoring implemented?
8. **Risk definition and mitigation.** Are risks defined and mitigated? Is a risk audit procedure started?

**YES**

**NO**  
Stop research or redefine research.

Q3. Is this processing a high risk processing?

## Criteria for high risk processing (WP29 - DPIA Guideline\*\*):

1. Evaluation or scoring
2. Automated-decision making with legal or similar significant effect
3. Systematic monitoring
4. Sensitive data or data of a highly personal nature
5. Data processed on a large scale
6. Matching or combining datasets
7. Data concerning vulnerable data subjects
8. Innovative use or applying new technological or organisational solutions
9. When the processing itself prevents data subjects from exercising a right or using a service or a contract

**YES**

**NO**

Proceed - measures required for safe-guarding privacy.

**Action**

## Prior consultation (GDPR\*, Article 36):

1. The Data Protection Officer shall, on behalf of the researcher, consult the supervisory authority, prior to the processing (the research) when the processing would result in a high risk *in the absence of measures* to mitigate the risk.

**Action**

## Principles relating to processing of personal data (GDPR\*, Article 5):

Demonstrate compliance with the principles: lawfulness, fairness, transparency, purpose limitation, data minimisation, accuracy, storage limitation, integrity, confidentiality and accountability.

\* Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation). Online available at: <http://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32016R0679&from=EN>

\*\* Article 29 Data Protection Working Party: *Guidelines on Data Protection Impact Assessment (DPIA) and determining whether processing is "likely to result in a high risk" for the purposes of Regulation 2016/679.* Adopted on 4 April 2017. As last Revised and Adopted on 4 October 2017. Online available at: [https://ec.europa.eu/newsroom/document.cfm?doc\\_id=47711](https://ec.europa.eu/newsroom/document.cfm?doc_id=47711)

# GDPR- consenso



## DARIAH consent wizard



## Welcome to the DARIAH ELDAH Consent Form Wizard (CFW)!

Since the coming into effect of the General Data Protection Regulation (GDPR), researchers must consider their subjects' right to privacy when conducting their research while considering their subjects' right to privacy.

### (What this tool is)

The aim of the CFW is to support humanities researchers with their specific professional activity.

This tool will guide you through a questionnaire that will cover your specific purpose and the data categories you intend to collect.

Please be aware that the validity of the generated output will depend on the comprehensive answers you provide. The more comprehensive your answers, the more comprehensive the output will be.

After answering the questionnaire, the CFW will output a consent form template. You will be able to use this text template for creating your own consent forms.

Since we will not store the generated output ourselves, please provide your result as an example for other CFW users, please provide your result as an example for other CFW users, please provide your result as an example for other CFW users.

### (... and what it is not)

The consent forms provided by this tool will observe the Art. 17 GDPR advice.

**BE AWARE THAT THIS TOOL DOES NOT PROVIDE FORMAL LEGAL ADVICE. IT IS AT YOUR OWN RISK. TO MAKE SURE THAT YOU ARE COMPLETELY COMPLIANT WITH APPLICABLE LEGISLATION, CONSULT A LAWYER IN YOUR COUNTRY.**

The CFW provides consent form templates for several academic scenarios in which you may need to collect data about people (i.e. "process personal data"). The use cases presented here were identified by the working group **ELDAH** ("Ethics and Legality in Digital Arts and Humanities") through surveys of needs and demands of the **DARIAH-EU** research community. If you find your use scenario to be missing, do not hesitate to contact us: [eldah@darlah.eu](mailto:eldah@darlah.eu)

### What are you planning to do?

- Gather data from and/or about living people for research purposes
- Communicate through mailinglists or other (digital) communication media
- Gather data and/or consent from participants as the host of an academic event

Continue

### In what form are you gathering/recording data from/about your participants?

- Written survey (pen and paper)
- Online survey
- Oral interview (sound recording)
- Oral or video interview (transcription)
- Video interview

### What types of data do you collect from the participants?

Please be aware that the GDPR requires you to minimize the personal data collected to only what is necessary for your research. Please do not collect data you don't need just because you feel a need for completion.

#### Generic data categories

- Name, surname
- IP address
- E-mail address
- Age / date of birth
- Address / place of residence
- Gender
- Marital status
- Educational background / title
- Affiliation / professional situation / occupation

04/10/2018

# [anonimizzare]

## Anonymisation

[UK Data service](#)

Anonymisation is a valuable tool that allows data to be shared, whilst preserving privacy. The process of anonymising data requires that identifiers are changed in some way such as being removed, substituted, distorted, generalised or aggregated.

A person's identity can be disclosed from:

- **Direct identifiers** such as names, postcode information or pictures
- **Indirect identifiers** which, when linked with other available information, could identify someone, for example information on workplace, occupation, salary or age

You decide which information to keep for data to be useful and which to change. Remove key variables, applying pseudonyms, generalising and removing contextual information from textual files, and blurring image or video data could result in important details being missed or incorrect inferences being made. See [example 1](#) and [example 2](#) for balancing anonymisation with keeping data useful for qualitative and quantitative data.

Anonymising research data is best planned early in the research to help reduce anonymisation costs, and should be considered alongside obtaining informed consent, data sharing or imposing access restrictions. Personal data should never be disclosed as research information, unless a participant has given consent to do so, ideally in writing.

Quantitative data

Qualitative data

Step-by-step

Anonymising **quantitative data** may involve removing or aggregating variables or reducing the precision or detailed textual meaning of a variable.

### Primary anonymisation techniques

- **Remove direct identifiers** from a dataset. Such identifiers are often not necessary for secondary research.

*Example:* Remove respondents' names or replace with a code; remove addresses, postcode information, institution and telephone numbers.

- **Aggregate or reduce the precision** of a variable such as age or place of residence. As a general rule, report the lowest level of geo-referencing that will not potentially breach respondent confidentiality. The exact scale of data collected, but very detailed geo-references like full postcodes of small towns or villages are likely to be problematic. Coded data which may be potentially revealing can be aggregated into broader categories. If aggregation of a disclosive variable is not possible, consider removing from the dataset.

*Example:* Record the year of birth rather than the day; record postcode sectors (first 3 or 4 digits) rather than aggregate detailed 'unit group' standard occupational employment codes up to 'minor group' codes by removing the last digit.

- **Generalise the meaning** of a detailed text variable by replacing disclosive free-text responses with more general text.

*Example:* Detailed areas of medical expertise could be replaced by 'general practitioner'. The expertise variable could be replaced by more generic coded responses such as 'one area of medical speciality', etc.

- **Anonymise relational data** where relations between variables in related or linked datasets or in combination with other publicly available outputs may disclose identities.

*Example:* In confidential interviews on farms the names of farmers have been replaced with codes and other confidential information on the nature of the farm businesses and their locations have been disguised to anonymise the data.

However, if related biodiversity data collected on the same farms, using the same farmer codes, contain detailed locations for biodiversity data alone the location would not be confidential. Farmers could be identified by combining the two datasets.

The link between farmer codes and biodiversity location data should be removed, for example by using separate codes for farmer interviews and for farm locations.

- **Anonymise geo-referenced data** by replacing point coordinates with non-disclosing features or variables; or, preferably, keep geo-references intact and impose access restrictions on the data instead.

Point data may fix the position of individuals, organisations or businesses studied, which could disclose their identity. Point coordinates may be replaced by larger, non-disclosing geographical areas such as polygon features (km<sup>2</sup> grid, postcode district, county), or linear features (random line, road, river). Point data can also be replaced by meaningful alternative variables that typify the geographical position and represent the reason why the locality was selected for the research, such as poverty index, population density, altitude, vegetation type. In this way, the value of data is maintained.

# [anonimizzare]



Amnesia OpenAIRE

## High accuracy Data Anonymization.

Perform research and share your results that satisfy GDPR guidelines by using data anonymization algorithms.

GET STARTED



### Unlock sensitive data analysis

Use Amnesia to transform personal data to anonymous data that can be used for statistical analysis. Data anonymized with Amnesia are "statistically guaranteed" that they cannot be linked to the original data.

- ✓ Guarantees no links to the original data
- ✓ Offers k-anonymity & km-anonymity
- ✓ Allows minimal reduction of information quality



### Become GDPR compliant

Create anonymous datasets from personal data that are treated as statistics by GDPR. Anonymous data can be used without the need for consent or other GDPR restrictions, greatly reducing the effort needed to extract value from them.

- ✓ Guarantees anonymity
- ✓ Goes beyond pseudo-anonymization
- ✓ Anonymized data are not constrained by GDPR



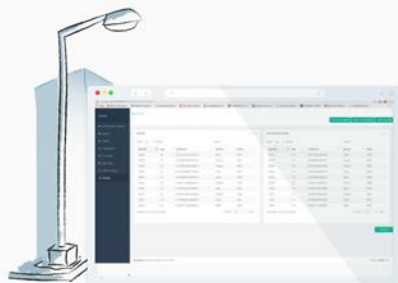
### High Usability & Flexibility

Anonymization tailored to user needs through a graphical interface. Guide the algorithm and decide trade-offs with simple visual choices. Developers can incorporate Amnesia anonymization engine to their project through a ReST API.

- ✓ Easy usage interface
- ✓ Adjustable settings
- ✓ Visualization of anonymization choices

## How it works

Get anonymous data in 3 steps



### 1 Insert your data

Amnesia accepts complex object relational data in delimited text files.

### 2 Select and Preview the data to anonymize

Visual representations of anonymization parameters and results allow non-expert users to tailor the anonymization process to their needs.

### 3 Download your data anonymized

The process is completed without any sensitive data leaving your premises!

# ...i dati vanno citati



DCC

DCC guides

Because good research needs good data



[DataCite How to](#)

[About us](#)

[Services](#)

[Resources](#)

DataCite Data help further research and assures reliable, predictable, and unambiguous access to research data in the following order to:

- support proper attribution and credit
- support collaboration and reuse of data
- enable reproducibility of findings
- foster faster and more efficient research progress, and
- provide the means to share data with future researchers

DataCite also looks to community practices that provide data citation guidance. The Joint Declaration Citation Principles is a set of guiding principles for data within scholarly literature, another dataset, or a research object (Data Citation Synthesis Group 2014). The FAIR Guiding Principles provide a guideline for those that want to enhance reuse of their data (Wilkinson 2016).

## Data Citation Examples

We recognise that the challenges associated with data publication vary across disciplines, and we encourage research communities to develop citation systems that work well for them. Our recommended format for data citation is as follows:

Creator (PublicationYear). Title. Publisher. Identifier

It may also be desirable to include information about two optional properties, Version and ResourceType (as appropriate). If so, the recommended form is as follows:

Creator (PublicationYear). Title. Version. Publisher. ResourceType. Identifier

- Principles of data citation
- Data citation for authors
  - Ways of referencing data
  - Elements of a data citation
  - Digital Object Identifiers
  - Contributor identifiers
  - Granularity
  - Citing unreleased data
  - Citing physical data



...pausa?

