DATA MANAGEMENT, «FAIR», EOSC …QUESTI SCONOSCIUTI

Elena Giglia

Ferrara, 1 marzo 2019

elena.giglia@unito.it

The road ahead today

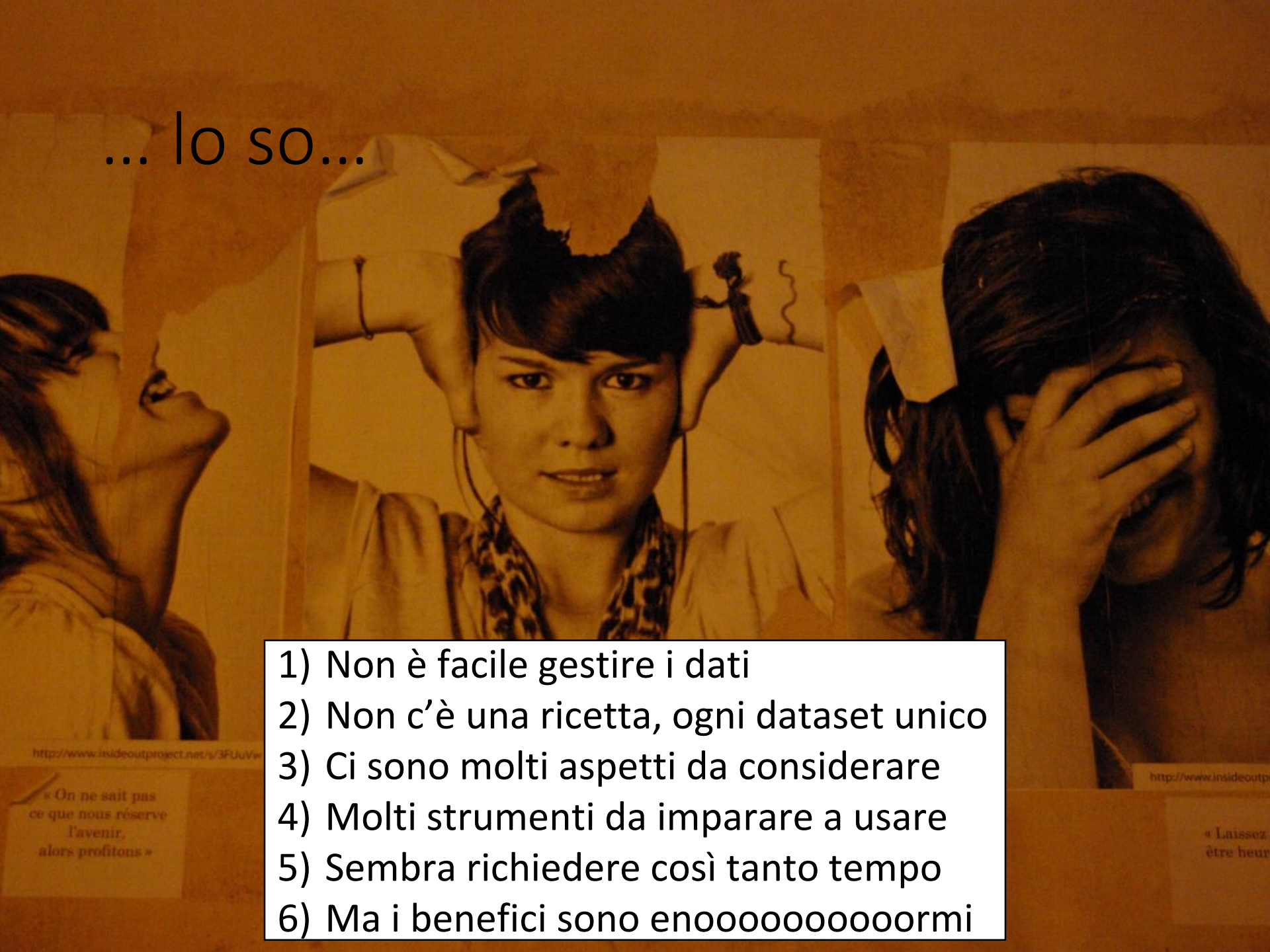
…perché occuparci dei dati?

…managed, FAIR, Open: diversi

… DMP: uno strumento utile, non un peso

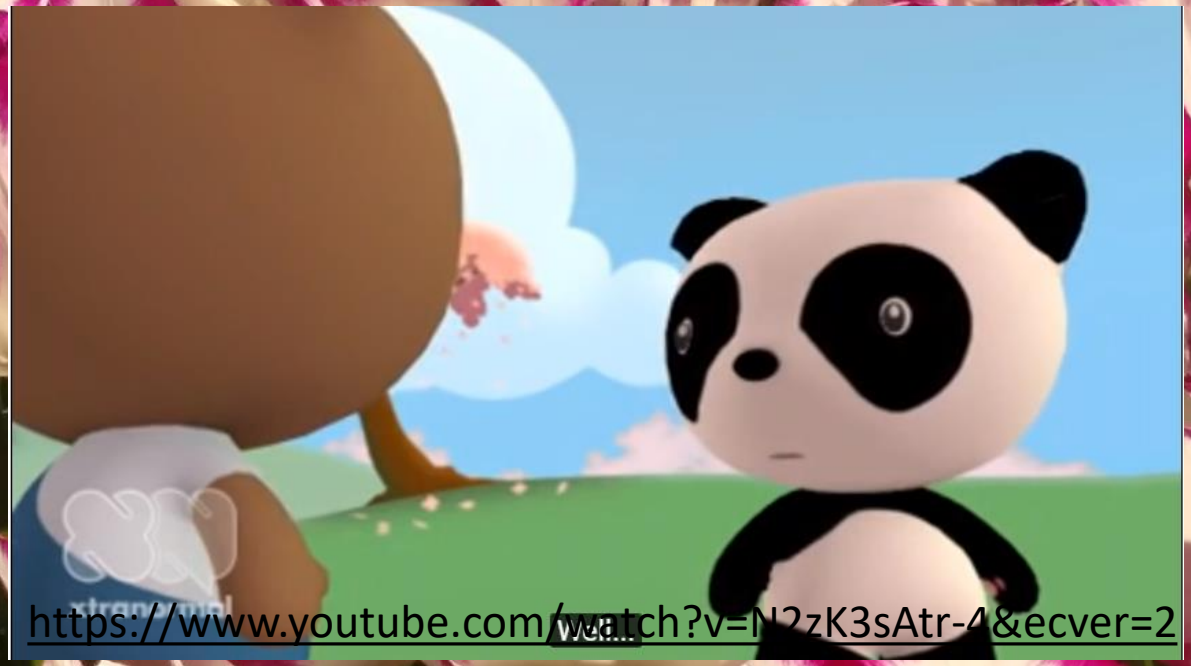
…corsi, servizi, strumenti per gestire i dati

… lo so…



1) Non è facile gestire i dati
2) Non c'è una ricetta, ogni dataset unico
3) Ci sono molti aspetti da considerare
4) Molti strumenti da imparare a usare
5) Sembra richiedere così tanto tempo
6) Ma i benefici sono enooooooooormi

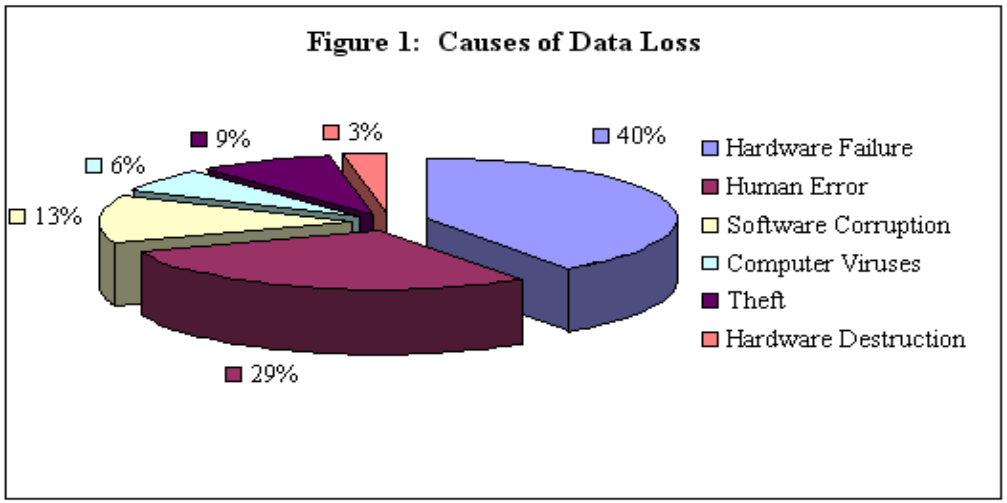
I dati. Perché vanno curati?



… è l'incubo del data steward:

- nessun backup
- nessun software di accompagnamento
- nessuna legenda dati

Perché i dati sono fragili / 1

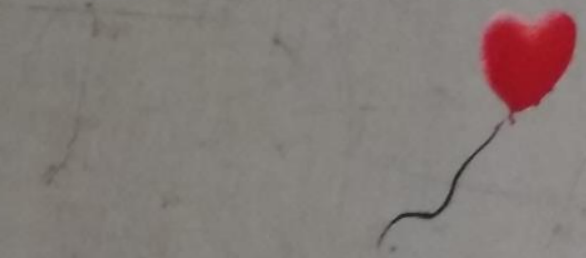


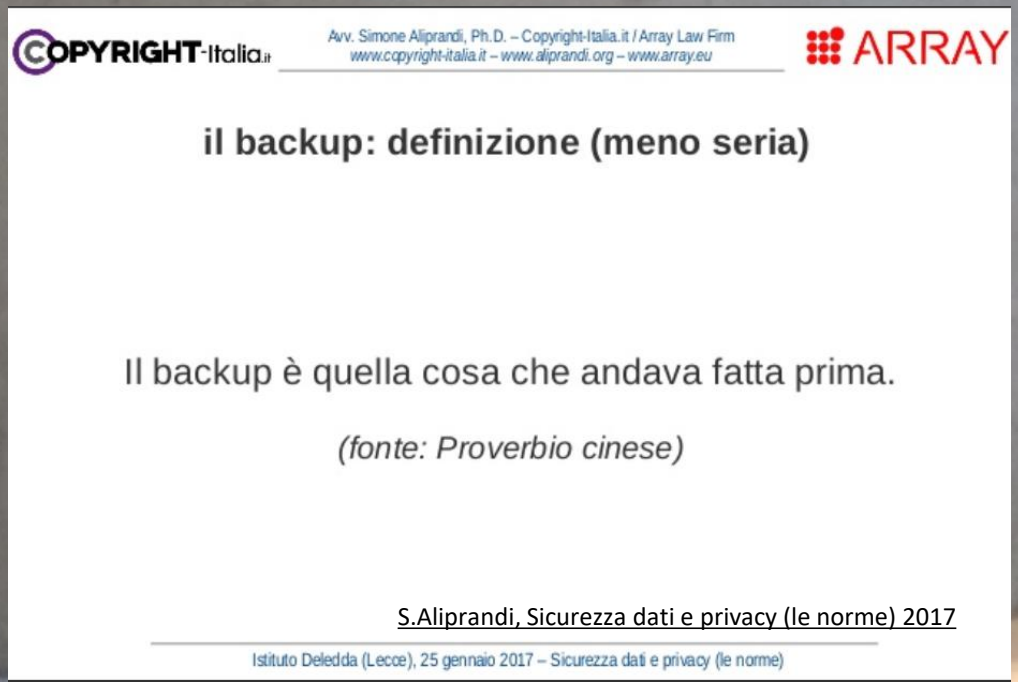
Source: Author's estimates based on data from Safeware, The Insurance Agency, Inc., "2000 Safeware Loss Study," 2001; and ONTRACK Data International, Inc., "Understanding Data Loss," 2003.

D.M. Smith The cost of lost data, 2003

rché i dati sono fragili / 2

PMRblog, 2011





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

… i dati sono fragili

Scientists losing data at a rapid rate

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

Elizabeth Gibney & Richard Van Noorden

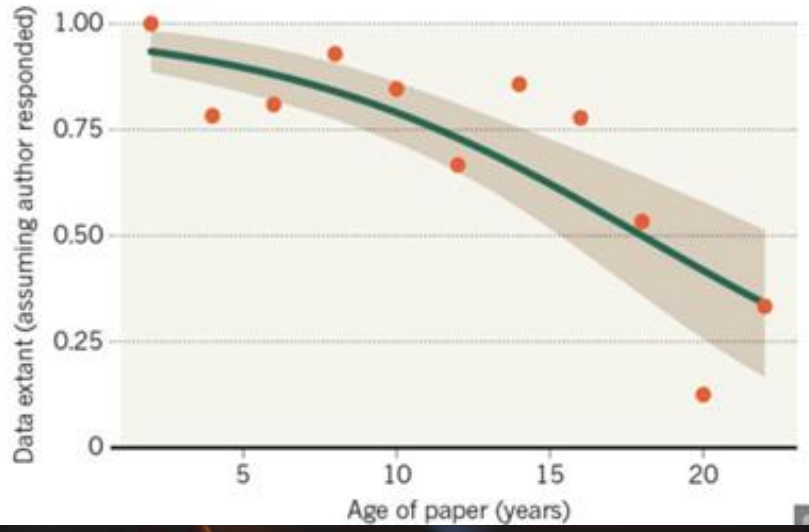
19 December 2013

Rights & Permissions

80% saranno persi in 20 anni

MISSING DATA

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



http://www.nature.com/news/scientists-losing-data-at-a-rapid-rate-1.14416

…ECCO A COSA SERVE IL DATA MANAGEMENT PLAN.

NON È SOLO L'ENNESIMA NOIA BUROCRATICA

…DOVE conservate i dati?

Perché occuparsi dei dati?

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

EVITARE DI PERDERLI

ORGANIZZARLI PER RENDERE PIÙ EFFICACE LA RICERCA

(SE APERTI) ESSERE PIÙ VISIBILI

ALCUNI SONO UNICI E IRRIPETIBILI (meteorologia)

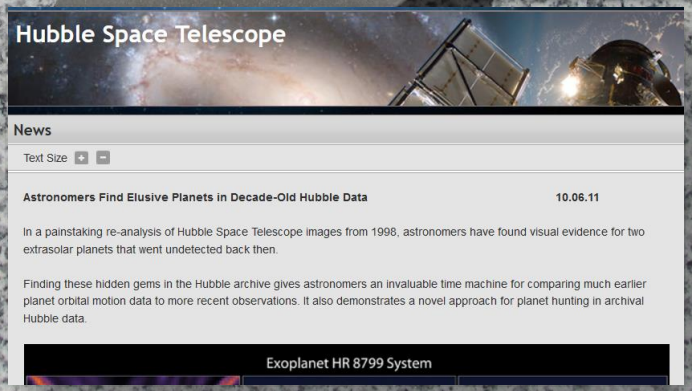
PERMETTERE VALIDAZIONE E CONTROLLI

(SE APERTI) FAVORIRE COLLABORAZIONI

MIGLIORARE INTEGRITÀ DELLA RICERCA

(SE APERTI) FAVORIRE RIUSO INEDITO

ESSERE RIPRODUCIBILI



«the coolest thing to do with your data will be thought of by someone else» [R.Pollock]

Perché occuparsi 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, http://goo.gl/ExaGW

I dati creano ponti fra le discipline

Perché occuparsi dei dati?

Box 1. Some Research Practices that May Help Increase the Proportion of True Research Findings

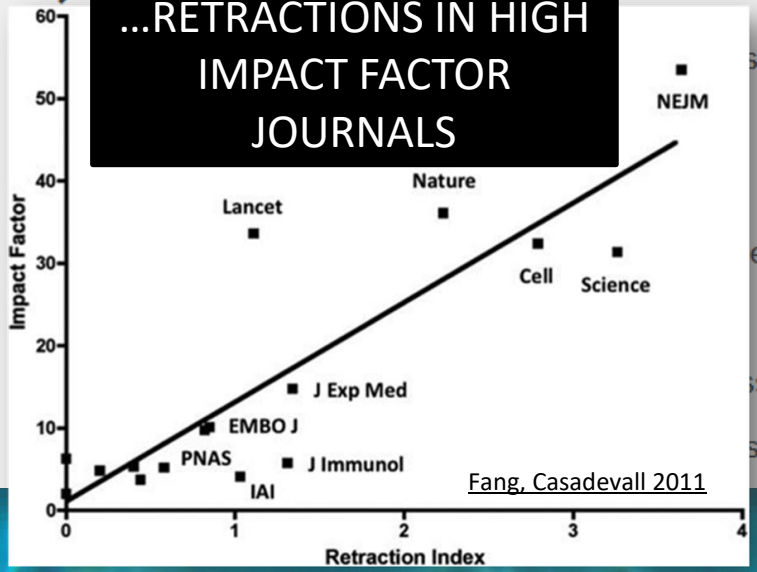
- Large-scale collaborative research
- Adoption of replication culture
- Registration (of studies, protocols, analysis codes, datasets, raw data, and results)
- Sharing (of data, protocols, materials, software, and

…es or "successes"

…ssemination of research

…s and statistical literacy

…RETRACTIONS IN HIGH IMPACT FACTOR JOURNALS



Fang, Casadevall 2011

PLOS | MEDICINE 2005

OPEN ACCESS

ESSAY

Why Most Published Research Findings Are False

John P. A. Ioannidis

Published: August 30, 2005 • https://doi.org/10.1371/journal.pmed.0020124

PLOS |

OPEN ACCESS

ESSAY

How to Make More Published Research True

John P. A. Ioannidis

Published: October 21, 2014 • https://doi.org/10.1371/journal.pmed.1001747

Tony Ross-Hellauer ha ritwittato

Gr@h@m Steel @McDawg · 8 set

If you missed Prof John **Ioannidis**'s Plenary at #osfair2017

Traduci dalla lingua originale: inglese

OSFair2017

OSFair2017

youtube.com

2017

OSFPREPRINTS https://dx.doi.org/10.17605/OSF.IO/2DXU5

The Preregistration Revolution

Brian Nosek, Charles Ebersole, Alexander DeHaven, David Mellor

Created on: June 16, 2017 | Last edited: March 12, 2018

nature human behaviour

Altmetric: 2520 Citations: 63 More detail »

Perspective | OPEN

A manifesto for reproducible science

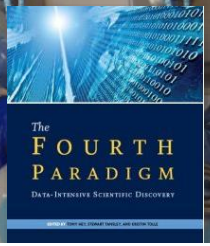
Marcus R. Munafò, Brian A. Nosek, Dorothy V. M. Bishop, Katherine S. Button, Christopher D. Chambers, Nathalie Percie du Sert, Uri Simonsohn, Eric-Jan Wagenmakers, Jennifer J. Ware & John P. A. Ioannidis

Manifesto, Jan 2017

http://journals.plos.org/plosmedicine/article?id=10.1371/journal.pmed.1001747

Perché occuparci dei dati?

2009 Jim Gray

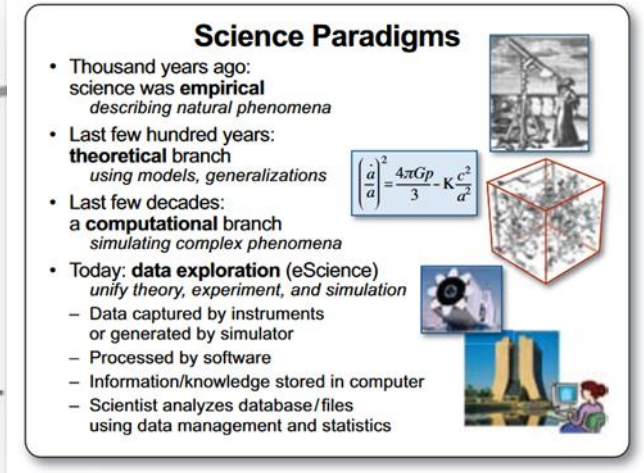


2018 JISC

Executive summary

Introduction

Researchers are creating, gathering and using data in hitherto-unimagined volumes. These vast data resources dramatically increase the capacity of science to infer patterns in phenomena, whether physical, chemical, biological or human, or in the complex systems that are at the heart of most global challenges



LA SCIENZA È DATA INTENSIVE. PUNTO

The challenge is clear to us: if we do not act, there might be a looming crisis on the horizon. The vast majority of all data in the world (in fact up to 90%) has been generated in the last two years. Computers have long surpassed individuals in their ability to perform pattern recognition over large data sets. Scientific data is in dire need of openness, better handling, careful management, machine actionability and sheer re-use. One of the sobering conclusions of our consultations was that research infrastructure and communication appear to be stuck in the 20th century paradigm of data scarcity. We should see this step-change in science as an enormous opportunity and not as a threat. The EOSC is a positive 'Cloud on the Horizon' to be realised by 2020. Ultimately, actionable knowledge and translation of its benefits to society will be handled by humans in the 'machine era' for decades to come, machines are just made to serve us.

2016

Perché occuparci dei dati?

The Vienna Declaration on the European Open Science Cloud

Vienna, 23 November 2018

PERCHÉ ORA ABBIAMO EOSC

Vienna, Nov.23, 2018

We, Ministers, delegates and other participants attending the launch event of the European Open Science Cloud (EOSC):

1. Recall the challenges of data driven research in pursuing excellent science as stated in the "EOSC Declaration" signed in Brussels on 10 July 2017.

2. Reaffirm the potential of the European Open Science Cloud to transform the research landscape in Europe. Confirm that the vision of the European Open Science Cloud is that of a research data commons, inclusive of all disciplines and Member States, sustainable in the long-term.

3. Recognise that the implementation of the European Open Science Cloud is a process, not a project, by its nature iterative and based on constant learning and mutual alignment. Highlight the need for continuous dialogue to build trust and consensus among scientists, researchers, funders, users and service providers.

4. Highlight that Europe is well placed to take a global leadership position in the development and application of cloud services for Science. Rea and open to the world, reaching out over time to

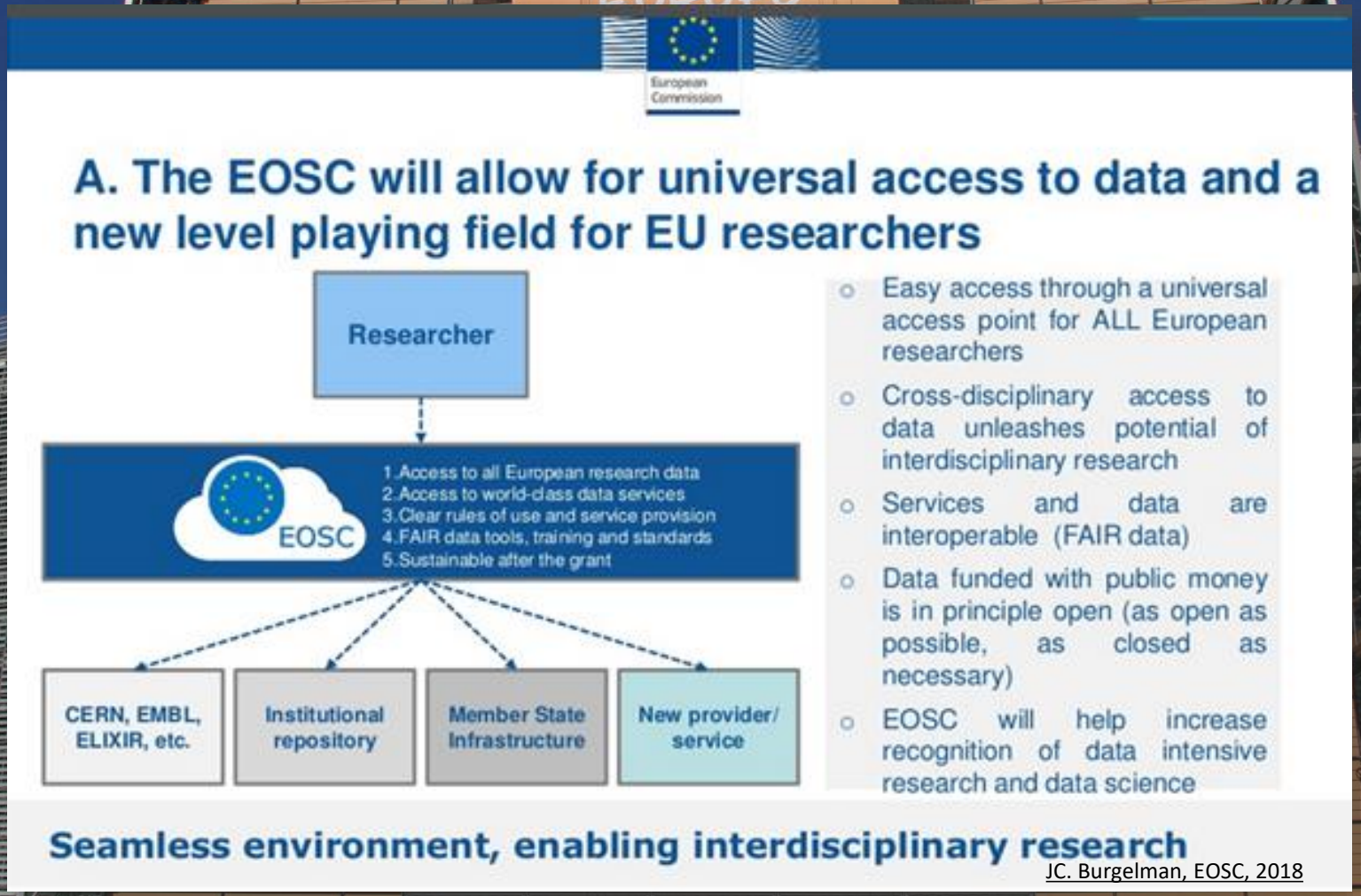
SEAMLESS ACCESS TO OPEN BY DEFAULT
FAIR DATA

5. Recall that the Council roadmap and the federated

9. Call for the European Open Science Cloud to provide all researchers in Europe with seamless access to an open-by-default, efficient and cross-disciplinary environment for storing, accessing, reusing and processing research data supported by FAIR data principles.

Science Cloud a reality, hinting at the need to further strengthen the ongoing dialogue across institutions and with stake-holders, for a new governance framework to be launched in Vienna, on 23 November 2018.

[EOSC: cioè?]



JC. Burgelman, EOSC, 2018

[EOSC – cioè?]

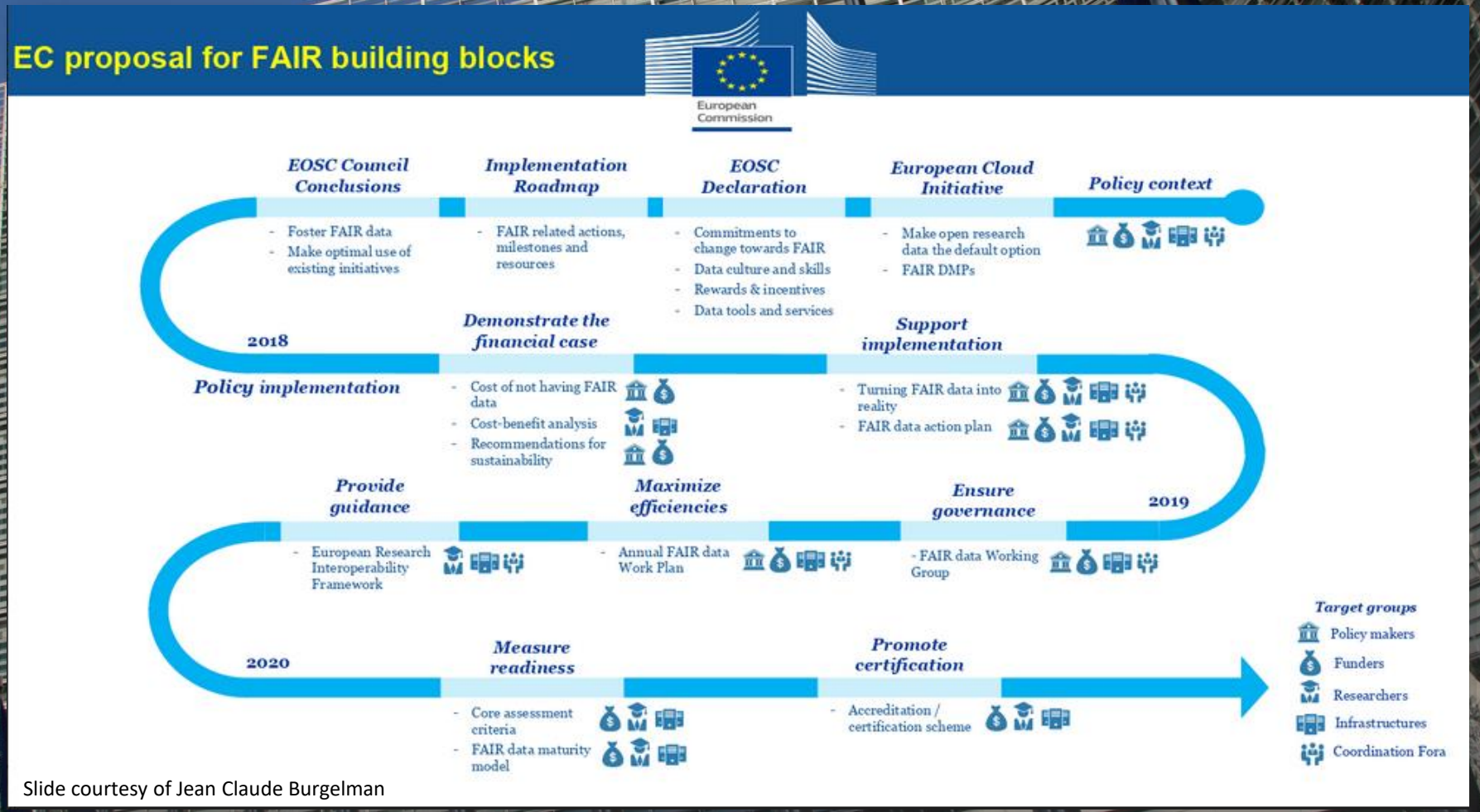
THE EUROPEAN OPEN SCIENCE CLOUD? SOME NUANCES AND DEFINITIONS

Imagine a federated, globally accessible environment where researchers, innovators, companies and citizens can publish, find and re-use each other's data and tools for research, innovation and educational purposes. Imagine that this all operates under well-defined and trusted conditions, supported by a sustainable and just value for money model. This is the environment that must be fostered in Europe and beyond to ensure that European research and innovation contributes in full to knowledge creation, meet global challenges and fuel economic prosperity in Europe. This we believe encapsulates the concept of the European Open Science Cloud (EOSC), and indeed such a federated European endeavour might be expressed as the European contribution to an Internet of FAIR Data and services.

The European Open Science Cloud is a supporting environment for Open Science and not an 'open Cloud' for science.

The EOSC aims to accelerate the transition to more effective Open Science and Open Innovation in a Digital Single Market by removing the technical, legislative and human barriers to the re-use of research data and tools, and by supporting access to services, systems and the flow of data across disciplinary, social and geographical borders. The term European Open Science Cloud requires some reflection to dispel incorrect associations and clarify boundaries; in fact the term 'cloud' is a metaphor to help convey the idea of seamlessness and a commons.

[EOSC – European Open Science Cloud]



Slide courtesy of Jean Claude Burgelman

[EOSC Declaration]

Data culture and FAIR data

- [**Data culture**] European science must be grounded in a common culture of data stewardship, so that research data is recognised as a significant output of research and is appropriately curated throughout and after the period conducting the research. Only a considerable cultural change will enable long-term reuse for science and for innovation of data created by research activities: no disciplines, institutions or countries must be left behind.

- [**Open access by-default**] All researchers in Europe must enjoy access to an open-by-default, efficient and cross-disciplinary research data environment supported by FAIR data principles. Open access must be the default setting for all results of publicly funded research in Europe, allowing for proportionate limitations only in duly justified cases of personal data protection, confidentiality, IPR concerns, national security or similar (e.g. 'as open as possible and as closed as necessary').

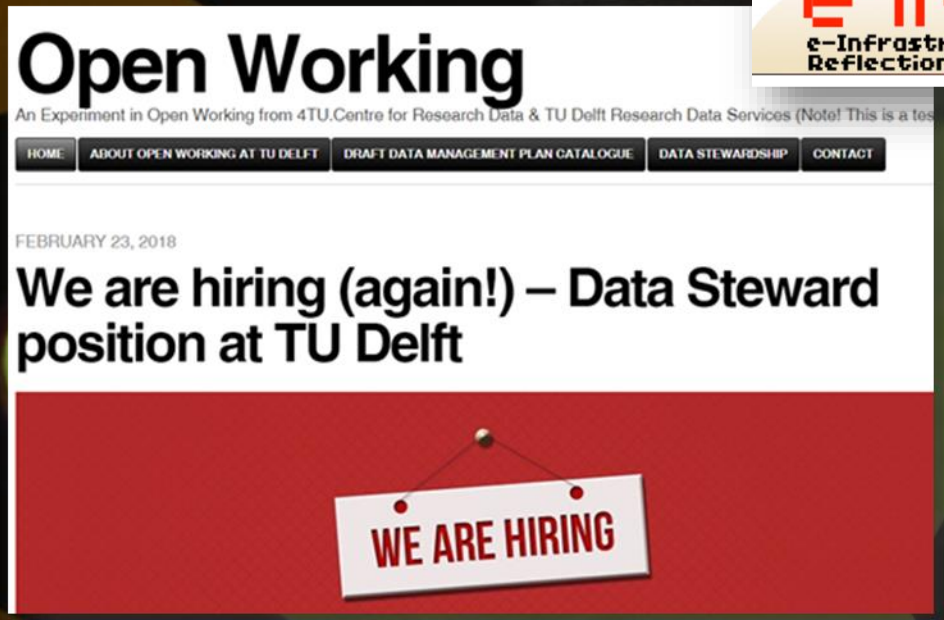
- [**Skills**] The necessary skills and education in research data management, data stewardship and data science should be provided throughout the EU as part of higher education, the training system and on-the-job best practice in the industry. University associations, research organisations, research libraries and other educational brokers play an important role but they need substantial support from the European Commission and the Member States.

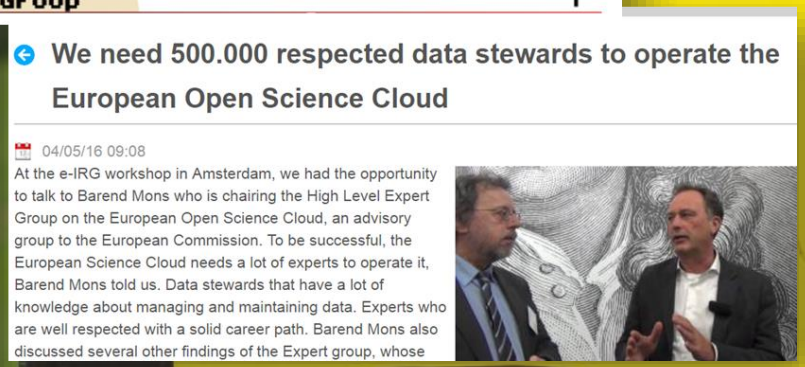
Oct. 2017

[EOSC è anche data stewardship]

Report, 2016

The number of people with these skills needed to effectively operate the EOSC is, we estimate, likely exceeding half a million within a decade. As we further argue below, we believe that the implementation of the EOSC needs to include instruments to help train, retain and recognise this expertise, in order to support the 1.7 million scientists and over 70 million people working in innovation[9]. The success of the EOSC depends upon it.





Perché occuparci di dati?

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

2013

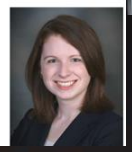
Pubblicato da keynesblog il 18 aprile 2013 in consigliati, Economia, ibt, Teoria economica

1. **l'esclusione selettiva di alcune osservazioni nei dati;**
2. **uno schema di bilanciamento dei dati non convenzionale;**
3. **un errore di codice nel foglio di calcolo originale utilizzato per selezionare i dati.**

DATA AB INITIO

AGING RESEARCH DATA RIGHT, FROM THE START

K.Birney, 2015



and think, "surely I've covered this one my blog?"
up when I wrote December's Exit Strategy post.

and, as you don't want to be stuck with
t. Proprietary, out-of-date, little used formats

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 that when properly calculated, the average real GDP growth rate for countries carrying a public-debt-to-GDP ratio of over 90 percent is actually 2.2 percent, not −0.1 percent as published in Reinhart and Rogoff. That is, contrary to RR, average GDP growth at public debt/GDP ratios over 90 percent is not dramatically different than when debt/GDP ratios are lower.

We also show how the relationship between public debt and GDP growth varies significantly by time period and country. Overall, the evidence we review contradicts Reinhart and Rogoff's claim to have identified an important stylized fact, that public debt loads greater than 90 percent of GDP consistently reduce GDP growth.

Perché occuparci dei dati?

1995

WaveLab and Reproducible Research

Jonathan B. Buckheit and David L. Donoho

Stanford University, Stanford CA 94305, USA

An article about computational science in a scientific publication is **not** *the scholarship itself, it is merely* advertising *of the scholarship. The actual scholarship is the complete software development environment and the complete set of instructions which generated the figures.*

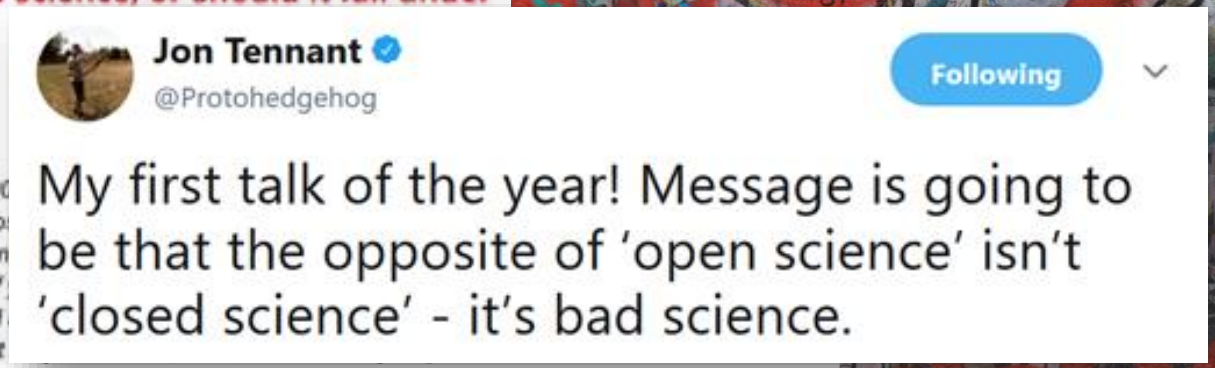
UN ARTICOLO SENZA I DATI È SOLO LA PUBBLICITÀ DELLA RICERCA

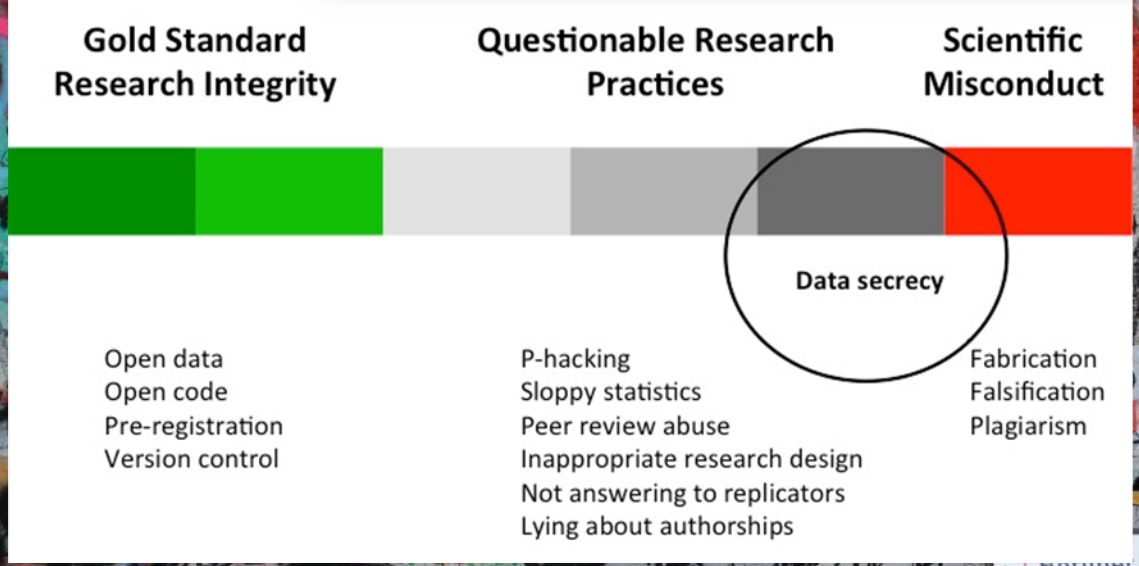
2015

https://memegenerator.net/instance/64979477/case-closed-judge-judy-data-or-it-didnt-happen

No data?

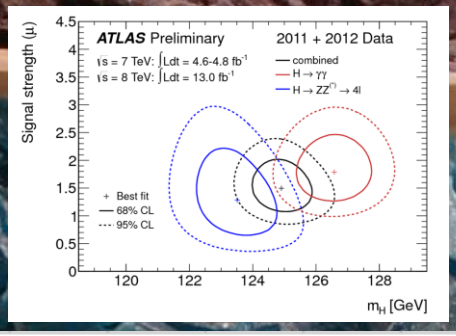
https://twitter.com/alastairdunning/status/968453078218395648





Parliamo di dati

«pezzi»
di conoscenza osservabili



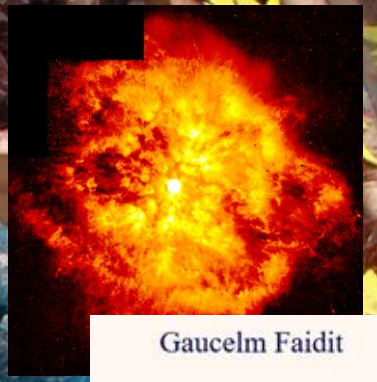


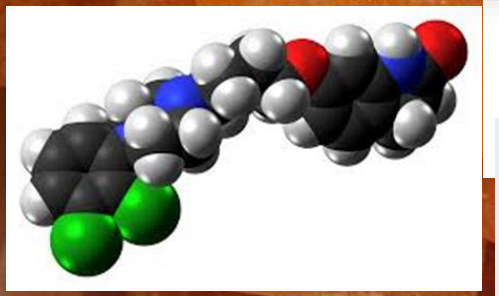
Table S1. Number of reads per prokaryotic operational taxonomic unit (OTU) and sample from the cohort.

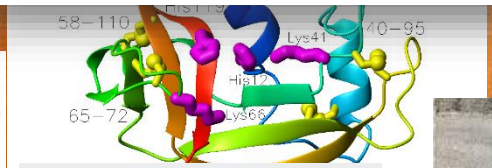
OTU	A01_TP1	A01_TP2	A01_TP3	A03_TP1	A03_TP3	A04_TP1	A04_TP2	A04_TP3	A05_T
OTU_1	261	76	1206	523	2131	25707	64473	60665	
OTU_2	49	52	117	43035	206	119	1152	539	
OTU_9	148	162176					22858	1898	
OTU_6	21	17					1457	29	
							19	85	
							646	214	
							292	37	
							18	170	
							6	4	

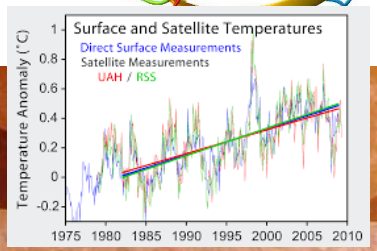
Gaucelm Faidit

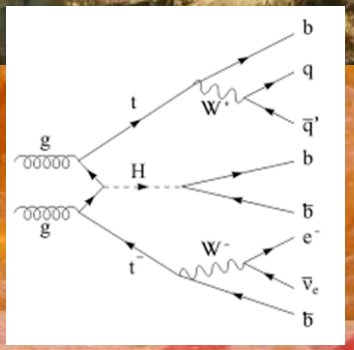
I.

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 planc e languis
e sospir nueg e dia.

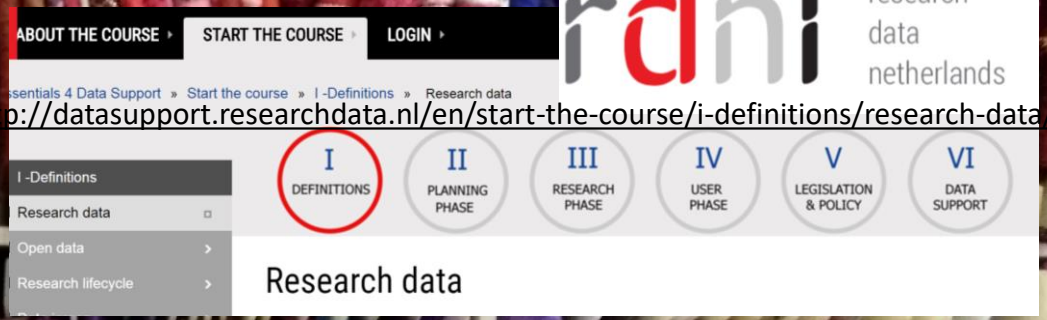








Un po' di glossario

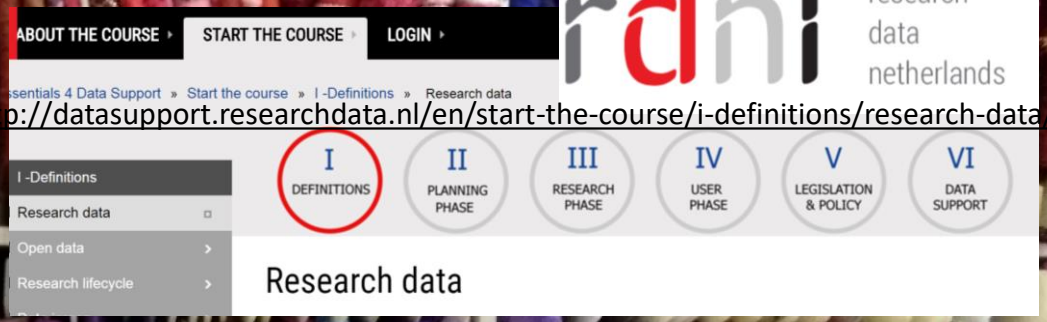


http://datasupport.researchdata.nl/en/start-the-course/i-definitions/research-data/

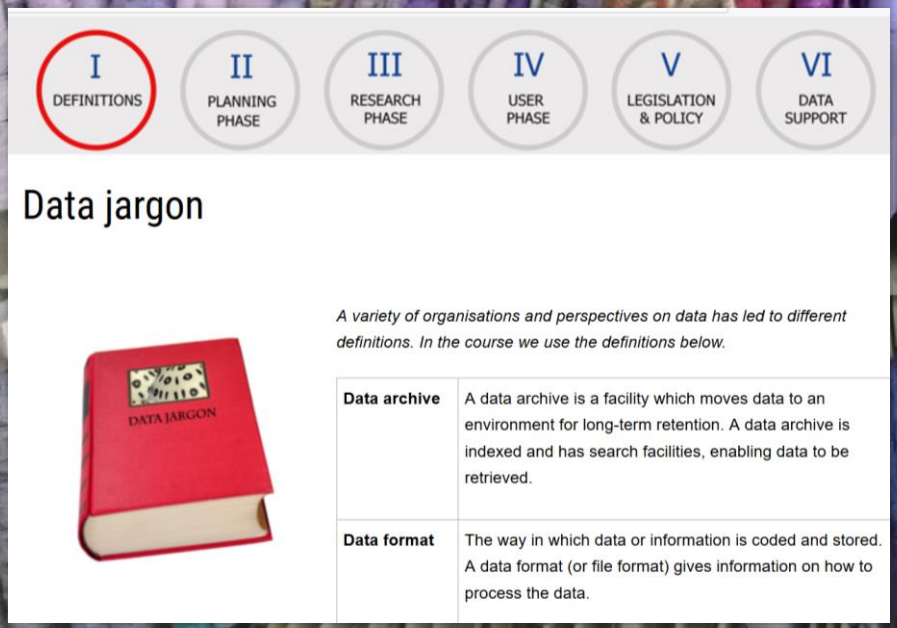
1 2 3 4

Research data is the material underpinning a research assertion. (4)

Un po' di glossario



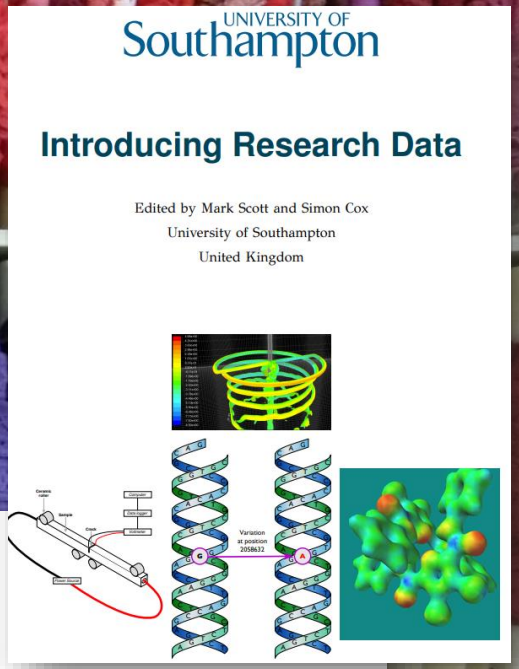
http://datasupport.researchdata.nl/en/start-the-course/i-definitions/research-data/



RDF	RDF is a standard model for data interchange on the Web (see http://www.w3.org/RDF/).
Research data	Data are facts, observations or experiences on which an argument or theory is based. (see http://ands.org.au/guides/what-is-research-data.pdf).
Resolver	A system that brings about the link between a persistent identifier and the location where the object is currently situated.
Text- and data mining	The computer-based process of deriving or organising information from text or data. It works by copying large quantities of material, extracting the data, and recombining it to identify patterns, trends and hypotheses or by providing the means to organise the information mined. (see www.ipo.gov.uk/ipreview-doc-t.pdf).

Un po’ di glossario

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.**

https://eprints.soton.ac.uk/403440/1/introducing_research_data.pdf

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 disemmination 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`.

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.

Un po’ di glossario

DATA CURATION: CONSERVAZIONE SUL LUNGO PERIODO

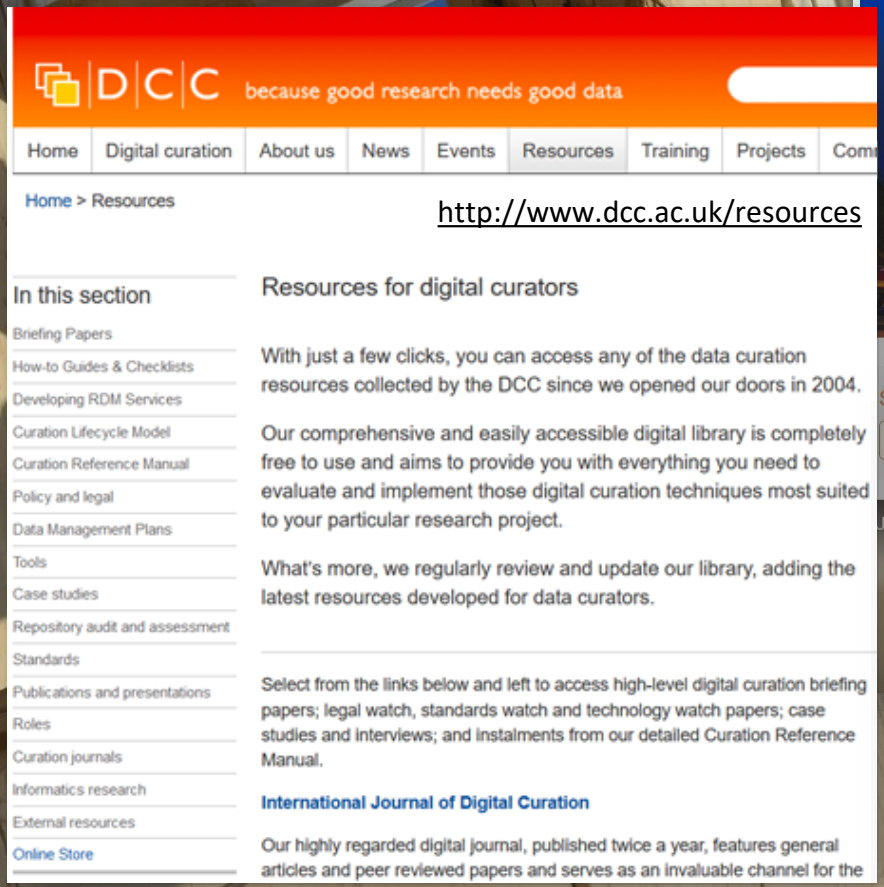
DATA MANAGEMENT: GESTIONE LUNGO TUTTO IL CICLO DI VITA

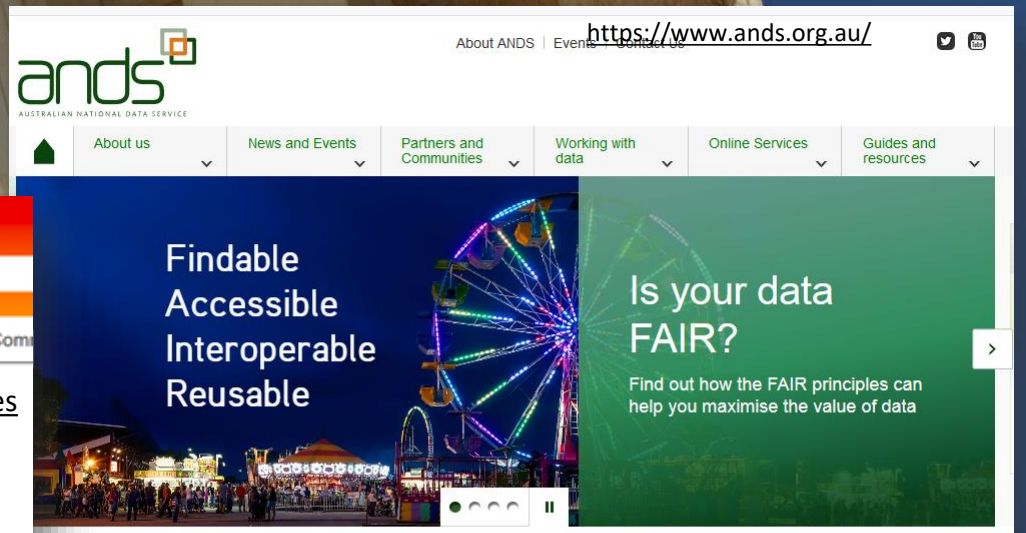
DATA STEWARD: ESPERTO DELLA MATERIA E DELLA GESTIONE DEI DATI

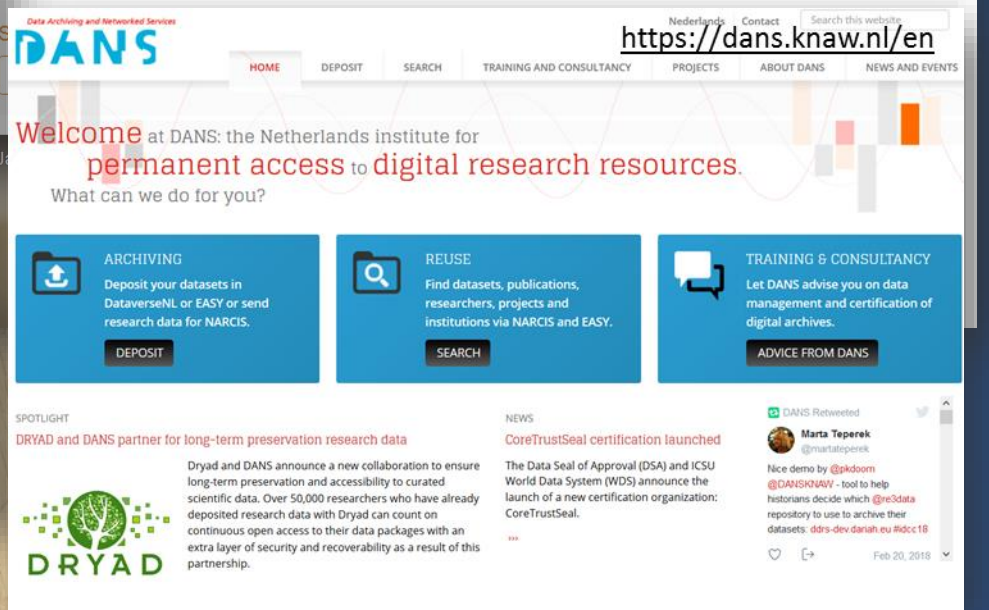
Digital curation involves maintaining, preserving and adding value to digital research data throughout its lifecycle.

http://www.dcc.ac.uk/resources/curation-lifecycle-model

Due pilastri, anzi tre



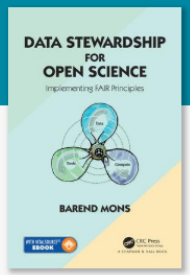




… e un maestro

Taylor & Francis Group
an informa business

https://www.taylorfrancis.com/books/9781498753180



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

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(+)).

Monsense and more... @barendmons · 2 h

Finally! Tomorrow the book goes to the printer: Data Stewardship for Open Science: Implementing FAIR Principles

Traduci dalla lingua originale: inglese

Data Stewardship for Open Science: Implementing …

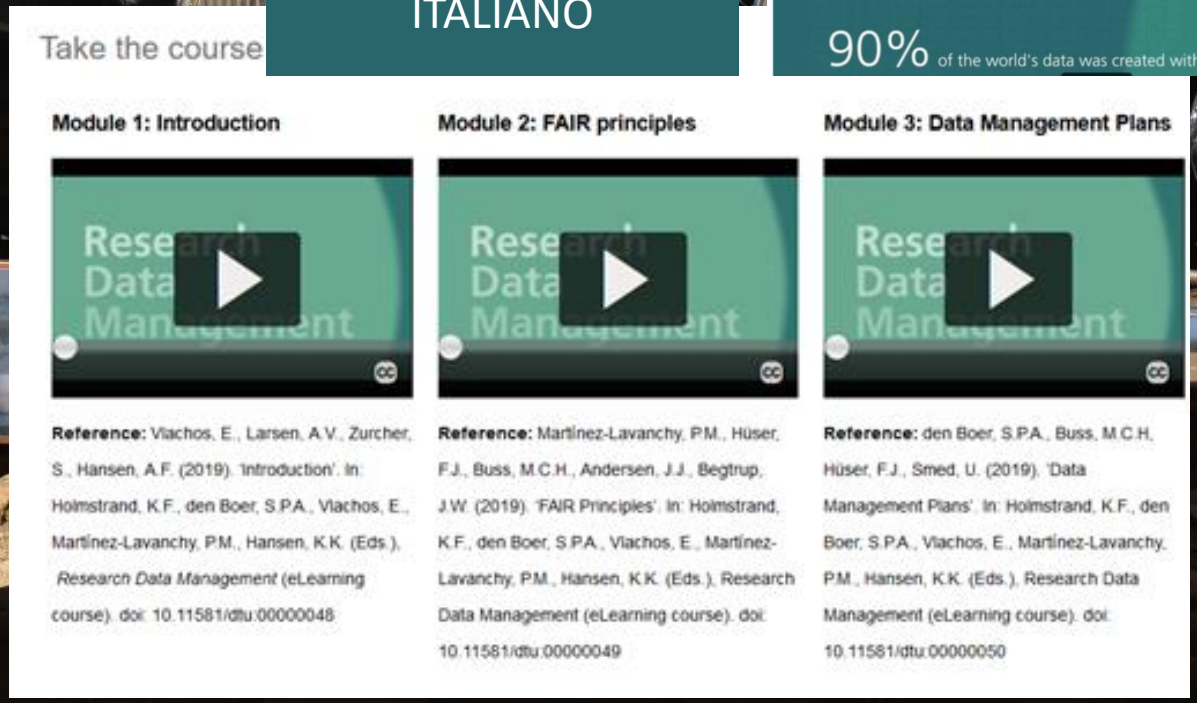
Data Stewardship for Open Science: Implementing FAIR Principles has been written with the intention of making scientists, funders, and innovators in all disciplines an...

crcpress.com

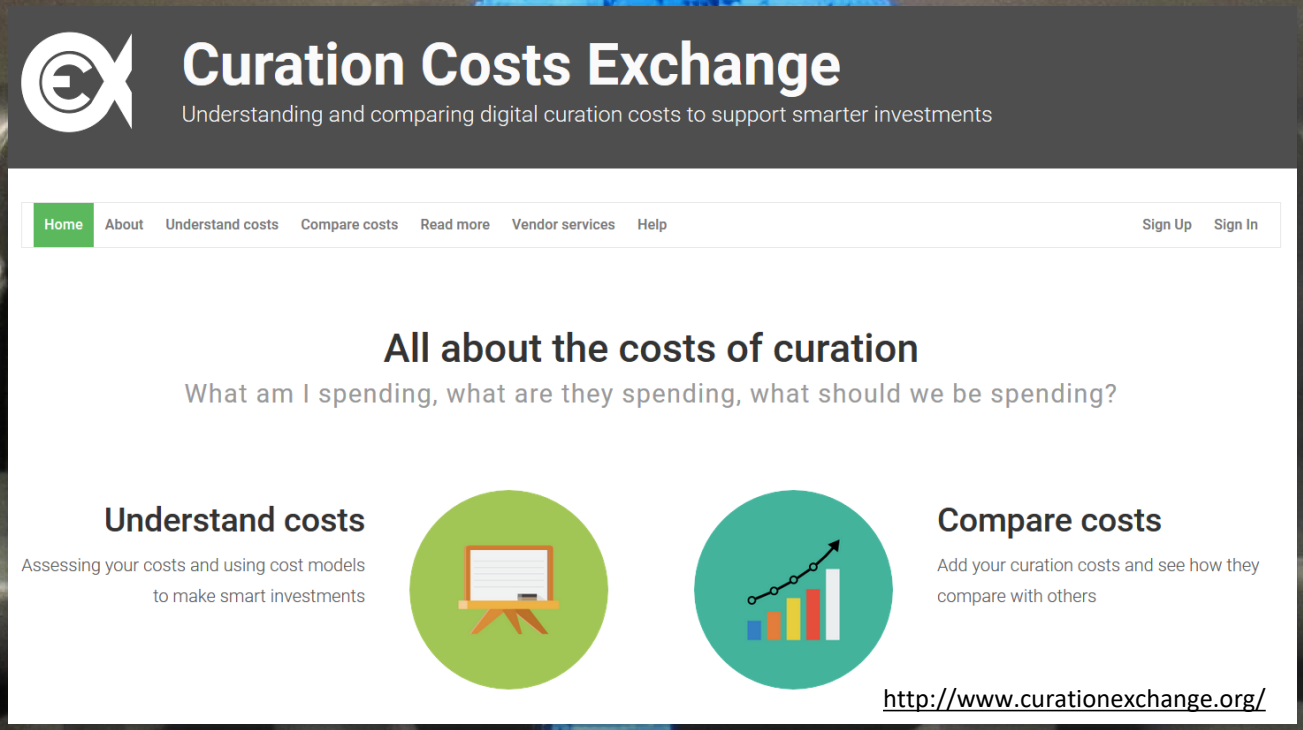
…una via veloce

LO TRADURREMO IN ITALIANO

https://vidensportal.deic.dk/RDMelearn



I costi



CI SONO COSTI PER CONSERVARE E GESTIRE I DATI…
MA PENSIAMO A QUANTO COSTEREBBE
NON CONSERVARLI E NON GESTIRLI

Il costo dei dati non-FAIR

Cost of not having FAIR research data

Cost-Benefit analysis for FAIR research data

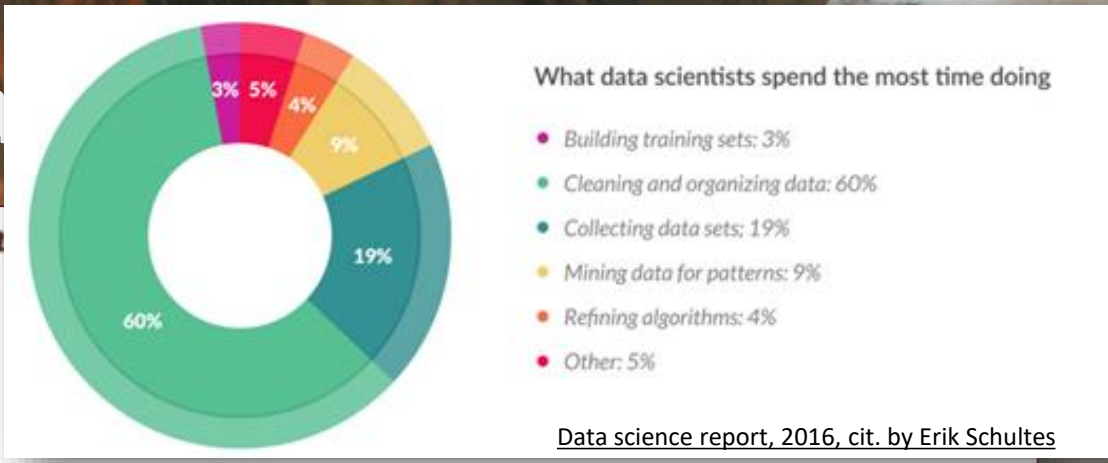
Following this approach, we found that the annual cost of not having FAIR research data costs the European economy at least €10.2bn every year. In addition, we also listed a number of consequences from not having FAIR which could not be reliably estimated, such as an impact on research quality, economic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

10,2 bn
16 bn

26,2 bn

Cost of not having FAIR data, 2018

FAIR data impa



Data science report, 2016, cit. by Erik Schultes

Cost of not having FAIR data, 2018


…un passo indietro…

[il fondamento]

Information Guide: Introduction to Ownership of Rights in Research Data. CREATe, University of Glasgow, 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.

http://eprints.gla.ac.uk/171314/

OpenAIRE

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 NON SONO «MIEI»
NON ESISTE COPYRIGHT
PERCHÉ NON SONO CREATIVI

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 … #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

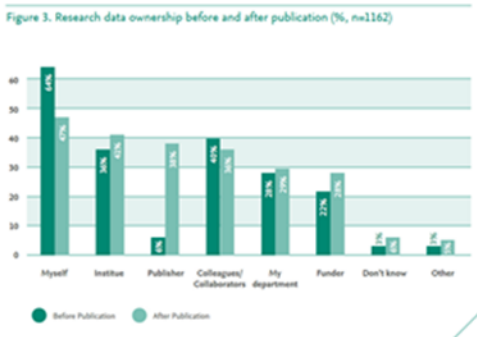
linkedin.com

11:18 - 12 apr 2017

14 Retweet **18** Mi piace

Lusoli, Apr.2017

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



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.

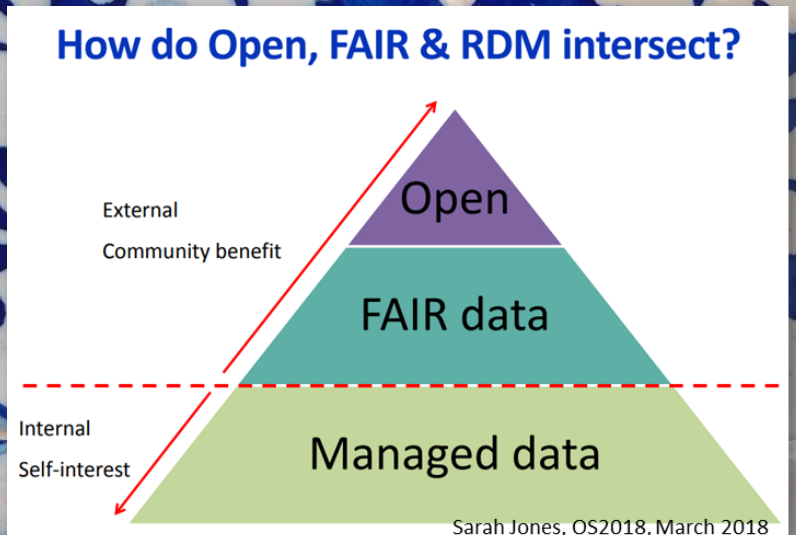
Of course, the study reports

[ricordandosi che ci sono 3 passi]

Open

FAIR

Gestione/Cura



1. I DATI VANNO CURATI

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.

ORGANIZZAZIONE (file naming, versioning…)

METADATI

BACKUP E STORAGE

CONSERVAZIONE SUL LUNGO PERIODO

ASPETTI LEGALI

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
I2. (meta)data use vocabularies that follow FAIR principles.
I3. (meta)data include qualified references to other (meta)data.

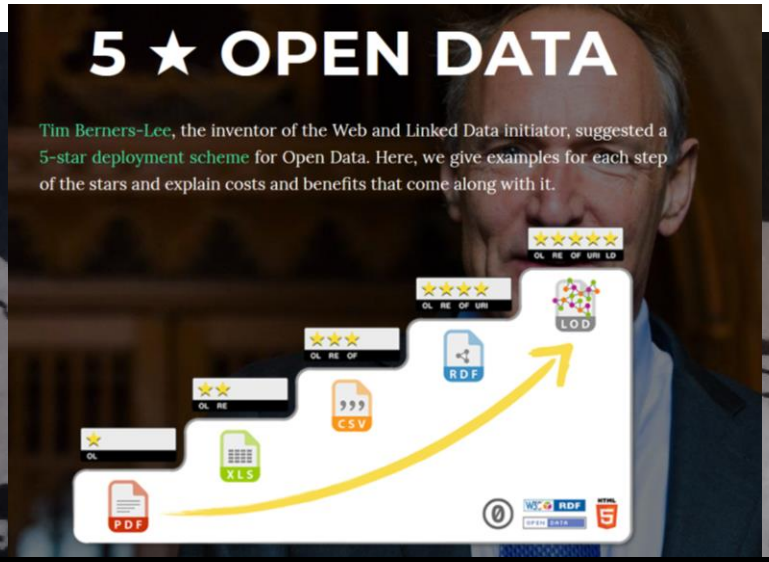
TO BE RE-USABLE:

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.

https://www.force11.org/group/fairgroup/fairprinciples

«ACCESIBLE» NON SIGNIFICA «OPEN». SIGNIFICA SOLO DICHIARARE LE CONDIZIONI SECONDO CUI I DATI SONO ACCESSIBILI

3. I DATI POSSONO ESSERE OPEN



★ make your stuff available on the Web (whatever format) under an open license[1]

★★ make it available as structured data (e.g., Excel instead of image scan of a table)[2]

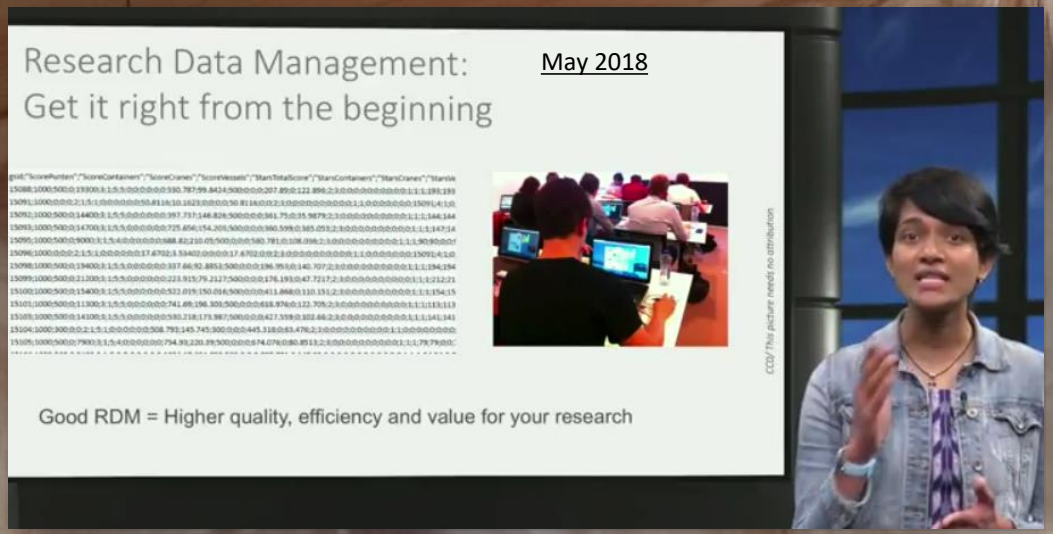
★★★ make it available in a non-proprietary open format (e.g., CSV instead of Excel)[3]

★★★★ use URIs to denote things, so that people can point at your stuff[4]

★★★★★ link your data to other data to provide context[5]

1. GESTIONE DEI DATI

… dall'inizio



Add a "version management" tab to your spreadsheet.

Now, let me expand on this idea.

Start by adding an extra "version management" tab to a new spreadsheet. In this sheet, carefully write down a version name (name of the file, typically) in the first column, in the second column the date, and in a third column an explanation of all changes you made to the sheet. Carefully fill out this sheet every single time you move something around, or tinker with the sheet.

If you're a starting PhD student, start doing this the very next time you build a new sheet. Thank me later.

If you already have multiheaded monstrous sheets: start by managing them in this way, and take a few extra hours to redefine the logic behind what you did earlier. Your dissertation writing self will thank you.

Data Management expert guide

cessda TRAINING

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?

https://www.cessda.eu/Training/Training-Resources/Library/Data-Management-Expert-Guide

Data Manageme

	Ad Hoc	One-Time	Active and Informative	Optimized for Re-Use
Planning your project	When it comes to my data, I have a "way of doing things" but no standard or documented plans.	I create some formal plans about how I will manage my data at the start of a project, but I generally don't refer back to them.	I develop detailed plans about how I will manage my data that I actively revisit and revise over the course of a project.	I have created plans for managing my data that are designed to stream its future use by myself others.
Organizing your data	I don't follow a consistent approach for keeping my data organized, so it often takes time to find things.	I have an approach for organizing my data, but I only put it into action after my project is complete.	I have an approach for organizing my data that I implement prospectively, but it not necessarily standardized.	I organize my data so th others can navigate, understand, and use it without me being present.
Saving and backing up your data	I decide what data is important while I am working on it and typically save it in a single location.	I know what data needs to be saved and I back it up after I'm done working on it to reduce the risk of loss.	I have a system for regularly saving important data while I am working on it. I have multiple backups.	I save my data in a manner and location designed maximize opportunities for re-use by myself and others.
Getting your data ready for analysis	I don't have a standardized or well documented process for preparing my data for analysis.	I have thought about how I will need to prepare my data, but I handle each case in a different manner.	My process for preparing data is standardized and well documented.	I prepare my data in such a way as to facilitate use by both myself and others in the future.
Analyzing your data and handling the outputs	I often have to redo my analyses or examine their products to determine what procedures or parameters were applied.	After I finish my analysis, I document the specific parameters, procedures, and protocols applied.	I regularly document the specifics of both my analysis workflow and decision making process while I am analyzing my data.	I have ensured that the specifics of my analysis workflow and decision making process can be understood and put into action by others.

Suppl. material 5: Draft Guide - Preparing doi

Authors: John Borghi

Data type: OpenDocument Text (.odt) file

Brief description: A draft guide that corresponds with the "Getting your data ready for analysis" row of the RDM rubric. Suggested points of customization are highlighted in yellow (discipline-specific) and red (institution-specific).

Filename: Draft Guide - Preparing.odt
Download file (59.52 kb)

Suppl. material 6: Draft Guide - Analyzing doi

Authors: John Borghi

Data type: OpenDocument Text (.odt) file

Brief description: A draft guide that corresponds with the "Analyzing your data and handling the outputs" row of the RDM rubric. Suggested points of customization are highlighted in yellow (discipline-specific) and red (institution-specific).

Filename: Draft Guide - Analyzing.odt
Download file (51.82 kb)

Suppl. material 7: Draft Guide - Sharing doi

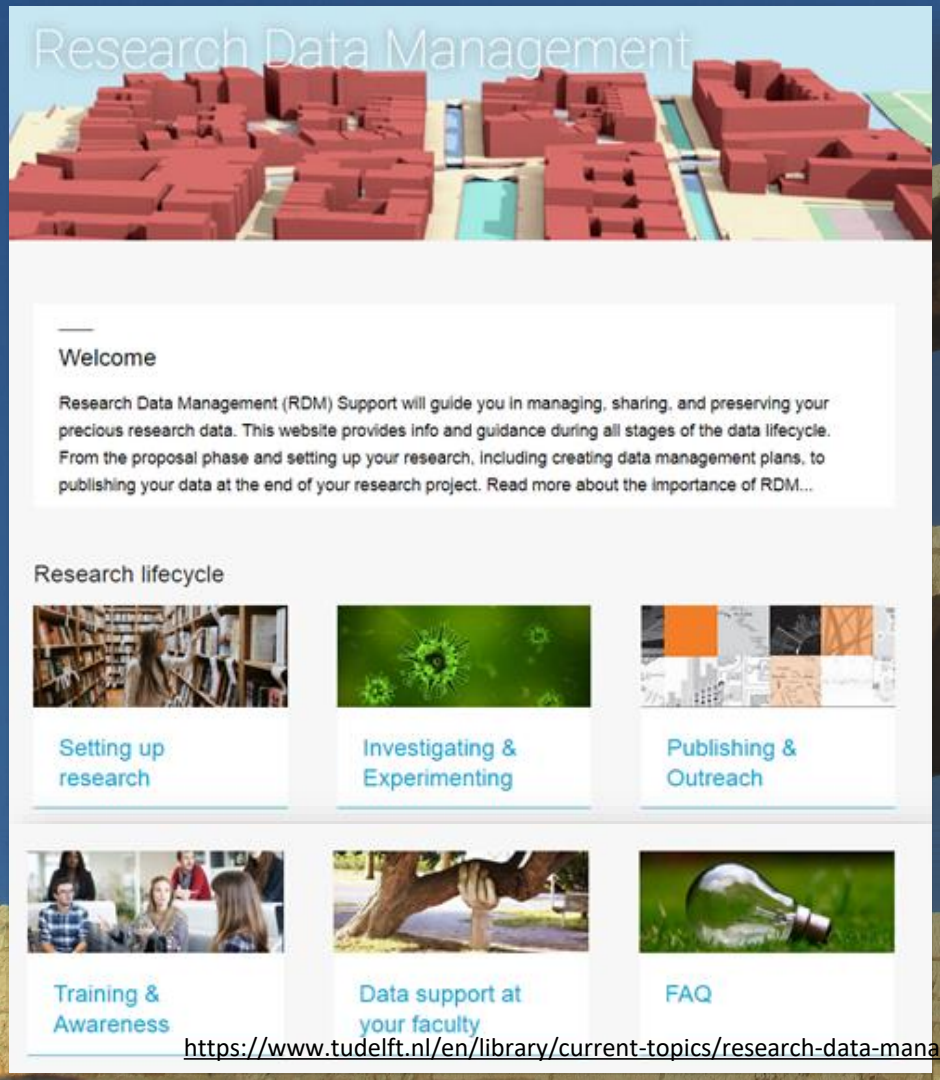
Authors: John Borghi

Data type: OpenDocument Text (.odt) file

Brief description: A draft guide that corresponds with the "Sharing and publishing your data" row of the RDM rubric. Suggested points of customization are highlighted in yellow (discipline-specific) and red

https://www.cessda.eu/

Supporto



https://www.tudelft.nl/en/library/current-topics/research-data-management/

Store your data

TU Delft offers its employees several options for storing and exchanging research data safely. Solutions range from basic storage to storage were you can store, organize and exchange data with colleagues all over the world.

The table below summarizes the current storage possibilities. The solution that fits your research data depends on your specific (security) needs.

Goal and solution

- I want to easily store data and share it selected others (in and outside of TU Delft).
- I want to store and backup personal data.
- I want to store, share and backup data with faculty colleagues.
- I want to store raw data.
- I want to store, backup and share data with other colleagues from TU Delft.
- I want to regularly exchange data internationally between groups.
- I want to share project information.

I want to store, backup, organise, annotate and share my data with others in and outside of TU Delft

DataverseNL is specifically designed to store, back-up, organise, annotate and share research data with colleagues all over the world. With this open source application you can grant multiple individuals controlled access to your data.

Supporto

THE UNIVERSITY of EDINBURGH

Information Services

RESEARCH DATA SERVICE

- **Research Data Service home**
- Why research data management?
- Training and support resources
- About the Research Data Service
- Before you begin
- Research in progress
- Approaching completion

Home > Information Services > Research Support > **Research Data Service**

Contact us

Why research data management?

Sharing and responsibly managing data produced or re-used in a research project is required by funders and publishers, and supports open scholarship.

Training and support resources

We offer a range of training programmes on research data, including regularly scheduled workshops, online courses and tailored events. Browse your options here.

About the Research Data Service

We provide a suite of tools and support that helps staff and students be effective with their research data before, during and after their project.

Before you begin

Create a data management plan to ensure that your research proposal is as strong as it can be.

Research in progress

Tools to help store, manage, protect and collaborate on research data during your project.

Approaching completion

Archive, preserve and share your data towards the end of a project, and link datasets to your publications with a DOI.

https://www.ed.ac.uk/information-services/research-support/research-data-service

Data management ABC – File naming

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, 'attch', 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.

https://www.ed.ac.uk/records-management/guidance/records/practical-guidance/naming-conventions

Data management ABC – File naming / 2

Data versioning

ands AUSTRALIAN NATIONAL DATA SERVICE

About us

Working with data

Data versioning

+SHARE

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 iden
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.

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

Numbering system 1

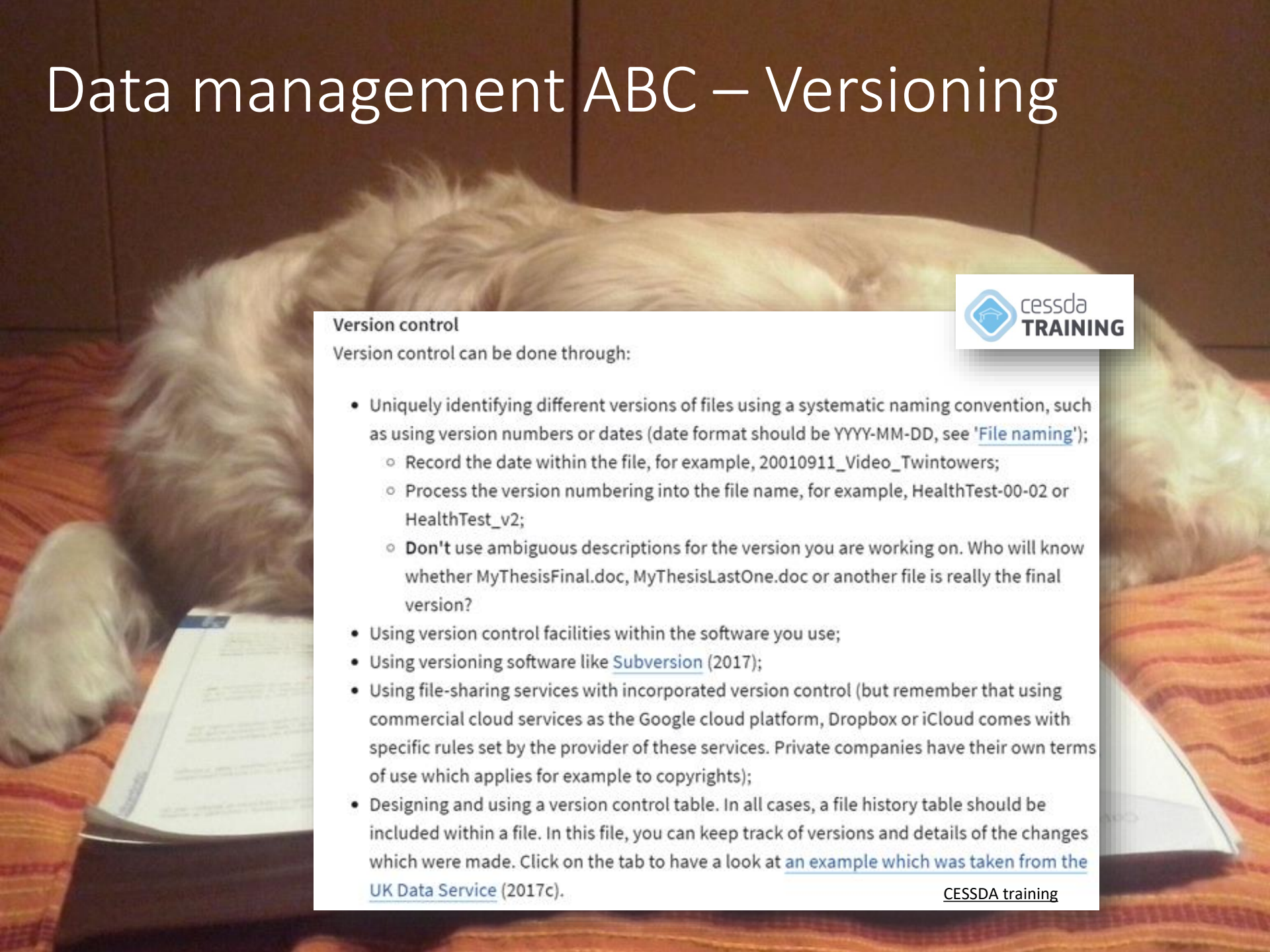
Data versioning follows a similar path to software versioning, usually applying a two-part numbering rule: Major.Minor (e.g. V2.1). Major data revision indicates a change in the formation and/or content of the dataset that may bring changes in scope, context or intended use. For example, a major revision 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



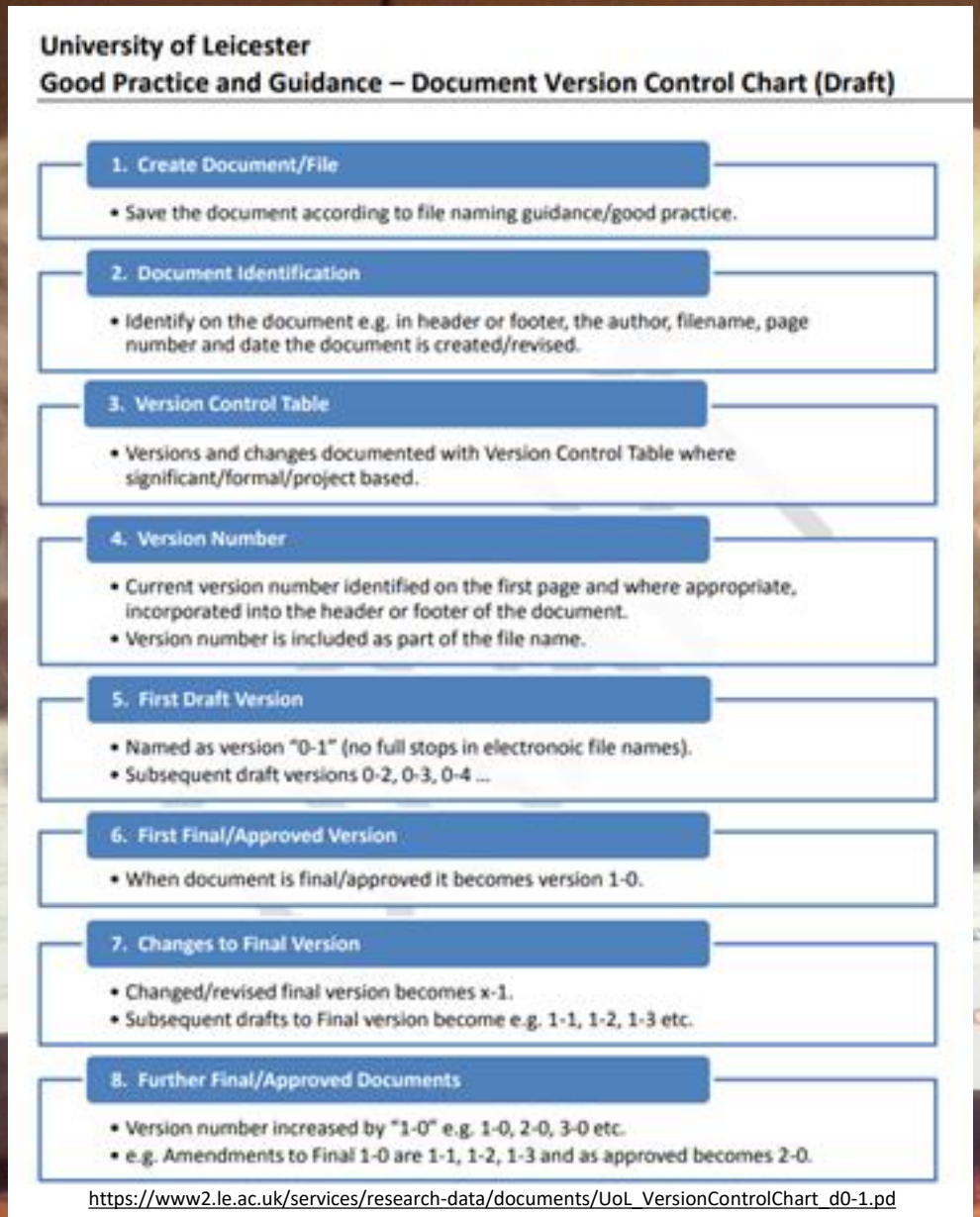
Version control

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



University of Leicester

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 electronoic 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

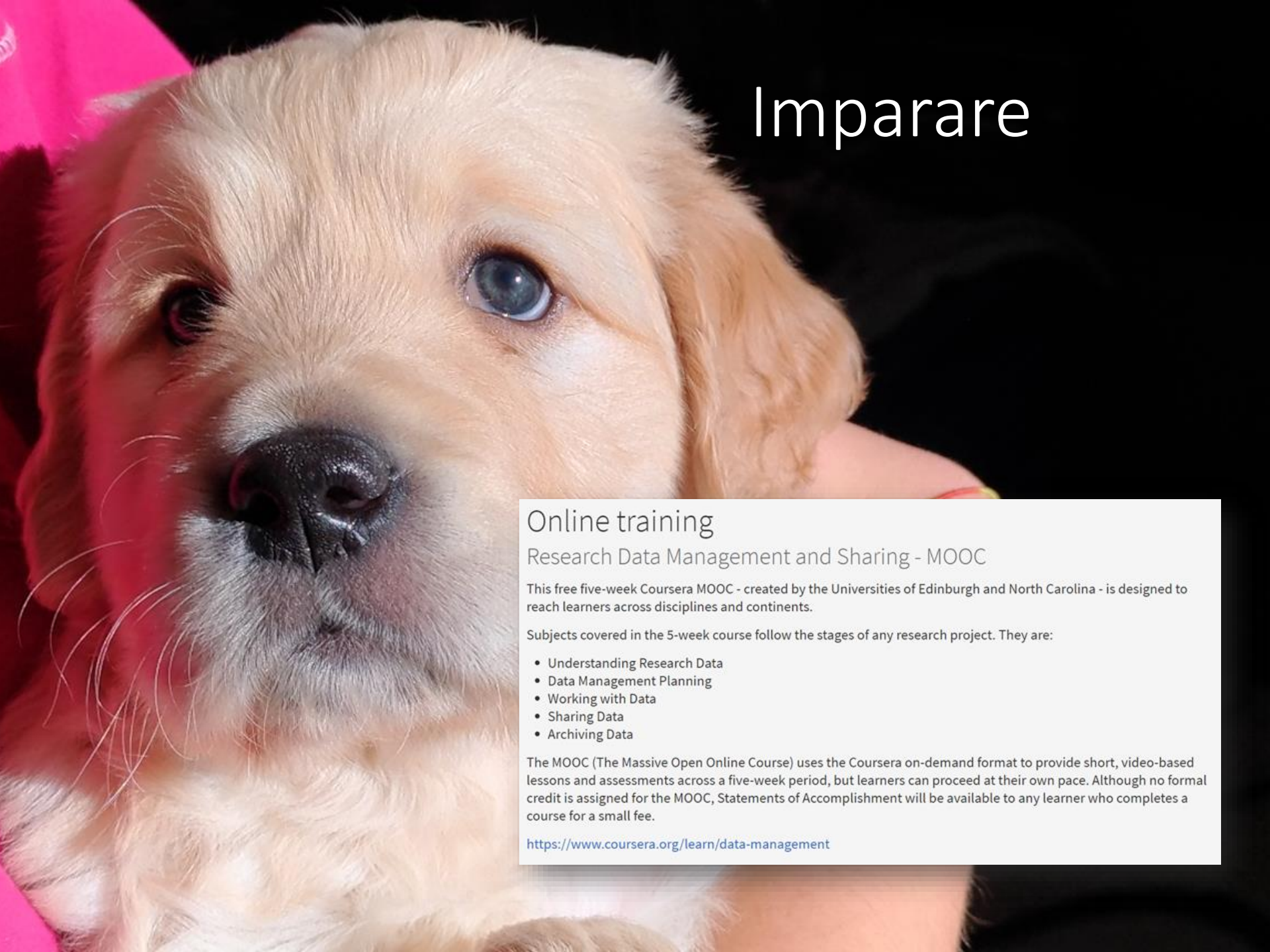
- Changed/revised final version becomes x-1.
- Subsequent drafts to Final version become e.g. 1-1, 1-2, 1-3 etc.

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.

https://www2.le.ac.uk/services/research-data/documents/UoL_VersionControlChart_d0-1.pd

Imparare



Online training

Research Data Management and Sharing - MOOC

This free five-week Coursera MOOC - created by the Universities of Edinburgh and North Carolina - is designed to reach learners across disciplines and continents.

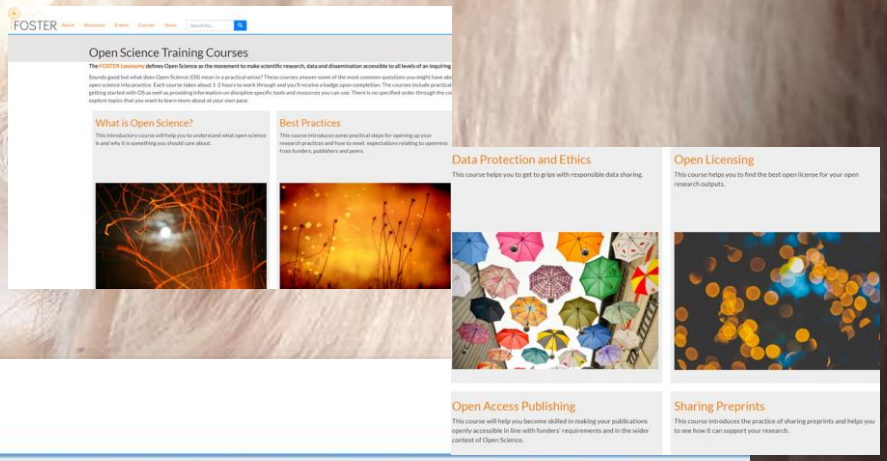
Subjects covered in the 5-week course follow the stages of any research project. They are:

- Understanding Research Data
- Data Management Planning
- Working with Data
- Sharing Data
- Archiving Data

The MOOC (The Massive Open Online Course) uses the Coursera on-demand format to provide short, video-based lessons and assessments across a five-week period, but learners can proceed at their own pace. Although no formal credit is assigned for the MOOC, Statements of Accomplishment will be available to any learner who completes a course for a small fee.

https://www.coursera.org/learn/data-management

Imparare a gestire



FOSTER About Resources Events Courses News Search for...

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

https://www.fosteropenscience.eu/node/2328

...parare 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 About Resources Events Courses News

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

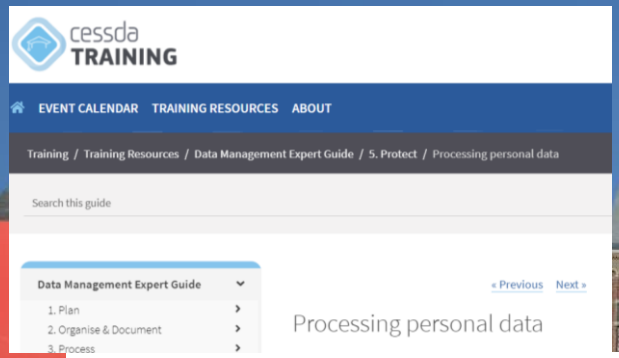
Start the Free Course

Full details

Level of knowledge: Introductory: no previous knowle
is required

Topics

[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, destruction or damage, using appropriate technical or organisational measures.

[leggi applicabili]

Science Europe 2018

Privacy

- **Personal Data Protection Acts** are present in all European countries and concern general laws regulating the protection of personal data. They are based on European Directive 95/46/EC.[9] This Directive will be replaced in the near future by the General Data Protection Regulation (GDPR),[10] which all EU Member States will have to implement in their national legislation by May 2018.
- **Obligations to Report Data Leakage Acts** are additions to the Personal Data Protection Acts. They deal with the publication of personal data and contain sanctions in the form of penalties.
- **Medical Treatment Agreement Acts** regulate the use and preservation of personal (patient) data in and for medical research.
- **Scientific Medical Research with Humans Acts** regulate scientific research in the medical field, in particular how to handle personal health-related data. These make ethical reviews compulsory for all medical research projects.

Intellectual Property Rights

- **Copyright Acts** regulate the rights of the creator of a work. One distinguishes between exploitation rights and personal intellectual rights ('moral rights').
- The **Database Rights Act** recognises the investments made in creating and/or compiling a database. It is based on European Directive 96/9/EC.[11]
- **Related Rights Acts** or **Neighbouring Rights Acts** mostly refer to the rights of performers, phonogram producers, and broadcasting organisations.
- **Patent Acts** are for the protection of patents. Publication of research results (including data) is restricted during the application stage of a patent.

Public data

- **Public Records Acts** (Public Archives Acts) oblige all public administration offices and services to preserve their documents and transfer these, after appraisal and selection, to public archives.
- **Public Sector Information Acts** (concerning re-usability of public data) are based on European Directive 2013/37/EU[12] that focuses on the economic aspects of the re-use of public information. It encourages Member States to make as much of this information as possible available for re-use. This also covers content held by museums, libraries, and archives, but does not apply

- **Freedom of information Acts** regulate and enable citizen access to documents held by public authorities or companies carrying out work for a public authority. They do not specifically deal with access to research data.
- **Heritage Acts** are relevant for archaeological research data in so far as that they regulate ownership of documentation (data) from archaeological excavations.
- **Statistical Information Acts** regulate the competencies of the statistics authorities in data gathering as well in access to data.
- **Land Registry Acts** (cadastral information) regulate the competencies of the national land registries and access to their data, with special provisions concerning personal data contained in their various databases.

Codes of Conduct/Ethical Issues

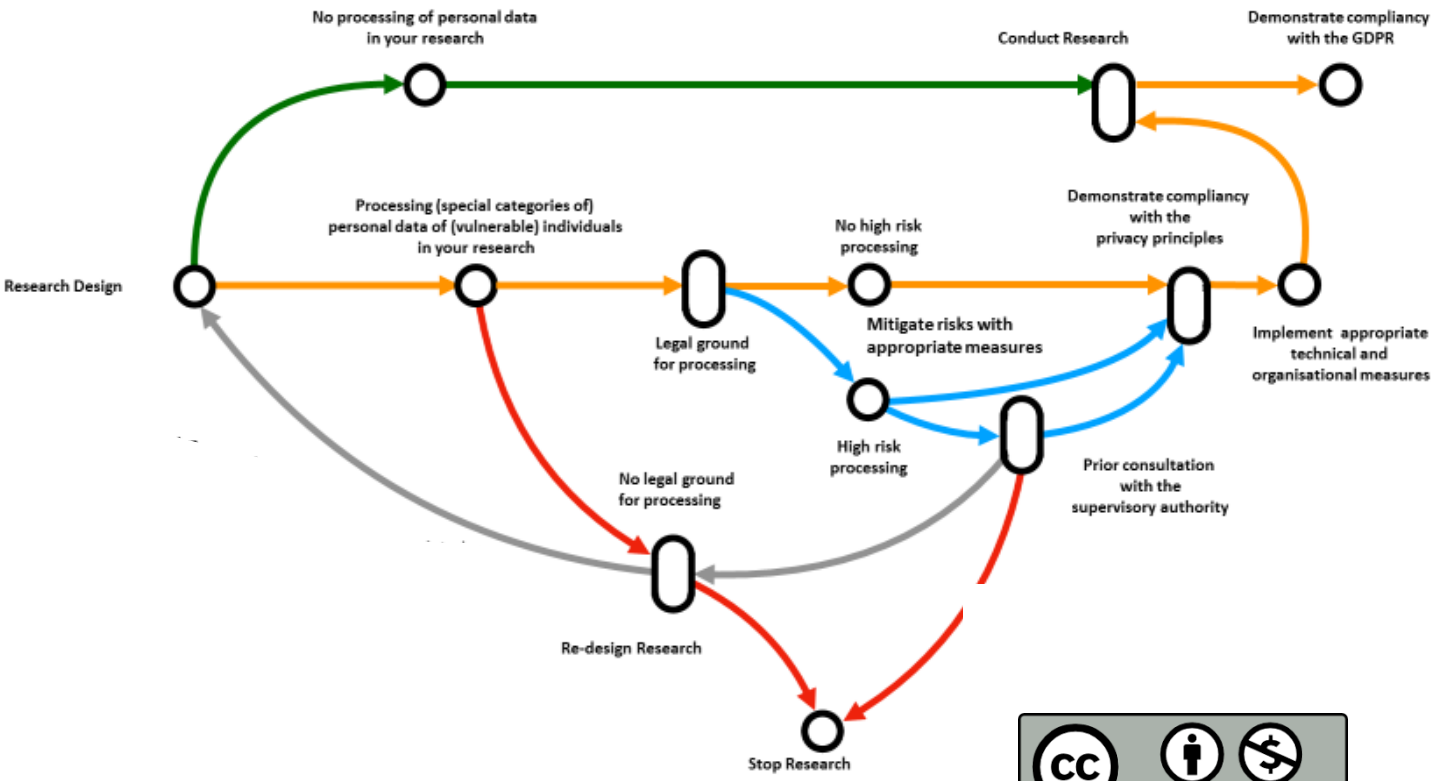
- **Codes of Conduct**, where these exist on a national level or in an institution, should be taken into account in DMPs. They contain the general principles of good academic teaching and research.
- **Codes of Practice** for the use of personal data in scientific and scholarly research are based on the Personal Data Protection Acts[13] and prescribe how to handle personal data in research practice.
- **Codes of Conduct** for Medical Research regulate how researchers should handle medical personal data. They may be based on Medical Treatment Agreement Acts.

[Dati e GDPR]


Ligue des Bibliothèques Européennes de Recherche
Association of European Research Libraries
About Us
Network
Strategy
Login
Scholarly Communication
Digital Skills & Services
Research Infrastructure
Events
Webinar Video: GDPR & What It Means For Researchers
Menu
May 22 2018
LIBER Webinar: GDPR & What It Means For Researchers
The Privacy Impact Assessment (PIA) Route Planner for Academic Research
Inspired by Harry Beck's London Metro Map
Research Design
Conduct Research
Legal ground for processing
No legal ground for processing
Re-design Research
Stop Research
Chat : (Everyone)

[Dati e 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 → **Q2. What is the legal ground for this processing?**

NO
Proceed - no measures required for safeguardingp rivacy.

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.

YES → **Q3. Is this processing a high risk processing?**

NO
Stop research or redefine research.

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.

"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 compliancy 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.

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?

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.

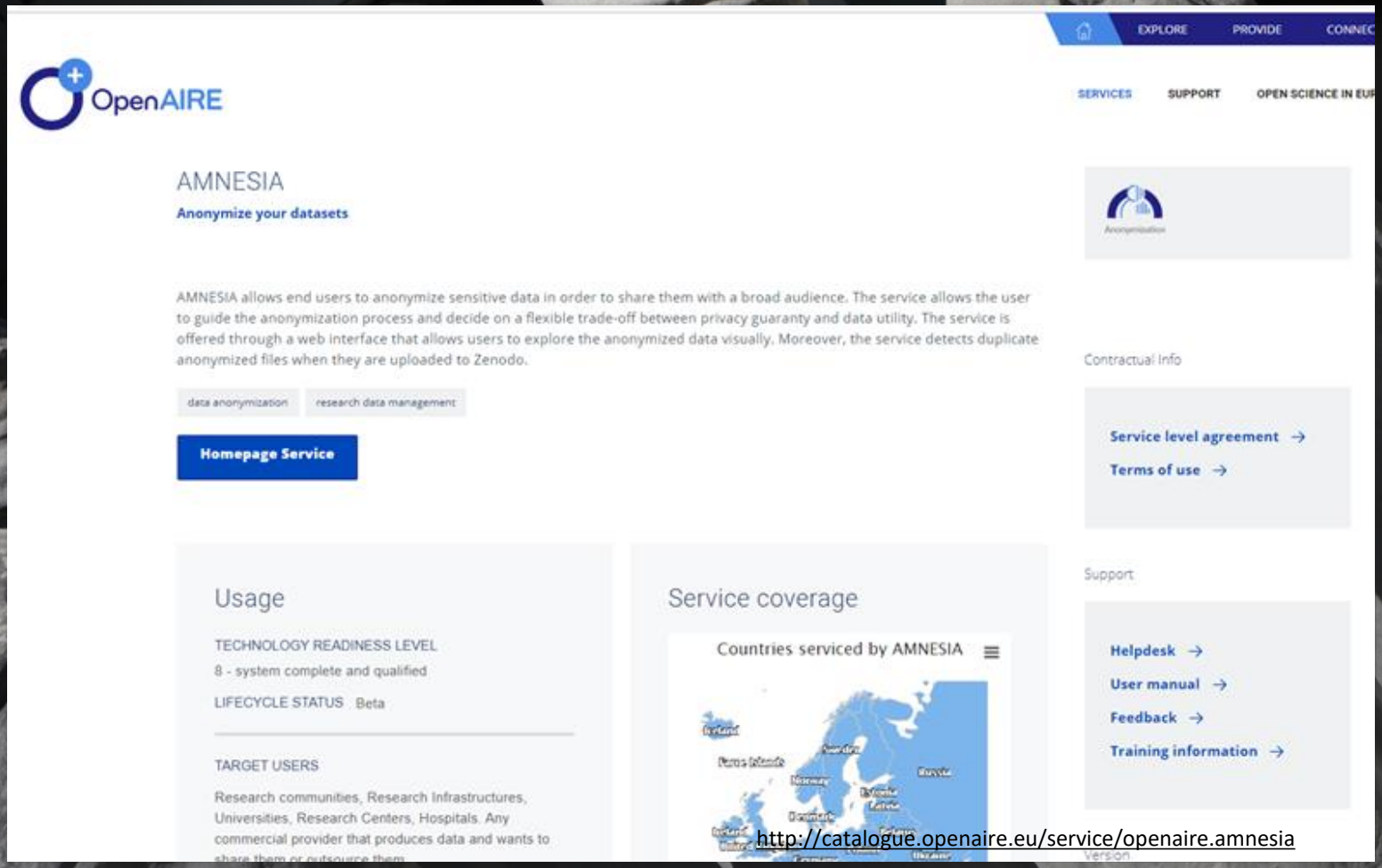
Action

Principles relating to processing of personal data (GDPR*, Article 5):
Demonstrate compliancy 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

[anonimizzare i dati]



http://catalogue.openaire.eu/service/openaire.amnesia

2. RENDERE I DATI FAIR

FAIR data

Findable

discoverable with machine readable metadata, identifiable and locatable by means of a standard identification mechanism

Accessible

available and obtainable to both human and machine

Interoperable

both syntactically parseable and semantically understandable, allowing data exchange and reuse among scientific disciplines, researchers, institutions, organisations and countries

Reusable

sufficiently described and shared with the least restrictive licences, allowing the widest reuse possible across scientific disciplines and borders, and the least cumbersome integration with other data sources

FAIR principles for dummies

GO FAIR | GO FAIR Initiative | Implementation Networks | FAIR Principles

FAIR Principles

Findable

The first step in (re)using data is to find them. Metadata and data should be easy to find for both humans and computers. Machine-readable metadata are essential for automatic discovery of datasets and services, so this is an esse

F1. (Meta)data are assigned a globally

F2. Data are described with rich meta

F3. Metadata clearly and explicitly in

F4. (Meta)data are registered or index

Accessible

Once the user finds the required data, s
including authentication and authorisat

A1. (Meta)data are retrievable by the protocol

A1.1 The protocol is open, free, a

A1.2 The protocol allows for an a necessary

What does this mean?

Principle F1 is arguably the most important because it will be hard to achieve other aspects of FAIR without globally unique and persistent identifiers. Hence, compliance with F1 will already take you a long way towards publishing FAIR data (see **10 ways identifiers can help with data integration**).

Globally unique and persistent identifiers remove ambiguity in the meaning of your published data by assigning a unique identifier to every element of metadata and every concept/measurement in your dataset. In this context, identifiers consist of an internet link (e.g., a URL that resolves to a web page that defines the concept such as a particular human protein: **http://www.uniprot.org/uniprot/P98161**). Many data repositories will automatically generate globally unique and persistent identifiers to deposited datasets. Identifiers can help other people understand exactly what you mean, and they allow computers to interpret your data in a meaningful way (i.e., computers that are searching for your data or trying to automatically integrate them). Identifiers are essential to the human-machine interoperation that is key to the vision of **Open Science**. In addition, identifiers will help others to properly cite your work when reusing your data.

Of course, identifiers are one thing, but their meaning is another (see principles I1-I3). F1 stipulates two conditions for your identifier:

1. It must be globally unique (i.e., someone else could not reuse/reassign the same identifier without referring to your data). You can obtain globally unique identifiers from a registry service that uses algorithms guaranteeing the uniqueness of newly minted identifiers.
2. It must be persistent. It takes time and money to keep web links active, so links tend to

https://www.go-fair.org/fair-principles/

FAIR data

Cloudy, increasingly FAIR; revisiting the FAIR Data guiding principles for the European Open Science Cloud

2017

Article type: Research Article

Authors: Mons, Barend[a; b; c; *] | Neylon, Cameron[d] | Velterop, Jan[e] | Dumontier, Michel[f] | da Silva Santos, Luiz Olavo Bonino[b; g] | Wilkinson, Mark D.[h]

3. What FAIR is...

FAIR refers to a set of principles, focused on ensuring that research objects are reusable, and actually will be reused, and so become as valuable as is possible. They deliberately do not specify technical requirements, but are a set of guiding principles that provide for a continuum of increasing reusability, via many different implementations. They describe characteristics and aspirations for systems and services to support the creation of valuable research outputs that could then be rigorously evaluated and extensively reused, with appropriate credit, to the benefit of both creator and user.

FAIR data

Cloudy, increasingly FAIR; revisiting the FAIR Data guiding principles for the European Open Science Cloud

2017

Article type: Research Article

Authors: Mons, Barend[a; b; c; *] | Neylon, Cameron[d] | Velterop, Jan[e] | Dumontier, Michel[f] | da Silva Santos, Luiz Olavo Bonino[b; g] | Wilkinson, Mark D.[h]

4. ...and what FAIR is not

FAIR is not a standard: The FAIR guiding principles are sometimes incorrectly referred to as a 'standard', even though the original publication explicitly states they are not [25]. The guiding principles allow many different approaches to rendering data and services Findable, Accessible, Interoperable, to serve the ultimate goal: the reuse of valuable research objects. Standards are prescriptive, while guidelines are permissive. We suggest that a variety of valuable standards can and should be developed, each of which is guided by the FAIR Principles. FAIR simply describes the qualities or behaviours required of data resources to achieve – possibly incrementally – their optimal discovery and scholarly reuse.

FAIR is not equal to RDF, Linked Data, or the Semantic Web The reference article in Scientific Data [25] emphasises the machine-actionability of data and metadata. This implies (in fact, requires) that resources that wish to maximally fulfil the FAIR guidelines must utilise a widely-accepted machine-readable framework for data and knowledge

FAIR is not just about humans being able to find, access, reformat and finally reuse data: The official press release following the publication of the FAIR Principles states the authors' position clearly: "The recognition that computers must be capable of accessing a data publication autonomously, unaided by their human operators, is core to the FAIR Principles. Computers are now an inseparable companion in every research endeavour". In recent surveys, the time reportedly spent by PhD students and other researchers in projects dealing with discovering and reusing multiple data sources – so called 'data munging' – has been pegged at 80% [19]. Were these colleagues and their machine-assistants only having to deal with FAIR data and services, this wasted time would be reduced to a fraction of what it is today. The avoidance of time-wasting would be a first return on investment in good data stewardship. To serve this potentially enormous cost reduction, FAIR compliant (meta)data and services should be actionable by machines without human supervision whenever and wherever possible.

FAIR is not equal to Open: The 'A' in FAIR stands for 'Accessible under well defined conditions'. There may be legitimate reasons to shield data and services generated with public funding from public access. These include personal privacy, national security, and competitiveness. The FAIR principles, although inspired by Open Science, explicitly and

FAIR / Open

Turning FAIR into reality, 2018

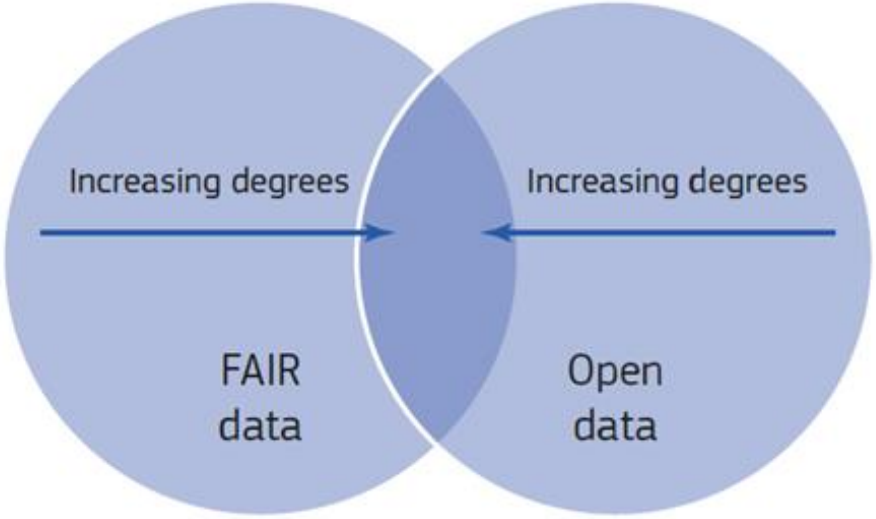


Figure 4. The relationship between FAIR and Open

Data can be FAIR or Open, both or neither. The greatest benefits come when data are both FAIR and Open, as the lack of restrictions supports the widest possible reuse, and reuse at scale. To maximise the benefits of making FAIR data a reality, and in the context of Open Science initiatives, the FAIR principles should be implemented in combination with a policy requirement that research data should be Open by default - that is, Open unless there is a good reason for restricting access or reuse. In recent European Commission formulations, the maxim 'as open as possible, as closed as necessary' has been introduced, which is a helpful articulation of the principles

Rec. 17: Align and harmonise FAIR and Open data policy

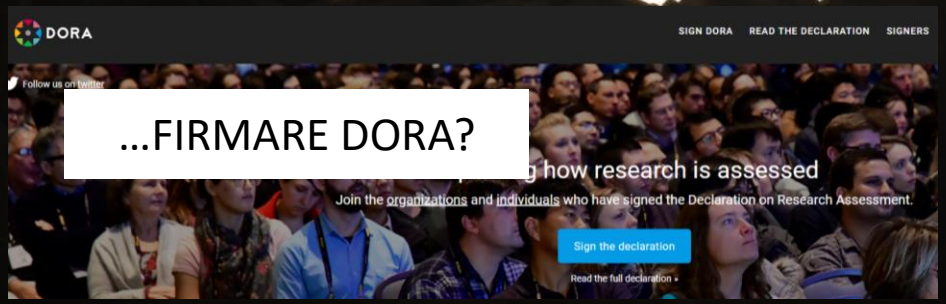
Policies should be aligned and consolidated to ensure that publicly-funded research data are made FAIR and Open, except for legitimate restrictions. The maxim 'as Open as possible, as closed as necessary' should be applied proportionately with genuine best efforts to share.

FAIR data – dati e valutazione

The system of incentives and rewards must also be addressed in a fundamental way. From the perspective of measuring and rewarding research contributions, the full diversity of outputs should be taken into account including FAIR data, code, workflows, models, and other digital research objects as well as their curation and maintenance. In the 21st century, traditional publications and journal articles are far from being the only significant contributions to the advancement of knowledge.

…SMETTETE DI ADORARE
L' IMPACT FACTOR

…E GLI ARTICOLI. ESISTONO I DATI, SOFTWARE, PROTOCOLLI



…FIRMARE DORA?

Turning FAIR into reality, 2018

FAIR data - componenti

Turning FAIR into reality, 2018

- POLICIES
- DATA MANAGEMENT PLANS
- IDENTIFICATIVI PERSISTENTI
- STANDARDS
- REPOSITORIES

FAIR Digital Objects sit in a wider **FAIR ecosystem** comprising services and infrastructures for FAIR. The realisation of FAIR relies on, at a minimum, the following essential components: policies, DMPs, identifiers, standards and repositories. In this ecosystem, data policies are issued by several stakeholders and help to define and regulate requirements for the running of data services. Data Management Plans provide a dynamic index that articulates the relevant information relating to a project and linkages with its various FAIR components. Persistent Identifiers are assigned to many aspects of the ecosystem including data, software, institutions, researchers, funders, projects and instruments. Specifications and standards are relevant in many ways, from metadata, vocabularies and ontologies for data description to transfer and exchange protocols for data access, and standards governing the certification of repositories or composition of DMPs. Repositories offer databases and data services and should be certified to ensure trust.

FAIR data – ecosistema distribuito

EOSC = DISTRIBUITO. I DATI RESTANO DOVE SONO

Turning FAIR into reality, 2018

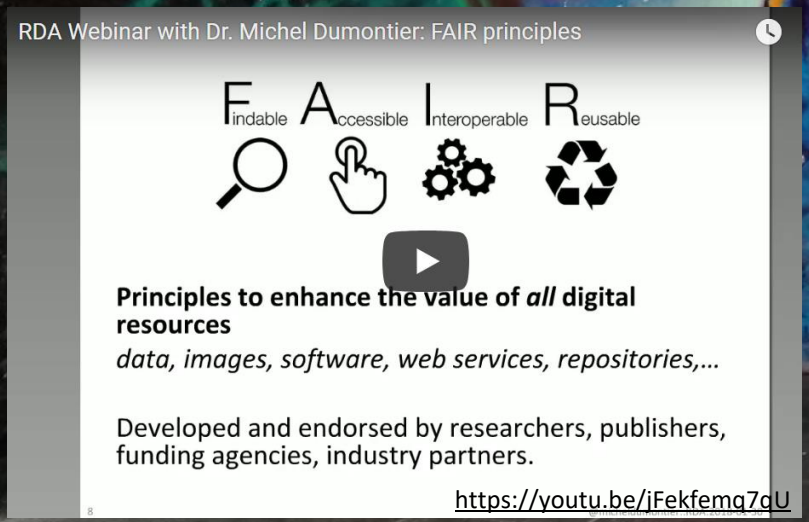
The future FAIR ecosystem will necessarily be highly distributed. It will require technical mechanisms for linking resources as well as collaboration mechanisms for coordination and for agreement about specifications and standards. EOSC will have an important role to play in each of these mechanisms. For the FAIR ecosystem to work, there need to be registries cataloguing the component services and automated workflows between them. Federations offer a means to establish agreements between repositories or registries to carry out certain tasks collaboratively and therefore will be essential to this distributed system. Data will increasingly remain at different locations for reasons such as the expense of copying data or because of legal or ethical restrictions. Distributed queries, managed by brokering software, will be used to virtually integrate data. The need for such

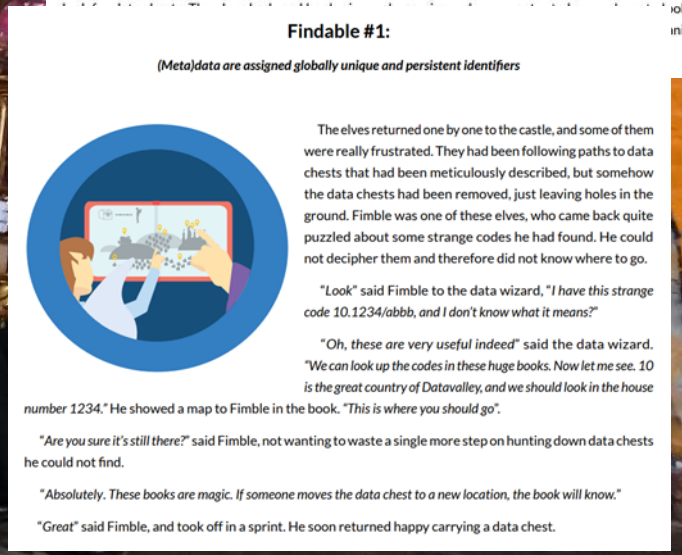
…FAIR data

Introduction

Once upon a time in the beautiful kingdom of Datamania lived a prince named Prince Fairhair. Though he was gentle as few, and good looking too, his father would not let him choose the love of his life on his own. No, he was destined to marry a woman from the neighbouring kingdom. He did not even know her name, only that she was referred to as My Fair Lady. Before the father of My Fair Lady could accept the marriage, he had a quest for Prince Fairhair. Only by fulfilling the quest, would he be able to marry the princess. His quest was to find out how to turn water into gold. A quest that would require gathering loads of data chests and look for clues that could lead to the recipe.

Luckily, Prince Fairhair was not alone in his quest. One of the castle wings housed a number of wizards who could help him decrypt and investigate the data chests. However, it was impossible for the data wizards to go and hunt for data themselves. Thus to assist them, a huge number of elves were trained to





Findable #1:

(Meta)data are assigned globally unique and persistent identifiers

The elves returned one by one to the castle, and some of them were really frustrated. They had been following paths to data chests that had been meticulously described, but somehow the data chests had been removed, just leaving holes in the ground. Fimble was one of these elves, who came back quite puzzled about some strange codes he had found. He could not decipher them and therefore did not know where to go.

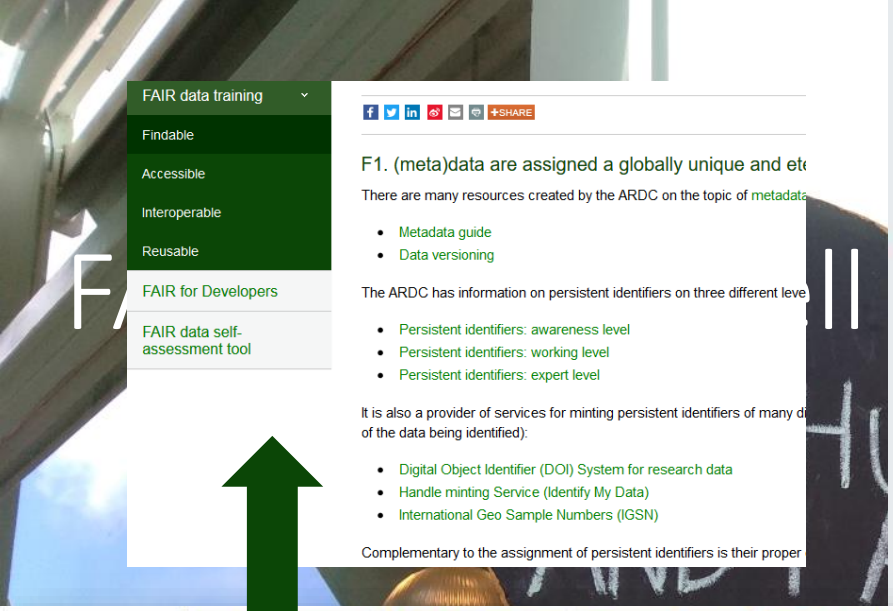
"Look" said Fimble to the data wizard, "*I have this strange code 10.1234/abbb, and I don't know what it means?*"

"Oh, *these are very useful indeed*" said the data wizard. "*We can look up the codes in these huge books. Now let me see. 10 is the great country of Datavalley, and we should look in the house number 1234.*" He showed a map to Fimble in the book. "*This is where you should go*".

"*Are you sure it's still there?*" said Fimble, not wanting to waste a single more step on hunting down data chests he could not find.

"*Absolutely. These books are magic. If someone moves the data chest to a new location, the book will know.*"

"*Great*" said Fimble, and took off in a sprint. He soon returned happy carrying a data chest.



https://www.ands.org.au/working-with-data/fairdata/training

Rob Hooft, DTL

FAIR Data management wizard

ELIXIR
Data Stewardship Wizard

Smart Data Management Plans for FAIR Open Science

For serious researchers and data stewards

Data Stewardship Wizard

Data integration 7

Data interpretation 3

Information and insight 14

Is there any pre-existing data?

Are there any data sets av

Data Stewardship for C

No

Yes

Will reference data be cr

Will any of the data that y
others)?

Data Stewardship for C

No

Yes

Will you be storing sample

Data Stewardship Wizard | Questionnaire Demo | Log In | Sign Up

Will you store licenses with the data?

What's up?

Always consider the use of your data beyond the original purpose. One of the issues with re-using other people's' data is that they cannot be assumed to be reusable from an ethical or legal standpoint without explicit permission. Assuming that unlicensed data are 'free to use for whatever purpose' is intrinsically wrong, and in the case of for instance Pharmaceutical industry can lead to court cases later on. Therefore, whenever you publish a data set or any other kind of information or digital object, it is important to define a license for reuse. For software many licences exist, and for data, increasingly, standard licenses are available or under development. Please note that a given license is also a defined concept and therefore deserves a Persistent Identifier and a URI pointing to where the license can be studied (machine readable licenses are also under development in some areas). This means that in the metadata, the license under which the data or the workflow can be reused is 'just another PID in the right place'. Users can then specify in their search or workflow container that 'only data with the following licenses should be included'.

For instance, if you include some data in your analysis that cannot be used for commercial purposes, that decision may render your entire results not usable for commercial purposes (at least in the view of some lawyers). This means that not licensing your data at all, even if you don't care who uses them and for what purpose is very counterproductive and will severely undermine the actual reuse of your data by others and in particular by industry. It will also lower the attribution-rate (usually part of the license conditions) and thus the citation and the impact score of your data.

Do

- Always carefully choose a license to be attached to your data upon publication.
- Include and clearly mark the licences PID as a concept + attributes in the metadata.
- Store and 'expose' the license as part of the metadata in Open Access environments where search engines can easily find the license, even of the data they describe are not (yet) FAIR or even highly restricted in access. The 'fact' that a data set with a specific license is 'out there' is a first step toward effective reuse of your data or information source.
- Make sure, especially when you restrict use of your data, that you are able to enforce the license you choose. Licenses that are not enforceable make no sense. (please note that the enforcement is usually not done by an individual research group but at institutional or repository level)

Don't

- Ever publish data without a license attached or choose a license lightly, without considerations of anticipated reuse of your data.
- Choose a license that is not transitive (i.e can not be transferred with subsets of the data), but make sure its transitivity does not unduly restrict the reuse of your data.
- Choose an unnecessary complicated license with many clauses and wherever possible one that is already widely adopted in the research community for either software

Top 10 FA

ARDC Australian Research Data Commons | Library Carpentry

Top 10 FAIR Data & Software Things

February 1, 2019

Sprinters:

Reid Otsuji, Stephanie Labou, Ryan Johnson, Guilherme Castelao, Bia Villas Boas, Anna-

Geoscience

Sprinters:

John Brown, Janice Chan, Niamh Quigley (Curtin Univer

Audience:

Researchers

Things

Findable

Thing 1: Data sharing and discovery

Thing 6: Vocabularies for data description

Thing 7: Identifiers and linked data

Thing 10: Spatial data

Accessible

Thing 2: Long-lived data: curation & preservation

Thing 3: Data citation for access & attribution

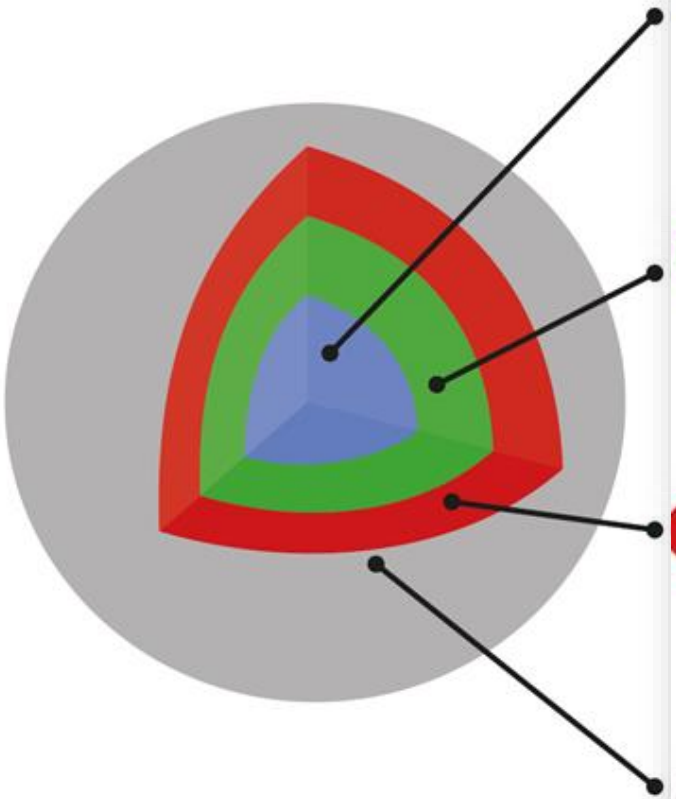
markup a dataset can be found here. The dataset is part of a project to reconstruct the domestic market for colonial goods in the Dutch Republic.

ACTIVITIES:
1. Try to find one or two terms that are relevant to your research using the resources that are mentioned above. You can aso use Swoogle to search for vocabularies related to your research. 2. Search for a term related to your research in the CIDOC Conceptual Reference Model (CRM) concept search. Were you able to find it? Tip 1: Search for "person" to get an idea of how the thesaurus works. Tip 2: All the terms used can be found in the last release of the model: http://www.cidoc-crm.org/get-last-official-release.

Thing 7: FAIR data modelling

The fourth and the fifth star in Berner Lee's model can be awarded when the data are stored in a format in which the topics their properties and their characteristics are identified using URIs whenever possible. More concretely, it implies that you record your data using the Resource Description Framework (RDF) format. RDF, simply put, is a technology which enables you to publish the contents of a database via the web. It is based on a simple data model which assumes that all statements about resources can be reduced to a basic form,

FAIR data: the ideal FAIR object



DIGITAL OBJECT

Data, code and other research outputs

At its most basic level, data or code is a bitstream or binary sequence. For this to have meaning and to be FAIR, it needs to be represented in standard formats and be accompanied by Persistent Identifiers (PIDs), metadata and documentation. These layers of meaning enrich the object and enable reuse.

IDENTIFIERS

Persistent and unique (PIDs)

Digital Objects should be assigned a unique and persistent identifier such as a DOI or URN. This enables stable links to the object and support citation and reuse to be tracked. Identifiers should also be applied to other related concepts such as the data authors (ORCIDs), projects (RAIDs), funders and associated research resources (RRIDs).

STANDARDS & CODE

Open, documented formats

Digital Objects should be represented in common and ideally open file formats. This enables others to reuse them as the format is in widespread use and software is available to read the files. Open and well-documented formats are easier to preserve. Data also need to be accompanied by the code use to process and analyse the data.

METADATA

Contextual documentation

In order for Digital Objects to be assessable and reusable, they should be accompanied by sufficient metadata and documentation. Basic metadata will enable data discovery, but much richer information and provenance is required to understand how, why, when and by whom the objects were created. To enable the broadest reuse, they should be accompanied by a plurality of relevant attributes and a clear and accessible usage license.

ECD, 2015

… quanto siete FAIR?

Findable

Does the dataset have any identifiers assigned? — No identifier

Is the dataset identifier included in all metadata records/files describing the data? — No

How is the data described with metadata? — The data is not described

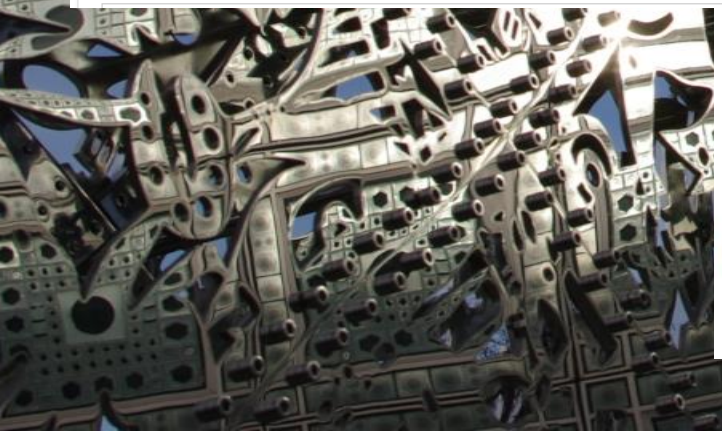
What type of repository or registry is the metadata record in? — The data is not described in any repository

Accessible

How accessible is the data? — No access to data or metadata

Is the data available online without requiring specialised protocols or tools once access has been approved? — No access to data

Will the metadata record be available even if the data is no longer available? — Unsure



https://www.ands-nectar-rds.org.au/fair-tool

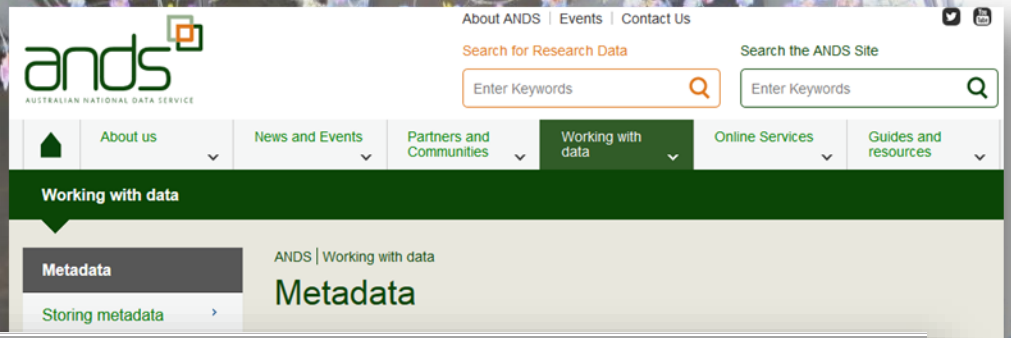
RDS ands nectar

home | news | events | programs | about

FAIR self-assessment tool

Welcome to the ARDC FAIR Data self-assessment tool. Using this tool you will be able to assess the 'FAIRness' of a dataset and determine how to enhance its FAIRness (where applicable).

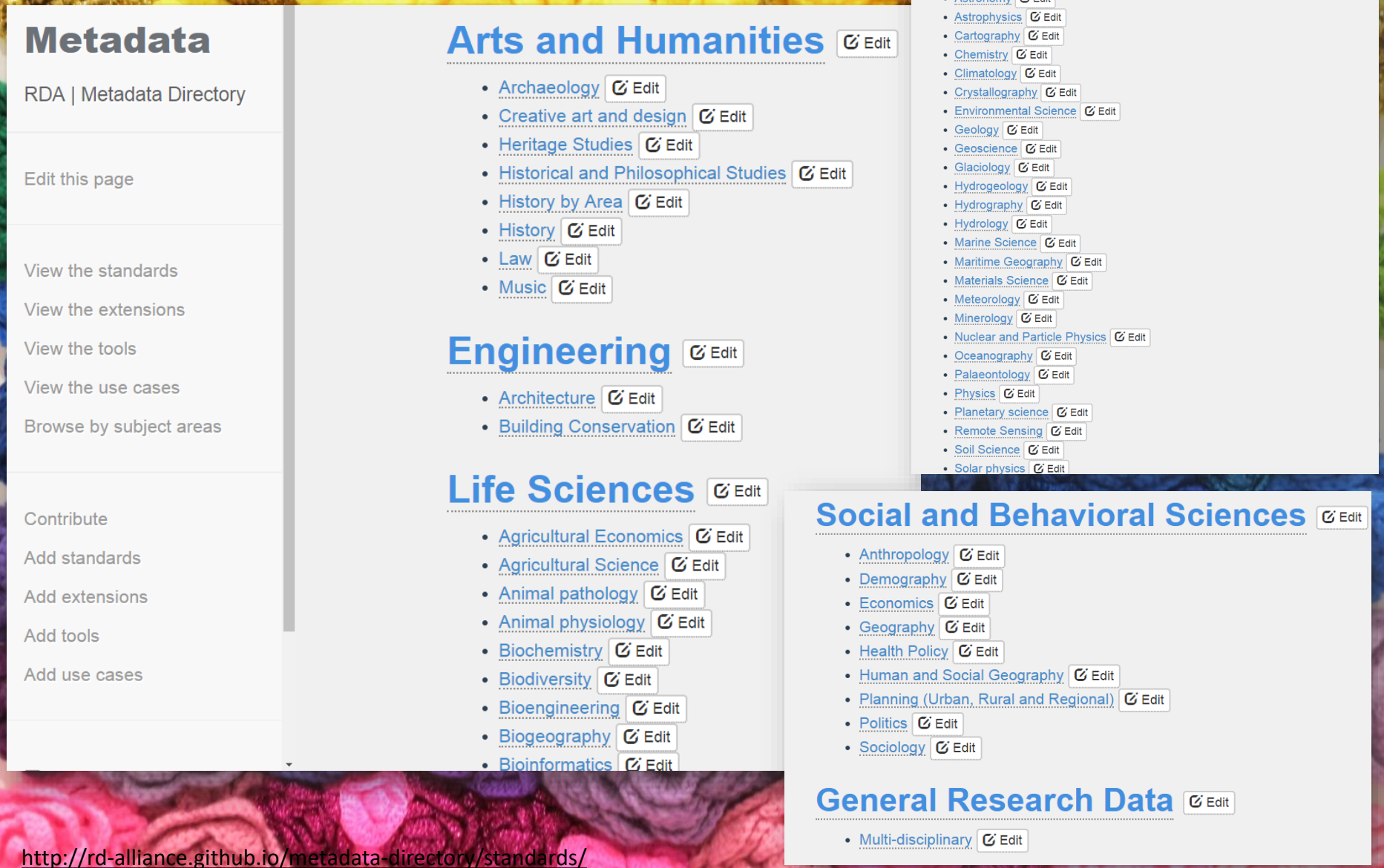
F = Findable: Metadata



Types of metadata	Goal	Example
Descriptive metadata	*The minimal metdata, required to find a digital object.* *If there are additional contextual metadata, a user will have a better idea on how to use the data*	*Author, title, abstract, date* *Contextuele metadata are for example location, time, data collection method (tools)*
Structural metadata	*These link the individual objects of a unity*	*Links to related digital objects, (e.g. the article written based on the linked research data)*
Technical metadata	*Information on the technical aspects of the data set*	*Data format, hardware/software used, calibration, version, authentication, encryption, metadata standard*
Administrative metadata	*Metadata focusing on user rights and management of digital objects*	*License, possible reasons for an embargo, waivers*

data

F = Findable. Metadata standards



Metadata

RDA | Metadata Directory

Edit this page

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

Contribute
Add standards
Add extensions
Add tools
Add use cases

Arts and Humanities Edit

- Archaeology Edit
- Creative art and design Edit
- Heritage Studies Edit
- Historical and Philosophical Studies Edit
- History by Area Edit
- History Edit
- Law Edit
- Music Edit

Engineering Edit

- Architecture Edit
- Building Conservation Edit

Life Sciences Edit

- Agricultural Economics Edit
- Agricultural Science Edit
- Animal pathology Edit
- Animal physiology Edit
- Biochemistry Edit
- Biodiversity Edit
- Bioengineering Edit
- Biogeography Edit
- Bioinformatics Edit

Physical Sciences & Mathematics Edit

- Astronomy Edit
- Astrophysics Edit
- Cartography Edit
- Chemistry Edit
- Climatology Edit
- Crystallography Edit
- Environmental Science Edit
- Geology Edit
- Geoscience Edit
- Glaciology Edit
- Hydrogeology Edit
- Hydrography Edit
- Hydrology Edit
- Marine Science Edit
- Maritime Geography Edit
- Materials Science Edit
- Meteorology Edit
- Minerology Edit
- Nuclear and Particle Physics Edit
- Oceanography Edit
- Palaeontology Edit
- Physics Edit
- Planetary science Edit
- Remote Sensing Edit
- Soil Science Edit
- Solar physics Edit

Social and Behavioral Sciences Edit

- Anthropology Edit
- Demography Edit
- Economics Edit
- Geography Edit
- Health Policy Edit
- Human and Social Geography Edit
- Planning (Urban, Rural and Regional) Edit
- Politics Edit
- Sociology Edit

General Research Data Edit

- Multi-disciplinary Edit

http://rd-alliance.github.io/metadata-directory/standards/

F = findable. Metadata tools

What CEDAR does

https://metadatacenter.org/

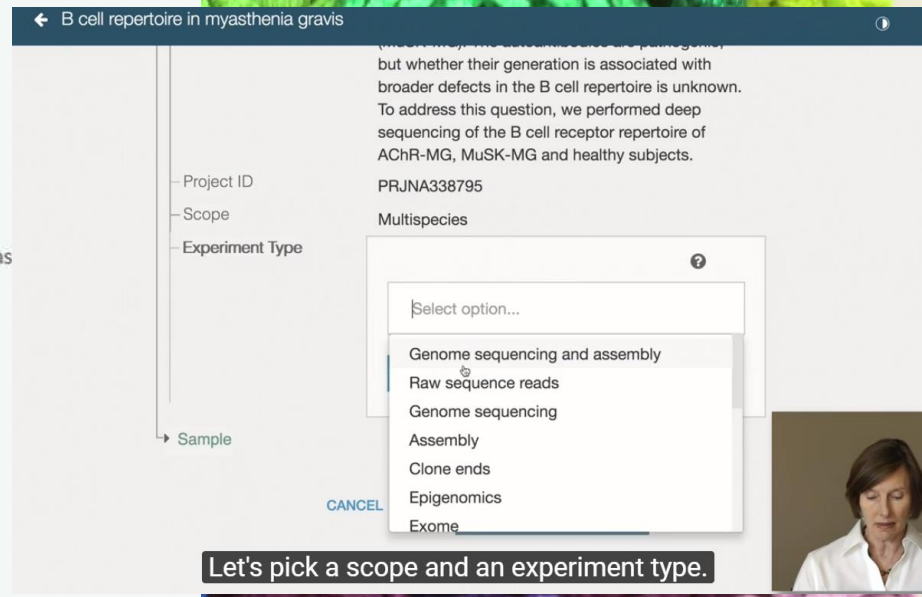
The CEDAR Workbench, as we refer to the suite of CEDAR tools, makes it easy to collect and use metadata. Eventually our tools wil
metadata record is created to its eventual processing, and even enhancement, by users and analysts. But for now, CEDAR tools he
to users, and download the information that users have provided.

What can CEDAR do for me already?

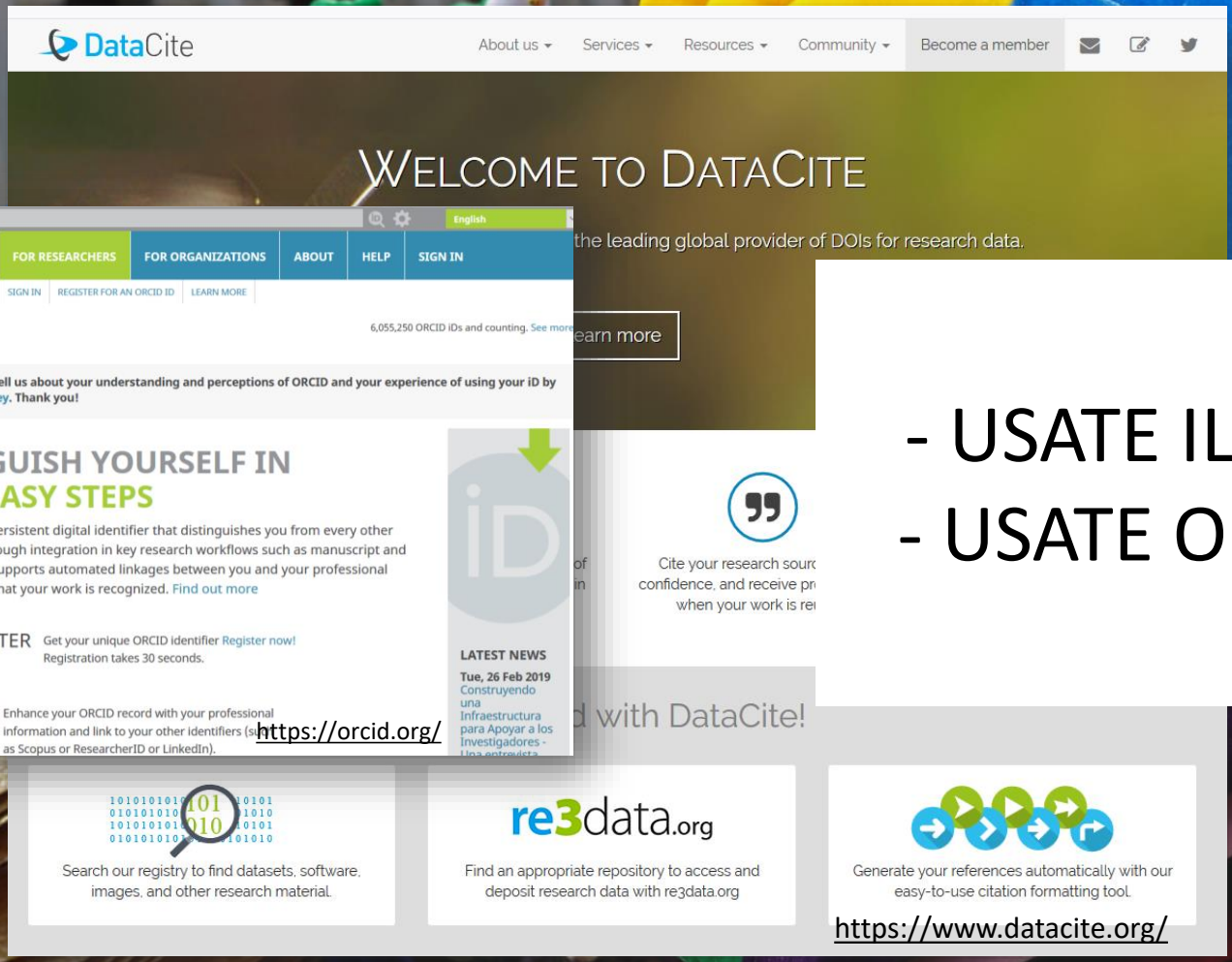
As of its production release, in February 2017, CEDAR addresses these scenarios:

- create user-friendly, shareable forms for collecting metadata, with features like
  - nested and repeatable elements and fields
  - reusable elements
  - control over tool tips, field titles, and field descriptions
- share your forms and metadata
  - provide a link to your metadata editors, so they can enter metadata responses based on your forms
  - share your forms and other content with individuals or a group
  - create and manage groups to make permissions simpler
- associate your questions (fields) and possible answers (values) with controlled terms
  - select any term or collection of terms from the NCBO BioPortal semantic repository
  - combine different terms from different controlled vocabularies into a single set of options
  - create your own terms, or term lists ('value sets') that can be re-used
- view responses meeting your (simple) search critieria, in several forms
  - CEDAR Metadata Editor's metadata view
  - an in-line JSON-LD format, used by CEDAR for all its metadata instances
  - download of JSON-LD files via the CEDAR REST API, for offline integration with your workflow
- use the Workbench Desktop interface to manage your content
  - use My Workspace to see your items, or Shared with Me to see other items you can access
  - select an item and control-click or use the 3-dot menu in the upper right to share it, copy it, delete it, or get info on it
- enable intelligent metadata suggestions in your template by using a field's Suggestions tab
  - CEDAR keeps track of metadata entered for that field
  - users will see a drop down list of the most popular metadata entries, and can select from them
- remotely access CEDAR content and capabilities using the CEDAR REST API

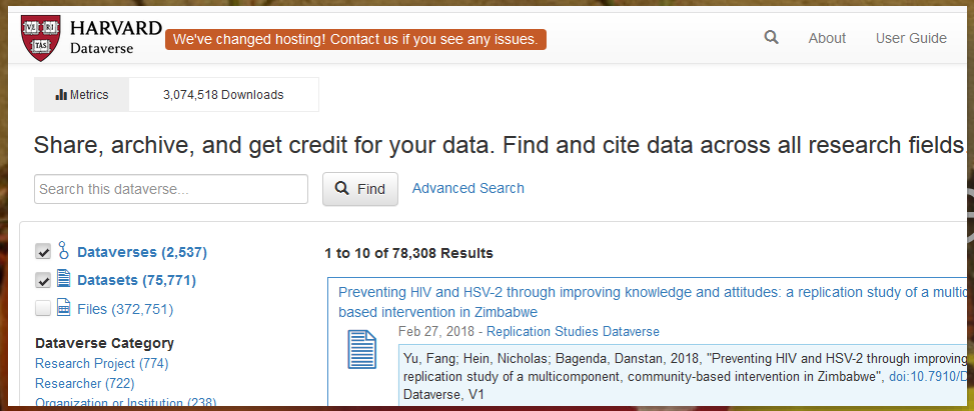
With these capabilities, you can capture simple or rich metadata for your project, build a repository of project metadata, or design
particular needs. Advanced users can even submit metadata entries through CEDAR's REST API.

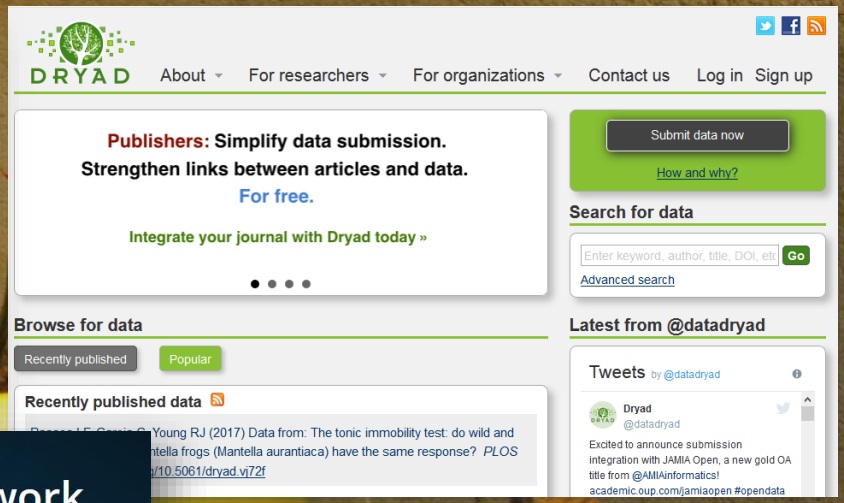


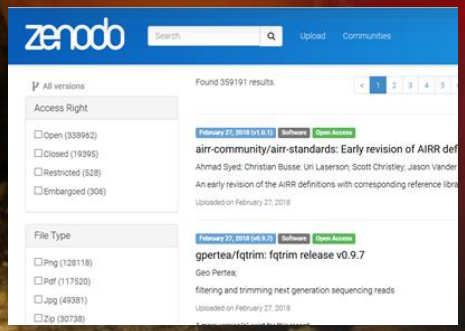
F = Findable. Persistent identifiers

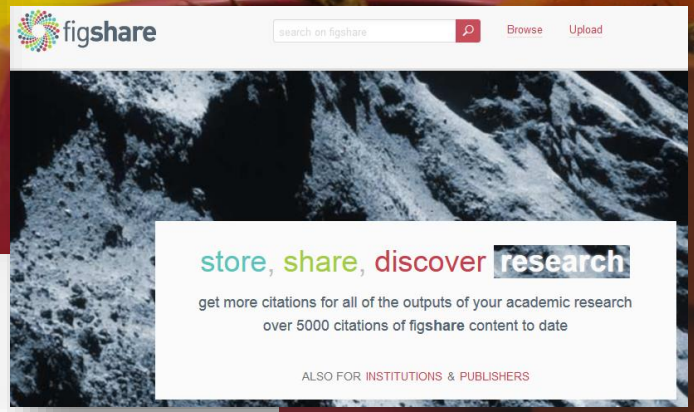


- USATE IL DOI
- USATE ORCID







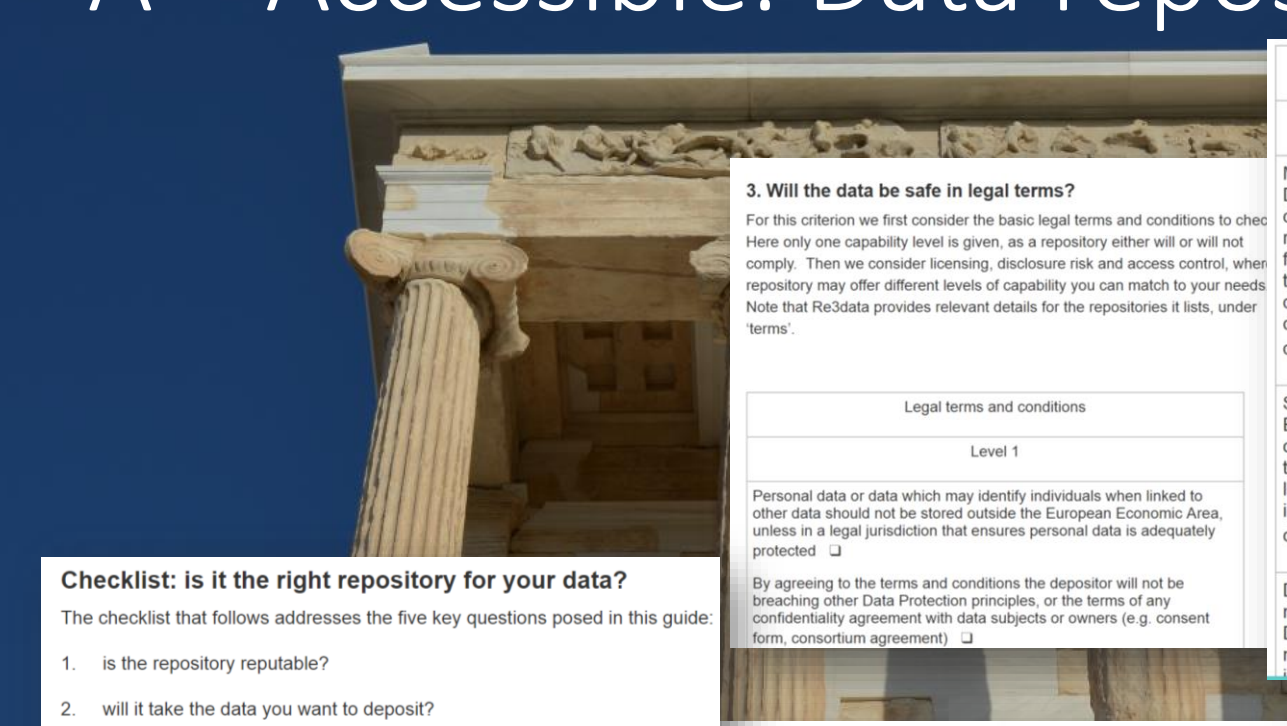


General depositories for research data

The following depositories are of interest to researchers in all domains:

- Zenodo (not-for-profit, hosted by CERN): https://zenodo.org:
- Dryad (not-for-profit membership organisation): http://www.datadryad.org
- Figshare (free service provided by private company): https://figshare.com
- Open Science Framework (not-for-profit, developed and maintained by the Center for Open Science[1]): https://osf.io
- Harvard Dataverse (not-for-profit, hosted by the Institute for Quantitative Social Studies IQSS at Harvard University): https://dataverse.harvard.edu

A = Accessible. Data repositories



3. Will the data be safe in legal terms?

For this criterion we first consider the basic legal terms and conditions to chec
Here only one capability level is given, as a repository either will or will not comply. Then we consider licensing, disclosure risk and access control, when repository may offer different levels of capability you can match to your needs
Note that Re3data provides relevant details for the repositories it lists, under 'terms'.

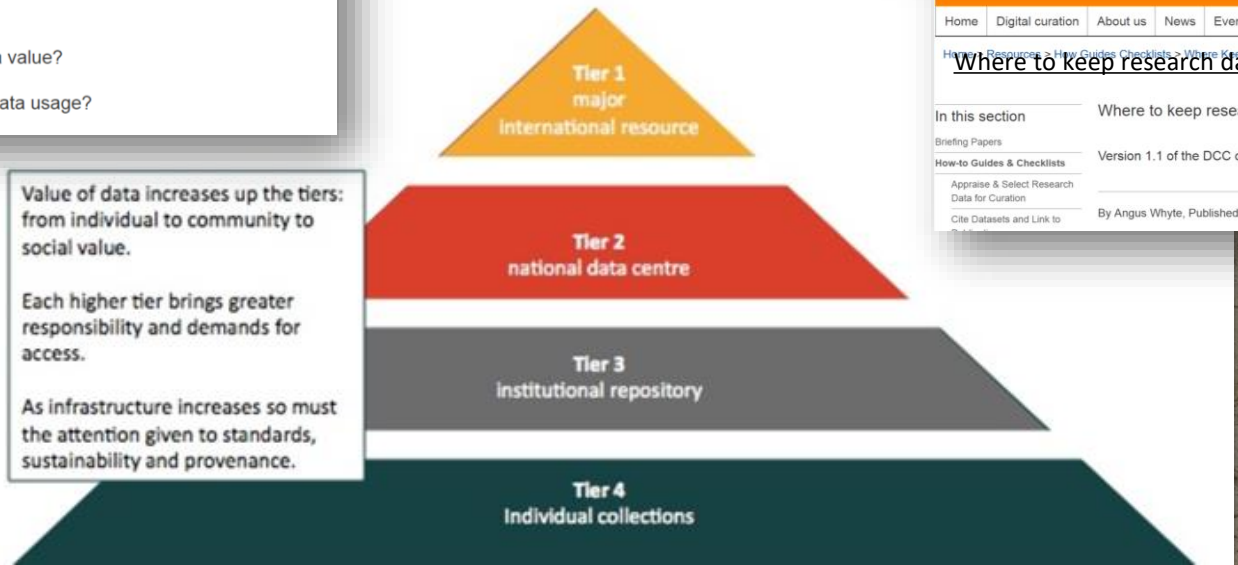
Legal terms and conditions
Level 1
Personal data or data which may identify individuals when linked to other data should not be stored outside the European Economic Area, unless in a legal jurisdiction that ensures personal data is adequately protected ❑
By agreeing to the terms and conditions the depositor will not be breaching other Data Protection principles, or the terms of any confidentiality agreement with data subjects or owners (e.g. consent form, consortium agreement) ❑

Findable, accessible and interoperable		
Level 1	Level 2	Level 3
Metadata publishing: Data collections are catalogued in a repository according to funder expectations so that they are discoverable by title, creator, and date of deposition ❑	Repository publishes other pertinent information as metadata fields to enhance cross-disciplinary discovery ❑	Metadata is catalogued to enhance reuse according to sector-leading standards, or to fulfil domain-specific purposes ❑
Stable identifiers: Enables a DOI or other open standard identifier to be assigned to a landing page for each ingested dataset/ collection ❑	Supports assignment of related persistent IDs per dataset/ collection ❑	Supports assignment of multiple persistent IDs at different levels of granularity within dataset/ collection ❑
Discovery metadata:Provides Datacite mandatory metadata and exposes	Provides metadata elements to enable broader discovery (e.g. geo-spatial) to	Exposes discovery metadata as Linked Open Data to optimise automatic

Checklist: is it the right repository for your data?

The checklist that follows addresses the five key questions posed in this guide:

1. is the repository reputable?
2. will it take the data you want to deposit?
3. will it be safe in legal terms?
4. will the repository sustain the data value?
5. will it support analysis and track data usage?

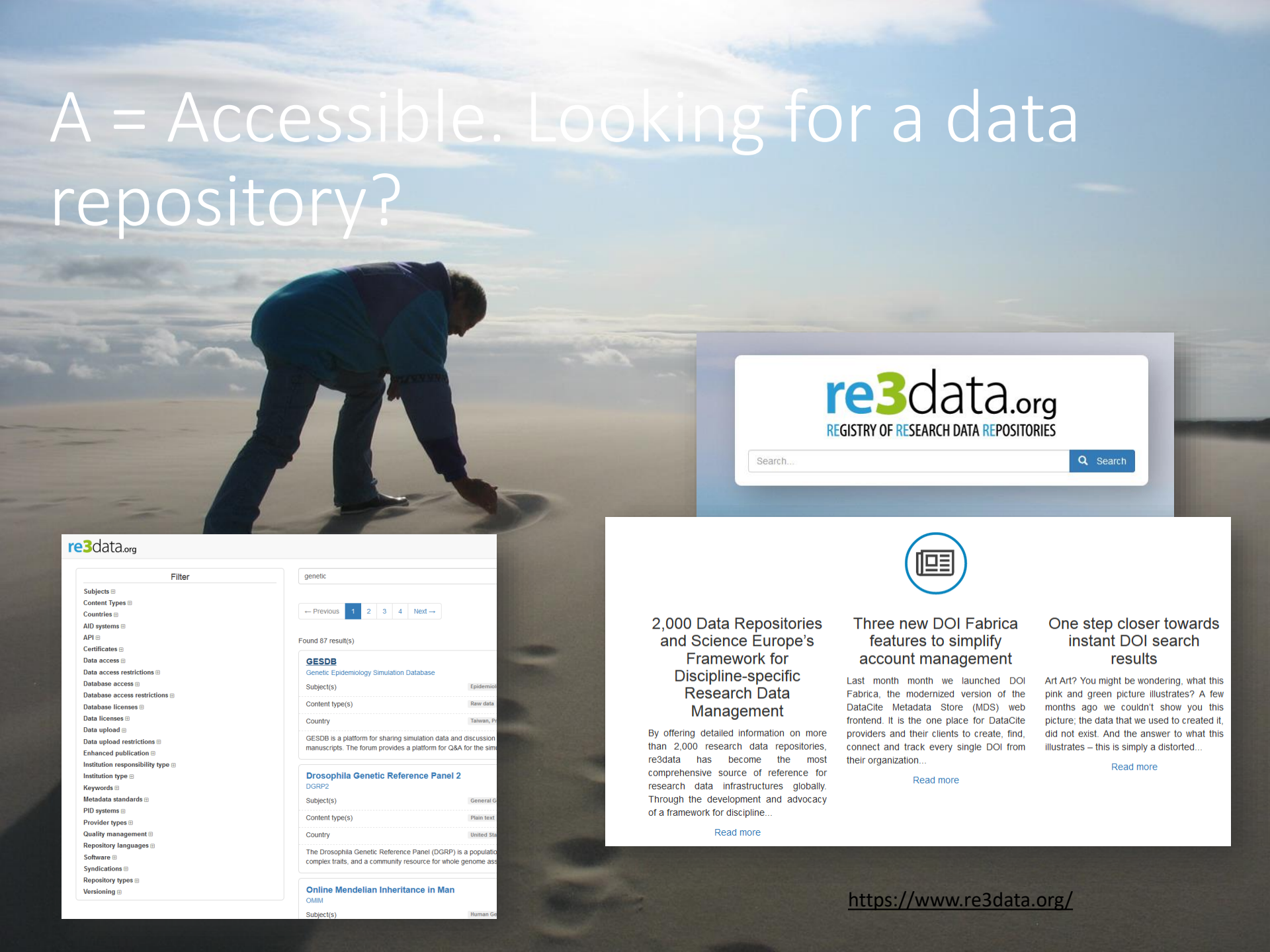


Value of data increases up the tiers: from individual to community to social value.

Each higher tier brings greater responsibility and demands for access.

As infrastructure increases so must the attention given to standards, sustainability and provenance.

A = Accessible. Looking for a data repository?



re3data.org

REGISTRY OF RESEARCH DATA REPOSITORIES

Search... Search

re3data.org

Filter

Subjects
Content Types
Countries
AID systems
API
Certificates
Data access
Data access restrictions
Database access
Database access restrictions
Database licenses
Data licenses
Data upload
Data upload restrictions
Enhanced publication
Institution responsibility type
Institution type
Keywords
Metadata standards
PID systems
Provider types
Quality management
Repository languages
Software
Syndications
Repository types
Versioning

genetic

← Previous 1 2 3 4 Next →

Found 87 result(s)

GESDB
Genetic Epidemiology Simulation Database
Subject(s)
Content type(s) Raw data
Country Taiwan, P
GESDB is a platform for sharing simulation data and discussion manuscripts. The forum provides a platform for Q&A for the simu

Drosophila Genetic Reference Panel 2
DGRP2
Subject(s) General G
Content type(s) Plain text
Country United Sta
The Drosophila Genetic Reference Panel (DGRP) is a populatio complex traits, and a community resource for whole genome ass

Online Mendelian Inheritance in Man
OMIM
Subject(s) Human Ge

2,000 Data Repositories and Science Europe's Framework for Discipline-specific Research Data Management

By offering detailed information on more than 2,000 research data repositories, re3data has become the most comprehensive source of reference for research data infrastructures globally. Through the development and advocacy of a framework for discipline...

Read more

Three new DOI Fabrica features to simplify account management

Last month month we launched DOI Fabrica, the modernized version of the DataCite Metadata Store (MDS) web frontend. It is the one place for DataCite providers and their clients to create, find, connect and track every single DOI from their organization...

Read more

One step closer towards instant DOI search results

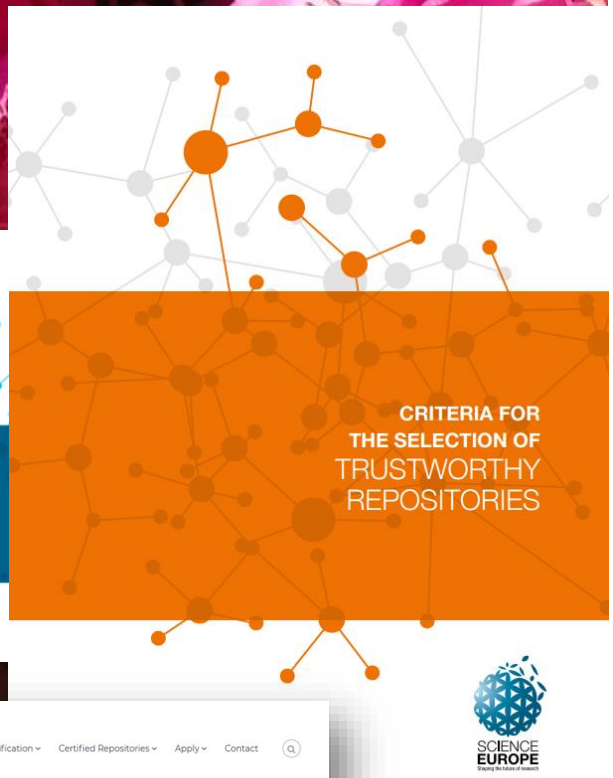
Art Art? You might be wondering, what this pink and green picture illustrates? A few months ago we couldn't show you this picture; the data that we used to created it, did not exist. And the answer to what this illustrates – this is simply a distorted...

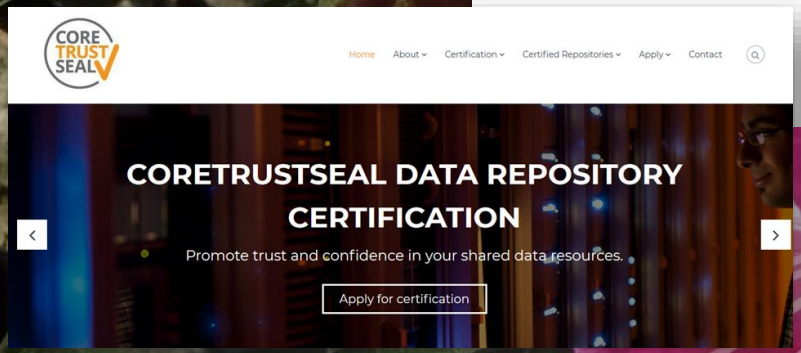
Read more

https://www.re3data.org/

[trusted repositories]

Dec. 2018





CRITERIA FOR THE SELECTION OF TRUSTWORTHY REPOSITORIES

Trustworthy repositories should meet the following minimum criteria:

1. **Provision of Persistent and Unique Identifiers (PIDs)**
   a. Allow data discovery and identification
   b. Enable searching, citing, and retrieval of data
   c. Provide support for data versioning

2. **Metadata**
   a. Enable finding of data
   b. Enable referencing to related relevant information, such as other data and publications
   c. Provide information that is publicly available and maintained, even for non-published, protected, retracted, or deleted data
   d. Use metadata standards that are broadly accepted (by the scientific community)
   e. Ensure that metadata are machine-retrievable

3. **Data access and usage licences**
   a. Enable access to data under well-specified conditions
   b. Ensure data authenticity and integrity
   c. Enable retrieval of data
   d. Provide information about licensing and permissions (in ideally machine-readable form)
   e. Ensure confidentiality and respect rights of data subjects and creators

4. **Preservation**
   a. Ensure persistence of metadata and data
   b. Be transparent about mission, scope, preservation policies, and plans (including governance, financial sustainability, retention period, and continuity plan)

A = Accessible. Data journals

Data Journals

Hier entsteht eine Liste von Data Journals, die vorwiegend Data Papers

- Atomic Data and Nuclear Data Tables (Elsevier)
- Biodiversity Data Journal (Pensoft Publishers)
- Biomedical Data Journal (Procon Ltd.)
- BMC Research Notes (Biomed Central)
- Chemical Data Collections (Elsevier)
- Data (MDPI)
- Data in Brief (Elsevier)
- Dataset Papers in Science (Hindawi Publishing Corporation)
- Earth System Science Data - ESSD (Copernicus Publications)
- Ecological Archives (Ecological Society of America - ESA)
- European Data Watch (European Data Watch)
- F1000Research (F1000 Research)
- Genomics Data (Elsevier)
- Geoscience Data Journal (Wiley)
- GigaScience (BioMed Central)
- Internet Archaeology (Internet Archaeology)
- Journal of Open Psychology Data (JOPD) (Ubiquity Press)
- Journal of Chemical & Engineering Data (ACS Publications)
- Journal of Physical and Chemical Data (AIP Publishing)
- Nuclear Data Sheets (Elsevier)
- Open Archaeology Data (Ubiquity Press)
- Open Data Journal for Agricultural Research (diverse)
- Open Health Data (Ubiquity Press)
- Open Journal of Bioresources (Ubiquity Press)
- Open Network Biology (BioMed Central)
- Research Data Journal for the Humanities and Social Sciences (Brill)
- Scientific Data (Nature Publishing Group)

Dataset Description

Object Name

- *walkers* – three files providing the data, metadata and field type definitions (.csv, .txt, .csvt respectively) for records made by individual walkers during stage-one fieldwalking.
- *counts* – three files providing the data, metadata and field type definitions (.csv, .txt, .csvt respectively) for pot-sherds countedduring stage-one fieldwalking.
- *pottery* – three files providing the data, metadata and field type definitions (.csv, .txt, .csvt respectively) for the main pottery database, assembled various artefact specialists.
- *petrography* – three files providing the data, metadata and field type definitions (.csv, .txt, .csvt respectively) for those sherds sampled for thin section petrography.
- *lithics* – three files providing the data, metadata and field type definitions (.csv, .txt, .csvt respectively) for the main lithics database.
- *other* – three files providing the data, metadata and field type definitions (.csv, .txt, .csvt respectively) for the main database of all non-ceramic and non-lithic finds.
- *structs* – three files providing the data, metadata and field type definitions (.csv, .txt, .csvt respectively) for the main database of all standing remains, except for terraces.
- *coast* – a vector polygon dataset (.shp and associated files) with the shape of Antkythera's coastline.
- *geology* –a vector polygon dataset (.shp and associated files) with the main bedrock units on Antkythera.
- *tracts* – a vector polygon dataset (.shp and associated files) with the main stage-one survey units.
- *grids* – a vector polygon dataset (.shp and associated files) with the main stage-two survey units.
- *terraces* – vector line dataset (.shp and associated files) with all observable agricultural terraces (i.e. the location

- *other* – primarily Andrew Bevan (UCL), with further assistance from James Conolly (Trent)
- *geology* – a combination of fieldwork by Ruth Siddall (UCL) and remote sensing by Andrew Bevan (UCL)

Repository Location
UK Archaeology Data Service Collection 1115 (doi: 10.5284/1012484)

Publication Date
05/02/2012

Language
English (a Greek language summary of the project methods and results can be found at www.ucl.ac.uk/asp/ or www.tuarc.trentu.ca/asp/).

License
Creative Commons CC-BY 3.0

Reuse Potential

Due to their unusual coverage of an entire landscape, these datasets would provided a good basis for developing a tutorial on survey, GIS and/or spatial analysis in archaeology. They also lend themselves to the comparative analysis of evidence from other intensive Mediterranean surveys that are in the public domain (e.g. http://dx.doi.org/ org/10.5284/100 dx.doi.org/10.528 to the fact that th cal. The ASP data locations, dates ar ally in the databas structures and ter

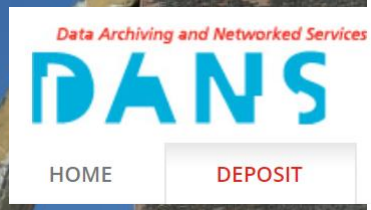
Data journals

Panayiota Polydoratou

Alexander Technological Educational Institute of Thessaloniki

European Commission Workshop
Alternative Open Access Publishing Models: Exploring New Territories in Communication
Brussels, 12 October 2015

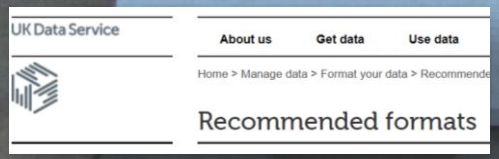
A = Accessible. Formati



Type	Preferred format(s)	Non-preferred format(s)
Text documents	• PDF/A (.pdf)	• ODT (.odt) • MS Word (.doc, .docx) • RTF (.rtf) • PDF (.pdf)
Plain text	• Unicode text (.txt)	• Non-Unicode text (.txt)
Markup language	• XML (.xml) • HTML (.html) • Related files: .css, .xslt, .js, .es	• SGML (.sgml)
Spreadsheets	• ODS (.ods) • CSV (.csv)	• MS Excel (.xls, .xlsx) • PDF/A (.pdf) • OOXML (.docx, .docm)
Databases	• SQL (.sql) • SIARD (.siard) • DB tables (.csv)	• MS Access (.mdb, .accdb) (v. 2000 or later) • dBase (.dbf) • HDF5 (.hdf5, .he5, .h5)
Statistical data	• SPSS Portable (.por) • SPSS (.sav) • STATA (.dta) • DDI (.xml) • data (.csv) + setup (.txt)	• SAS (.7dat; .sd2; .tpt) • R (* under examination)
Raster images	• JPEG (.jpg, .jpeg) • TIFF (.tif, .tiff) • PNG (.png) • JPEG 2000 (.jp2)	• DICOM (.dcm) (by mutual agreement)

https://dans.knaw.nl/en/deposit/information-about-depositing-data/before-depositing/file-formats

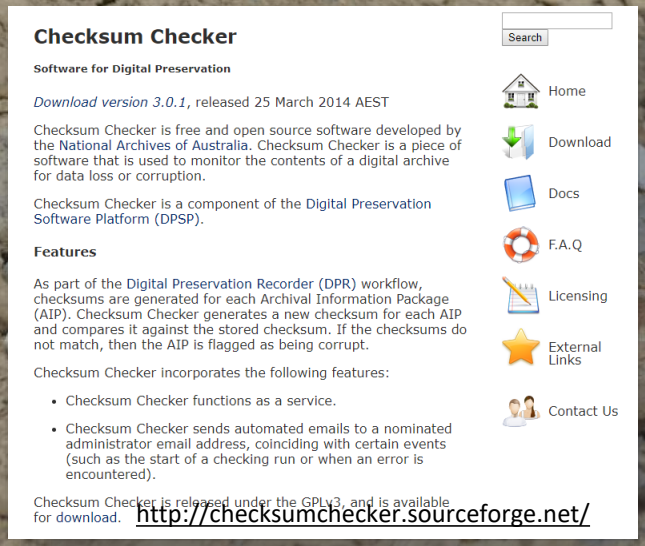
Type of data	Recommended formats	Acceptable formats
Tabular data with extensive metadata variable labels, code labels, and defined missing values	SPSS portable format (.por) delimited text and command ('setup') file (SPSS, Stata, SAS, etc.) structured text or mark-up file of metadata information, e.g. DDI XML file	proprietary formats of statistical packages: SPSS (.sav), Stata (.dta), MS Access (.mdb/.accdb)
Tabular data with minimal metadata column headings, variable names	comma-separated values (.csv) tab-delimited file (.tab) delimited text with SQL data definition statements	delimited text (.txt) with characters not present in data used as delimiters widely-used formats: MS Excel (.xls/.xlsx), MS Access (.mdb/.accdb), dBase (.dbf), OpenDocument Spreadsheet (.ods)
Geospatial data vector and raster data	ESRI Shapefile (.shp, .shx, .dbf, .prj, .sbx, .sbn optional) geo-referenced TIFF (.tif, .tfw) CAD data (.dwg) tabular GIS attribute data Geography Markup Language (.gml)	ESRI Geodatabase format (.mdb) MapInfo Interchange Format (.mif) for vector data Keyhole Mark-up Language (.kml) Adobe Illustrator (.ai), CAD data (.dxf or .svg) binary formats of GIS and CAD packages
Textual data	Rich Text Format (.rtf) plain text, ASCII (.txt) eXtensible Mark-up Language (.xml) text according to an appropriate Document Type Definition (DTD) or schema	Hypertext Mark-up Language (.html) widely-used formats: MS Word (.doc/.docx) some software-specific formats: NUD*IST, NVivo and ATLAS.ti
Image data	TIFF 6.0 uncompressed (.tif)	JPEG (.jpeg, .jpg, .jp2) if original created in this format GIF (.gif) TIFF other versions (.tif, .tiff) RAW image format (.raw) Photoshop files (.psd) BMP (.bmp) PNG (.png) Adobe Portable Document Format (PDF/A, PDF) (.pdf)
Audio data	Free Lossless Audio Codec (FLAC) (.flac)	MPEG-1 Audio Layer 3 (.mp3) if original created in this format Audio Interchange File Format (.aif) Waveform Audio Format (.wav)
Video data	MPEG-4 (.mp4) OGG video (.ogv, .ogg) motion JPEG 2000 (.mj2)	AVCHD video (.avchd)
Documentation and scripts	Rich Text Format (.rtf) PDF/UA, PDF/A or PDF (.pdf) XHTML or HTML (.xhtml, .htm) OpenDocument Text (.odt)	plain text (.txt) widely-used formats: MS Word (.doc/.docx), MS Excel (.xls/.xlsx) XML marked-up text (.xml) according to an appropriate DTD or schema, e.g. XHMTL 1.0



https://www.ukdataservice.ac.uk/manage-data/format/recommended-formats.aspx

A = Accessible. Conservazione

LONG TERM
OR SHORT TERM?



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.

Search

Home

Download

Docs

F.A.Q

Licensing

External Links

Contact Us

http://checksumchecker.sourceforge.net/

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

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).

A = Accessible. Conservazione

Portable devices | Cloud storage | Local storage | Networked drives

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

Advantages	Disadvantages/Risks	Precaut (sensiti data
• 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.	• 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.	Use in c encrypt passwo

Advantages	Disadvantages/Risks	Precautions for (sensitive) personal data
• Automatic backups. • Often automatic version control.	• 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.	• 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

Different needs, different tools.
During: you need also to share with the team.

I = Interoperable

To speed up discovery and uncover new insights, research data should be easily combined with other datasets by humans as well as computer systems.

INTEROPERABLE

10 Establish well documented machine-actionable APIs

Well documented and machine-actionable APIs - a set of subroutine definitions, protocols, and tools for building application software - allow for automatic indexing, retrieval and combining of (meta)data from different data repositories.

Document APIs well and make it possible to deliver the schema of the (meta)data model. Consider showing examples of how to successfully mine data from different endpoints and combine them into new data sets usable for new research.

11 Use open well-defined vocabularies

The description of metadata elements should follow community guidelines that use open, well defined and well known vocabularies. Such vocabularies describe the exact meaning of the concepts and qualities that the data represent.

Use vocabularies relevant to your field, and enrich and structure your research output accordingly from the start of your research project.

Give examples of vocabularies the research community may use, based on research domain specifics.

12 Document metadata models

Clearly documenting metadata models helps developers to compare and make mappings between metadata.

Publish the metadata models in use in your research infrastructure. Document technical specifications and define classes (groups of things that have common properties) and properties (elements that express the attributes of a metadata section as well as the relationships between different parts of the metadata). For metadata mapping purposes, list the mandatory and recommended properties.

13 Prescribe and use interoperable data standards

Using a data standard backed up by a strong community, increases the possibility to share, reuse and combine data collections.

Check with the repository where you want to deposit your data what data standards they use. Structure your data collection in this format from the start of your research project.

Clearly specify which data standard your institution uses, pool a community around them and maintain them especially with a perspective on interoperability. Good examples are CMDI (language studies) and the SIKB0102 Standard (archaeology).

14 Establish processes to enhance data quality

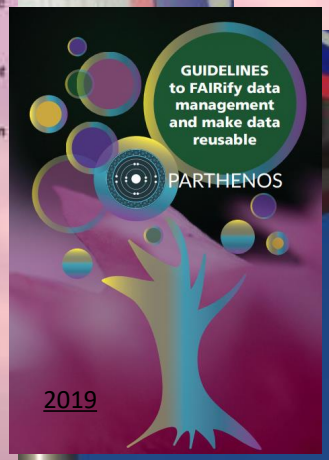
To boost (meta)data quality and, therefore, interoperability, establish (automatic) processes that clean up, derive and enrich (meta)data.

Establish procedures to minimise the risk of mistakes in collecting data. E.g. choose a date from a calendar instead of filling it in by hand.

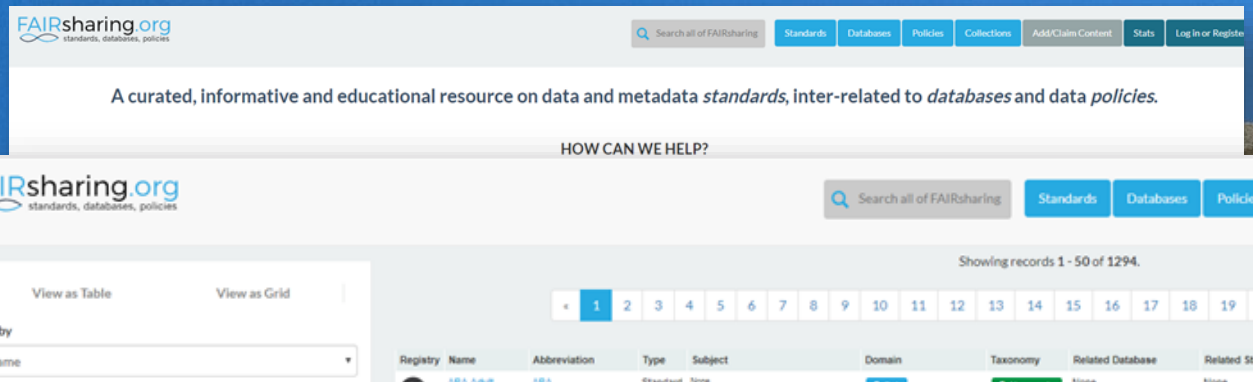
Invest in tools to help clean up (meta)data and to convert data into standardised and interoperable data formats. Combine efforts to develop workflows and software solutions for such automatic processes, e.g. by using machine learning tools.

15 Prescribe and use future-proof file formats

All data files held in a data repository should be in an open, international, standardised file format to ensure long-term interoperability in terms of usability, accessibility and sustainability.



[FAIRsharing / Interoperability]



https://fairsharing.org

R = Reusable. Documentazione

cessda TRAINING

Project-level documentation

Project-level documentation explains the aims of the study, what the research questions/hypotheses are, what methodologies were being used, what instruments and measures were being used, etc. In the accordion the questions which your project-level documentation should answer are stated in more detail:

- ⊕ 1. For what purpose was data created
- ⊕ 2. What does the dataset contain
- ⊕ 3. How was data collected
- ⊕ 4. Who collected the data and when
- ⊕ 5. How was the data processed
- ⊕ 6. What possible manipulations were done to the data
- ⊕ 7. What were the quality assurance procedures
- ⊕ 8. How can data be accessed

Data-level documentation

Data-level or object-level documentation provides information at the level of individual objects such as pictures or interview transcripts or variables in a database. You can embed data-level information in data files. For example, in interviews, it is best to write down the contextual and descriptive information about each interview at the beginning of each file. And for quantitative data variable and value names can be embedded within the data file itself.

⊖ Quantitative data

Variable-level annotation should be embedded within a data file itself. If you need to compile an extensive variable level documentation that can be created by using a structured metadata format.

Data-level documentation for quantitative data

For quantitative data document the following:

- **Information about the data file**
  Data type, file type and format, size, data processing scripts.
- **Information about the variables in the file**
  The names, labels and descriptions of variables, their values, a description of derived

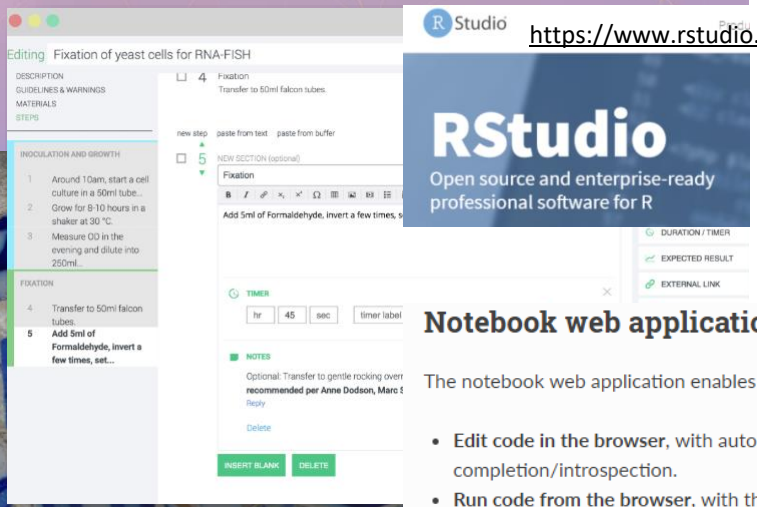
https://www.cessda.eu/Training/Training-Resources/Library/Data-Management-Expert-Guide/2.-Organise-Document/Documentation-and-metadata

R = Reusable. Documentazione

protocols.io

Make your science more reproducible

protocols.io is the #1 open access repository for science methods



RStudio

https://www.rstudio.com/

RStudio

Open source and enterprise-ready professional software for R

http://jupyter.org/index.h

What is an Open Notebook?

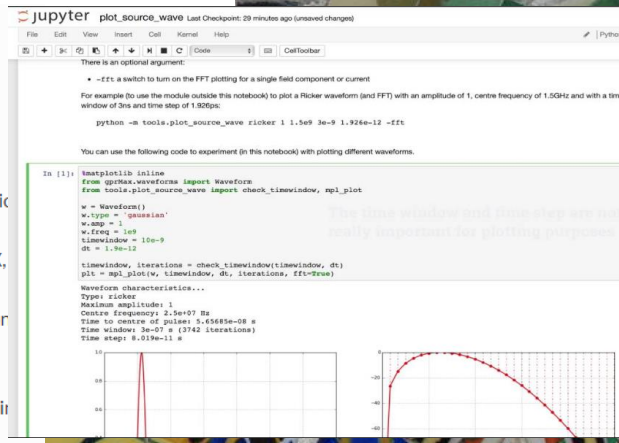
Open Notebooks are documents that contain equations, visualisations, narrative text and live code that can be executed independently and interactively, with output visible immediately beneath the input.

They bring together analysis descriptions and results, which can be executed to perform the data analysis in real time.

Notebook web application

The notebook web application enables users to:

- **Edit code in the browser**, with automatic syntax highlighting, indentation, and tab completion/introspection.
- **Run code from the browser**, with the results of computations attached to the code whic generated them.
- See the results of computations with **rich media representations**, such as HTML, LaTeX, SVG, PDF, etc.
- Create and use **interactive JavaScript widgets**, which bind interactive user interface cor and visualizations to reactive kernel side computations.
- Author **narrative text** using the Markdown markup language.
- Include mathematical equations using **LaTeX syntax in Markdown**, which are rendered i browser by MathJax.



…why not?

- protocols.io per depositare i metodi
- Open Lab Notebook per scrivere tutto insieme [time consuming la prima volta, poi…]

R= Reusable. Licenze

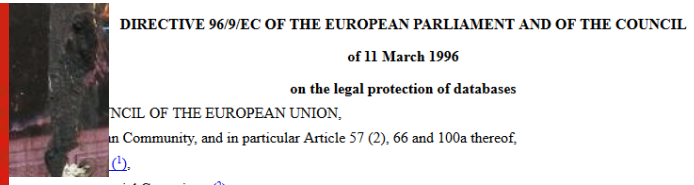
Copyright: protects the STRUCTURE, selection or arrangement of their contents" (Art. 3) NOT THE DATA

Sui generis database right: protects the «substantial effort» in OBTAINING data [NOT «CREATING»]… the right owner often is the institution

Database=a collection of independent works, data or other materials arranged in a systematic or methodical way (Art.1)

KEEP CALM AND

RICORDA: NESSUN COPYRIGHT SUI DATI (NON CREATIVI)


DIRECTIVE 96/9/EC OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL

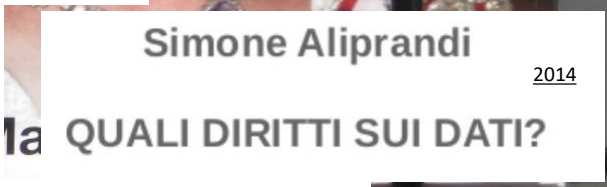
of 11 March 1996

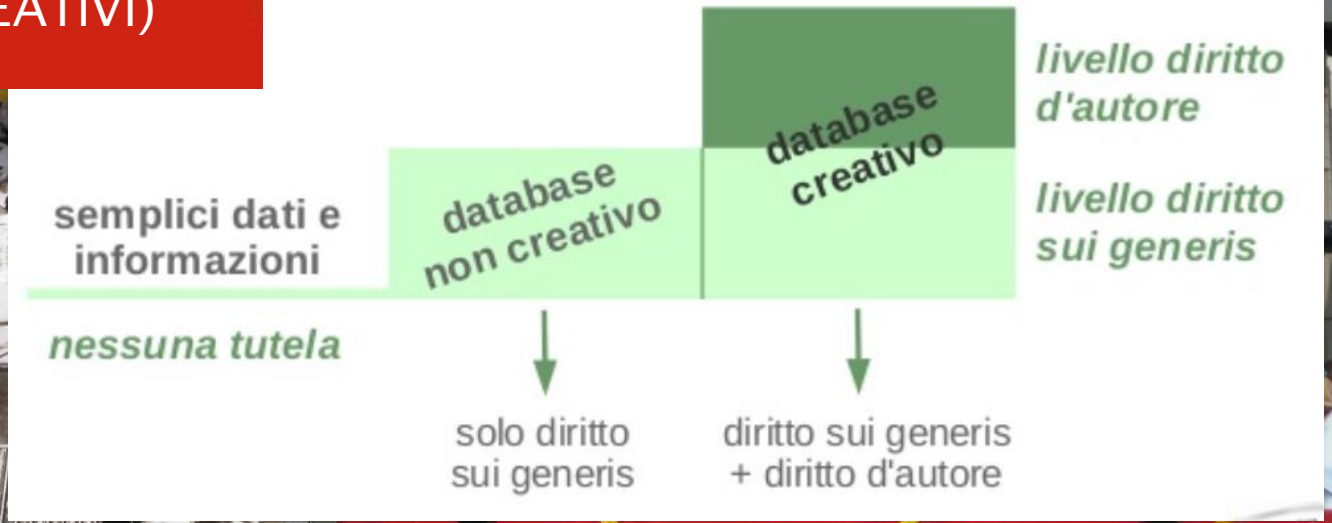
on the legal protection of databases

NCIL OF THE EUROPEAN UNION,

n Community, and in particular Article 57 (2), 66 and 100a thereof,

(1),





[non suoniamo tutti la stessa musica]

Obstacles to the trans-European archiving and sharing of research data

Making research data as openly available as possible is a widely recognised goal. For researchers working on an interdisciplinary project involving several countries, it can be difficult to fully comprehend in which ways open access to research data can be legally obtained. European national laws still diverge.

A report from Knowledge Exchange (Knowledege Exchange, 2011) concludes that it will remain difficult to predict when particular files of research data are protected because of:

- **Diversity in copyright protection**
  Even though most research data will fail to meet the criteria for copyright protection because they are not likely to be considered as "works" (they mainly concern facts), the lack of harmonisation of the criteria for copyright protection in Europe is tricky. E.g., whereas Germany, Denmark and the Netherlands have a relatively similar (higher) originality standard, the UK has a very low standard (skill, judgment and labour) making it

- **Diversity in copyright owner**
  If protection applies, the right holder's consent is required for sharing the data. However, the designation of the copyright owner is also different in different jurisdictions. Althoug in many cases the maker of the work will be considered to be the author and therefore th right holder, only Dutch and UK law designate the employer as the right holder if the wor was made in the course of employment.

CESSDA guide

⊕ Czech Republic

⊕ Finland

⊕ Switzerland

⊕ UK

[regole chiare dall'inizio]

Safe to be open 2016

… FISSATE REGOLE CHIARE DA SUBITO

- CHI HA I DIRITTI SUI DATI (SE ESISTE)
- CHI DEVE PROVVEDERE A CONSERVARE
- CHI PUÒ SFRUTTARE

R = Reusable. Licenze

DCC because good research needs good data

How to License Research Data

This guide will help you decide how to apply a licence to your research data, and which licence would be most suitable. It should provide you with an awareness of why licensing data is important, the impact licences have on future research, and the potential pitfalls to avoid. It concentrates on the UK context, though some aspects apply internationally; it does not, however, provide legal advice. The guide should interest both the principal investigators and researchers responsible for the data, and those who provide access to them through a data centre, repository or archive.

http://www.dcc.ac.uk/resources/how-guides/license-research-data

R = Reusable. Licenze

DCC because good research needs good data

How to License Research Data

This guide will help you decide how to apply a licence to your research data, and which licence would be most suitable. It should provide you with an awareness of why licensing data is important, the impact licences have on future research, and the potential pitfalls to avoid. It concentrates on the UK context, though some aspects apply internationally; it does not, however, provide legal advice. The guide should interest both the principal investigators and researchers responsible for the data, and those who provide access to them through a data centre, repository or archive.

Creative Commons at a glance

Good for

- very simple, factual datasets
- data to be used automatically

Watch out for

- versions: use v. 4 or later
- attribution stacking
- the NC condition: only use with dual licensing
- the SA condition as it reduces interoperability
- the ND condition as it severely restricts reuse

ODC-By at a glance

Good for

- most databases and datasets
- data to be used automatically
- data to be used for generating non-data products

Watch out for

- attribution stacking

ODC-ODbL at a glance

Good for

- most databases and datasets
- data to be used automatically
- data to be used for generating non-data products

Watch out for

- attribution stacking
- the copyleft condition as it reduces interoperability
- the DRM clause as it may put off some reusers

Public domain at a glance

Good for

- most databases and datasets
- data to be used by anyone or any tool
- data to be used for any purpose

Watch out for

- lack of control over how database is reused
- lack of protection against unfair competition

Creative Com

creative commons UK

FACT SHEET ON CREATIVE COMMONS & OPEN SCIENCE V.0.1

This information guide contains questions and responses to common concerns surrounding open science and the implications of licensing data under Creative Commons licences. It is intended to aid researchers, teachers, librarians, administrators and many others using and encountering Creative Commons licences in their work.

https://doi.org/10.5281/zenodo.840651

What is Open Science?

Open Science is the movement to make scientific research and data accessible to all for knowledge dissemination and public reuse.

How should I licence my data for the purposes of Open Science?

We recommend you use the CC0 Public Domain Dedication, which is first and foremost a waiver, but can act as a licence when a waiver is not possible.

CC ZERO LICENCE, 'NO RIGHTS RESERVED' LOGO

PUBLIC DOMAIN MARK LOGO

By applying CC0 to your data you enable everyone to freely reuse your data as they see fit by waiving (giving up) your copyright and related rights in that data.

You should keep in mind that there are many situations in which data is *not* protected as a matter of law. Such data can include facts, names, numbers – things that are considered 'non-original' and part of the public domain thus not subject to copyright protections. Similarly, your database (which is a structured collection of data) might be considered 'non-original' and thus ineligible for copyright, and it might additionally be excluded from other forms of protection (like the EU *sui generis* database right, also known as the 'SGDR', for non-original databases).

In these cases, using a Creative Commons licence such as a CC BY could signal to users that you claim a copyright in the non-original data despite the law, and perhaps despite your real intention.

Finally, if your data is in the public domain worldwide, you might state simply and obviously on the material that no restrictions attach to the reuse of your data and apply a Public Domain Mark.

When in doubt, consider which use may be appropriate according to the chart below:

CC0 & PUBLIC DOMAIN LICENCES WHICH LICENSE TO USE AND WHEN

CC ZERO	PUBLIC DOMAIN
'Creative arrangement' of data is original, but any copyright has been waived and content is made available copyright-free	'Creative arrangement' of data is not original; the author acknowledges this and communicates the data is in the public domain

nmons e Op

But I would like attribution when others use my dataset. In that case, shouldn't I use a CC BY licence?

We recommend that you avoid using a CC BY licence. Here's why:

While attribution is a genuine, recognisable concern, not only might using a CC BY licence be legally unenforceable when no underlying copyright or SGDR protects the work, but it may also communicate the wrong message to the world. A better solution is to use CC0 and simply ask for credit (rather than require attribution), and provide a citation for the dataset that others can copy and paste with ease. Such requests are consistent with scholarly norms for citing source materials.

Legally speaking, datasets that are ***not*** subject to copyright or related rights (and are thus in the public domain) cannot be the object of a copyright licence. Despite this, agreements based in contract law may be enforceable. Creative Commons licences, however, are copyright licences. Therefore, where the conditions for a copyright or related right are not triggered, copyright licences, such as the CC BY licence, are unenforceable.

In some cases, however, rights may exist (like the *sui generis* database right previously mentioned), and permission for others to use your dataset may be legally required. These rights are meant to protect the maker's investment, rather than originality. As such, database rights do not include the moral right of attribution. So by using a CC BY licence, you signal to users that you restrict access to your dataset beyond the protections provided by the law. We are not saying that this cannot be done, we are just saying that if you choose to do this, you should make sure you fully understand what it entails.

USARE CC0

- CHIEDERE CHE VENGA DATO CREDITO ALL'AUTORE
- PROPORRE GIÀ LA CITAZIONE-TIPO

(non citare la fonte è scorretto scientificamente)

It sounds like you're really pushing for the use of CC0 for open science datasets.

Exactly. Data is only open if anyone is free to use, reuse, and distribute it. This means it must be made available for both commercial and non-commercial purposes under non-discriminatory conditions that allow for it to be modified.

When data is made available for all reuse, others can create new knowledge from combining it. This leads to the enrichment of open datasets and further dissemination of knowledge. Accordingly, CC0 is ideal for open science as it both protects and promotes the unrestricted circulation of data.

And remember, it's bad science not to cite the source of data you use. To help others cite your data include a citation that users can copy and paste to give you credit for your hard work.

cannot be done, we are just saying that if you choose to do this, you should make sure you fully understand what it entails.

I'm uncomfortable with others using my research for commercial purposes. Should I use a non-commercial licence for my dataset?

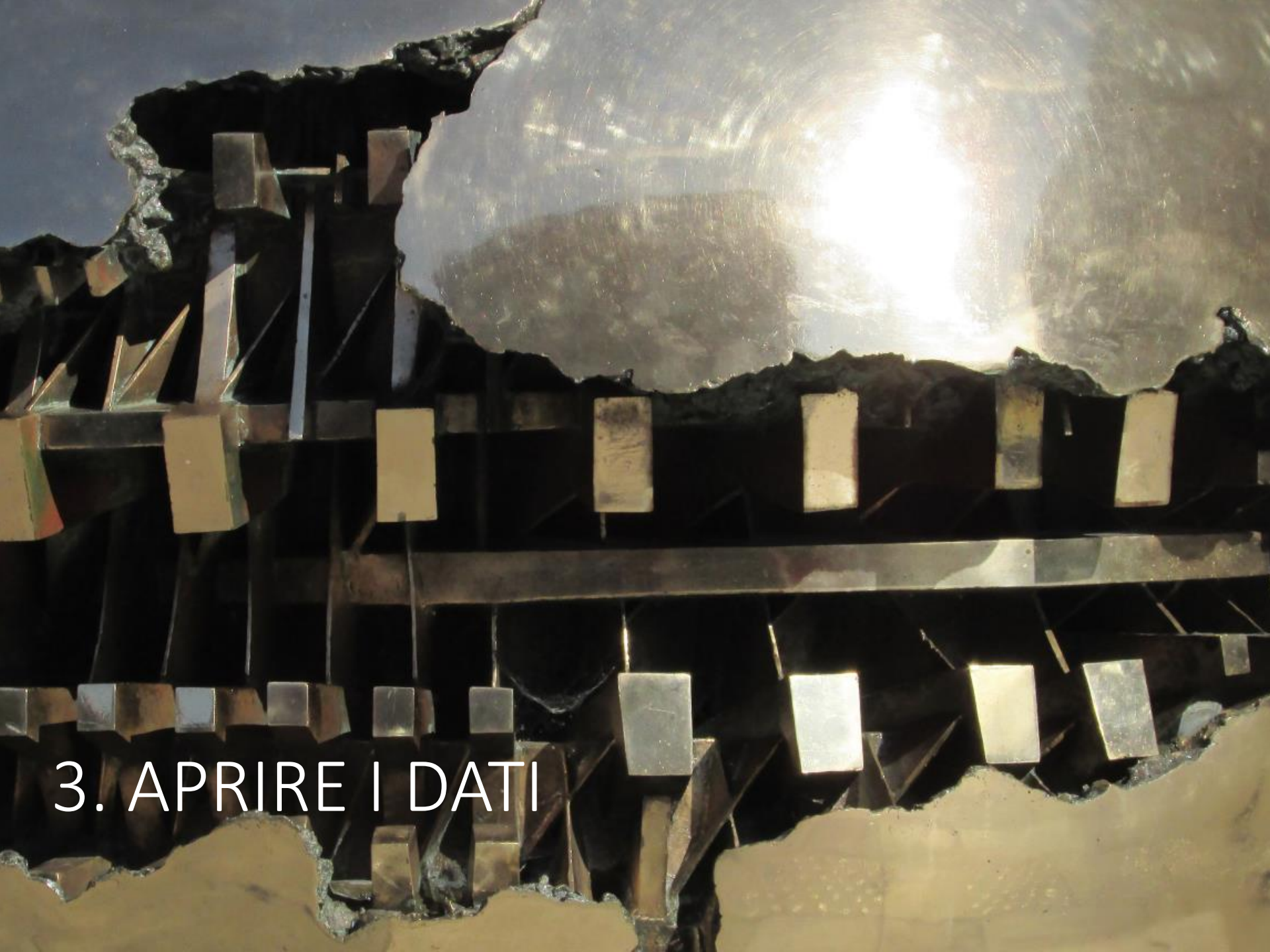
We recommend you avoid using a non-commercial licence. Here's why:

For legal purposes, drawing a line between what is and is not 'commercial' can be tricky; it's not as black and white as you might think. For example, if you release a dataset under a non-commercial licence, it would clearly prohibit an organisation

I'm uncomfortable permitting use of my research for any and all purposes. Should I use a 'No Derivatives' (ND) licence for my dataset?

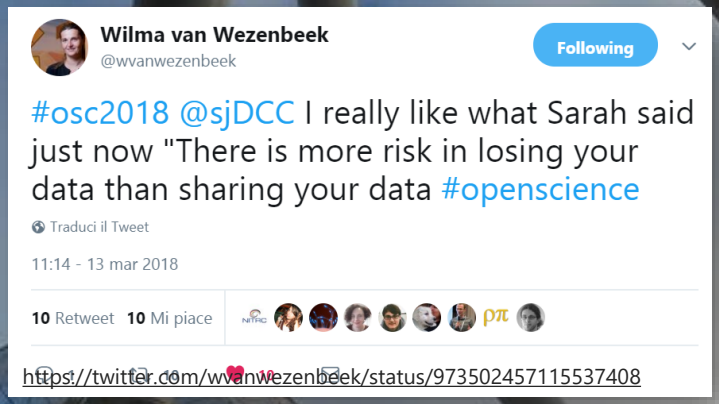
We recommend you avoid using a 'No Derivatives' licence. Here's why:

Similar to how a non-commercial licence might restrict meaningful reuse of your dataset, a ND licence can have the same effect: it may prevent someone from recombining and reusing your data for new research. For data to be truly Open Access, it must permit these important types of reuse.

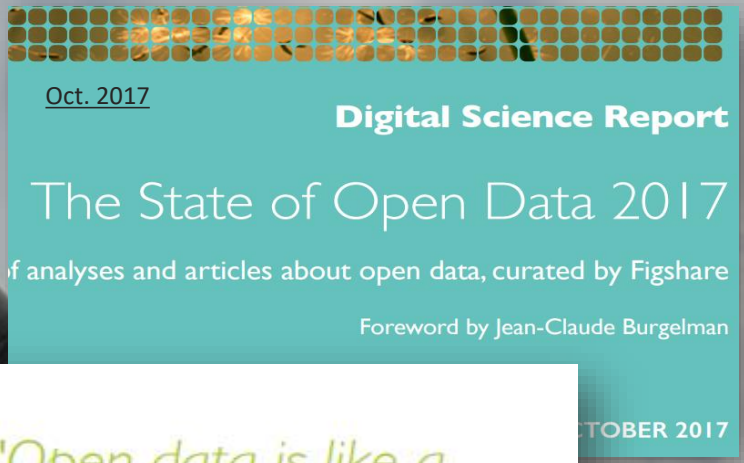


3. APRIRE I DATI

Perché Open Data?



https://twitter.com/wvanwezenbeek/status/973502457115537408



https://www.youtube.com/watch?v=HJb[illegible]ature=youtu.be

People will contact me to ask about stuff

YES, I KNOW. FRANKENSTEIN WAS THE DOCTOR, NOT THE MONSTER. FROM FLICKR BY CHOP SHOP GARAGE.

Christopher and Alex (C&A) say: "This is usually an objection of people who feel overworked and that [data sharing] isn't part of their job…" I would add to this that science is all about learning from each other – if a researcher is opposed to the idea of discussing their datasets, collaborating with others, and generally being a good science citizen, then they should be outed by their community as a poor participant.

People will misinterpret the data

C&A suggest this: "Document how it should be interpreted. Be prepared to help and correct such people; those that misinterpret it by accident will be grateful for the help." From the UK Data Archive: "Producing good documentation and providing contextual information for your research project should enable other researchers to correctly use and understand your data."

It's worth mentioning, however, a second point C&A make: "Publishing may actually be useful to counter willful misrepresentation (e.g. of data acquired through Freedom of Information legislation), as one can quickly point to the real data on the web to refute the wrong interpretation."

My data is not very interesting

C&A: "Let others judge how interesting or useful it is — even niche datasets hav people that care about them." I'd also add that it's impossible to decide wheth dataset has value to future research. Consider the many datasets collected bef "climate change" was a research topic which have now become invaluable to documenting and understanding the phenomenon. From the UK Data Archive:

I might want to use it in a research paper

Anyone who's discussed data sharing with a researcher is familiar with this excuse. The operative word here is *might*. How many papers have we all considered writing, only to have them shift to the back burner due to other obligations? That said, this is a real concern.

C&A suggest the embargo route: "One option is to have an automatic or optional embargo; require people to archive their data at the time of creation but it becomes public after X months. You could even give the option to renew the embargo so only things that are no longer cared about become published, but nothing is lost and eventually everything can become open." Researchers like to have a say in the use of their datasets, but I would caution to have any restrictions default to sharing. That is, after X months the data are automatically made open by the repository.

I would also add that, as the original collector of the data, you are at a huge advantage compared to others that might want to use your dataset. You have knowledge about your system, the conditions during collection, the nuances of your methods, et cetera that could never be fully described in the best metadata.

I'm not sure I own the data

My data is too complicated.

C&A: "Don't be too smug. If it turns out it's not that complicated, it could harm your professional [standing]." I would add that if it's too complicated to share, then it's too complicated to reproduce, which means it's arguably not real scientific progress. This can be solved by more documentation.

My data is embarrassingly bad

C&A: "Many eyes will help you improve your data (e.g. spot inaccuracies)… people will accept your data for what it is." I agree. All researchers have been on the back end of making the sausage. We know it's not pretty most of the time, and we can accept that. Plus it helps you strive will be at managing and organizing data during your next collection phase.

It's not a priority and I'm busy

Good news! Funders are *making* it your priority! New sharing mandates in the OSTP memorandum state that any research conducted with federal funds must be accessible. You can expect these sharing mandates to drift down to you, the researcher, in the very near future (6-12 months).

CARLY STRASSER

Open Science + Data Sharing Advocate

Home

Blog Posts

Previous Research

http://carlystrasser.net/closed-data-excuses-excuses/

Closed Data… Excuses, Excuses

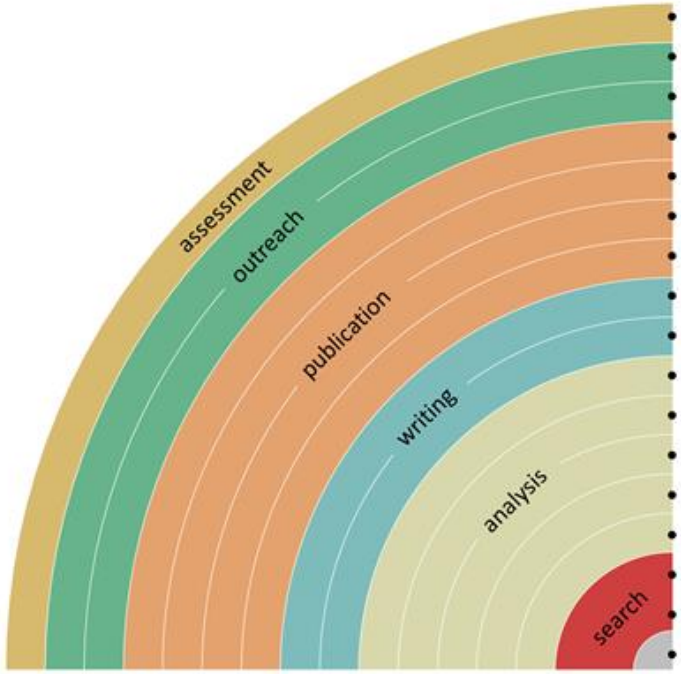
… «as open as possible»…

…effetti collaterali di essere «open»

…EVITARE DUPLICAZIONI INUTILI…
A MAGGIOR RAGIONE SE PUBBLICAE DATI NEGATIVI

Open Science: perché solo i dati?

You can make your workflow more open by …



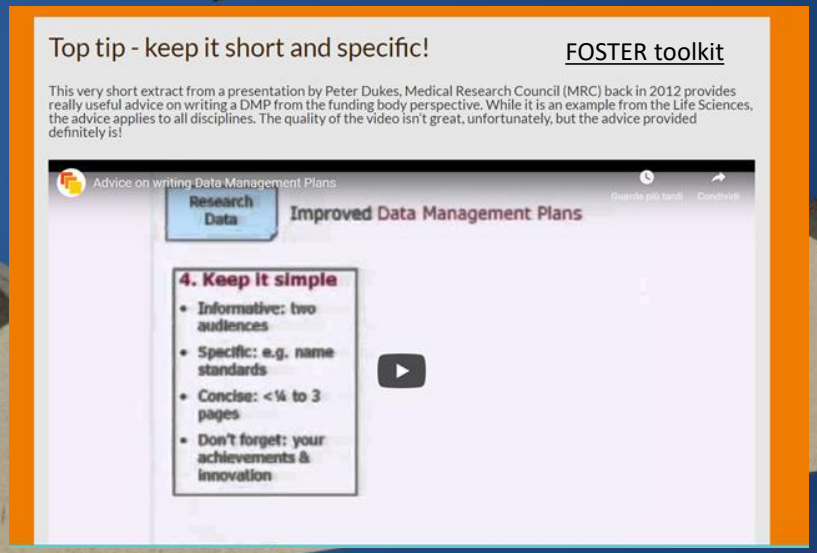
- adding alternative evaluation, e.g. with altmetrics
- communicating through social media, e.g. Twitter
- sharing posters & presentations, e.g. at FigShare
- using open licenses, e.g. CC0 or CC-BY
- publishing open access, 'green' or 'gold'
- using open peer review, e.g. at journals or PubPeer
- sharing preprints, e.g. at OSF, arXiv or bioRxiv
- using actionable formats, e.g. with Jupyter or CoCalc
- open XML-drafting, e.g. at Overleaf or Authorea
- sharing protocols & workfl., e.g. at Protocols.io
- sharing notebooks, e.g. at OpenNotebookScience
- sharing code, e.g. at GitHub with GNU/MIT license
- sharing data, e.g. at Dryad, Zenodo or Dataverse
- pre-registering, e.g. at OSF or AsPredicted
- commenting openly, e.g. with Hypothes.is
- using shared reference libraries, e.g. with Zotero
- sharing (grant) proposals, e.g. at RIO

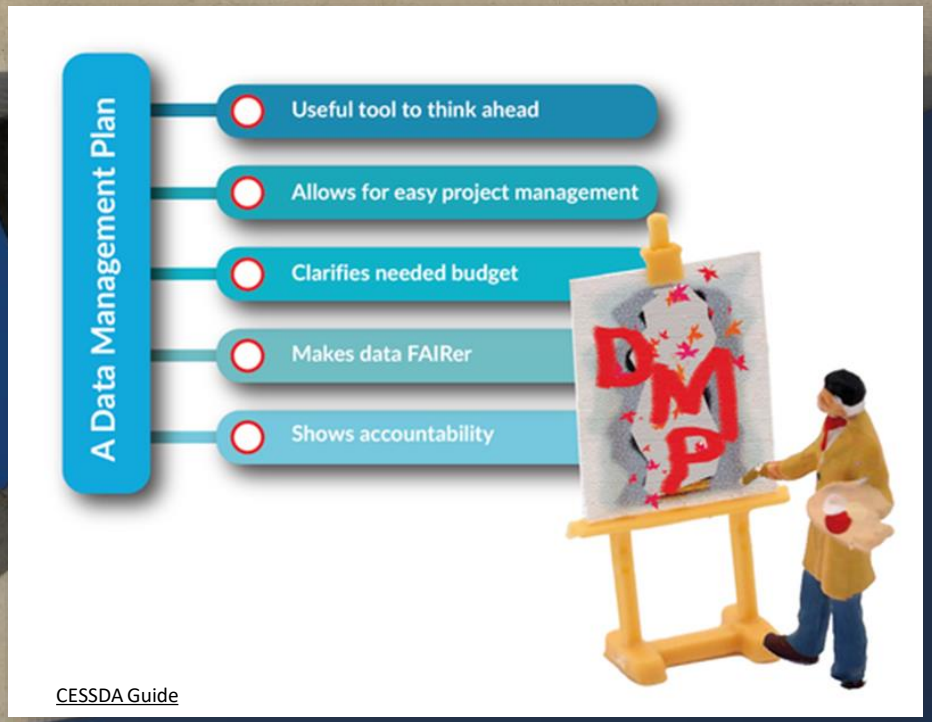
DOI: 10.5281/zenodo.1147025

UN MODO STRUTTURATO DI PENSARE AI DATI

REGOLE CHIARE=MENO ERRORI DA SUBITO

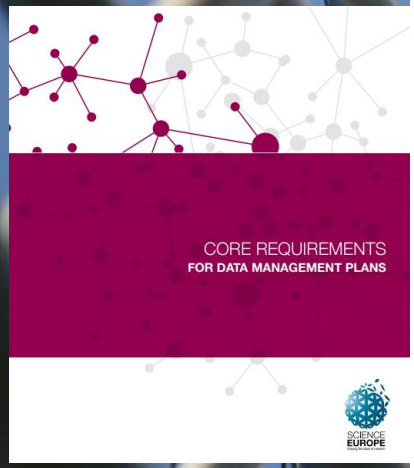
È UN «LIVING DOCUMENT», CRESCE COL PROGETTO





DOVE METTERE TUTTE QUESTE INFORMAZIONI? NEL DATA MANAGEMENT PLAN

DMP Core Requirements



CORE REQUIREMENTS
FOR DATA MANAGEMENT PLANS

When developing solid data management plans, researchers are required to deal with the following topics and answer the following questions:

1. **Data description and collection or re-use of existing data**
   a. How will new data be collected or produced and/or how will existing data be re-used?
   b. What data (for example the kinds, formats, and volumes) will be collected or produced?

2. **Documentation and data quality**
   a. What metadata and documentation (for example the methodology of data collection and way of organising data) will accompany data?
   b. What data quality control measures will be used?

3. **Storage and backup during the research process**
   a. How will data and metadata be stored and backed up during the research process?
   b. How will data security and protection of sensitive data be taken care of during the research?

4. **Legal and ethical requirements, codes of conduct**
   a. If personal data are processed, how will compliance with legislation on personal data and on data security be ensured?
   b. How will other legal issues, such as intellectual property rights and ownership, be managed? What legislation is applicable?
   c. How will possible ethical issues be taken into account, and codes of conduct followed?

Dec. 2018

5. **Data sharing and long-term preservation**
   a. How and when will data be shared? Are there possible restrictions to data sharing or embargo reasons?
   b. How will data for preservation be selected, and where will data be preserved long-term (for example a data repository or archive)?
   c. What methods or software tools will be needed to access and use the data?
   d. How will the application of a unique and persistent identifier (such as a Digital Object Identifier (DOI)) to each data set be ensured?

6. **Data management responsibilities and resources**
   a. Who (for example role, position, and institution) will be responsible for data management (i.e. the data steward)?
   b. What resources (for example financial and time) will be dedicated to data management and ensuring that data will be FAIR (Findable, Accessible, Interoperable, Re-usable)?

DMP: le domande


cessda

Adapt your Data Management Plan

A list of Data Management Questions based on the Expert Tour Guide on Data Management

PLAN

Overview

Title of the project

Date of this plan

Description of the project
- What is the nature of the project?
- What is the research question?
- What is the project time line?

Origin of Data
- What kind of data will be used during the project?
- If you are reusing existing data: What is the scope, volume and format? How are different data sources integrated?
- If you are collecting new data can you clarify why this is necessary?

Principal researchers
- Who are the main researchers involved?
- What are their contact details?

Collaborating researchers (if applicable)
- What are their contact details and their roles in the project?

Funder (if applicable)
- If funding is granted, what is the reference number of the funding granted?

Data producer
- Which organisation has the administrative responsibility for the data?

Project data contact
- Who can be contacted about the project after it has finished?

Data owner(s)
- Which organisation(s) own(s) the data?
- If several organisations are involved, which organisation owns what data?

Roles
- Who is responsible for updating the DMP and making sure that it's followed?
- Do project participants have any specific roles?
- What is the project time line?

Costs
- Are there costs you need to consider to buy specific software or hardware?
- Are there costs you need to consider for storage and backup?
- Are potential expenses for (preparing the data for) archiving covered?

ORGANISE & DOCUMENT

Organising and documenting your data

Data collection
- How will the data be collected?
- Is specific software or hardware or staff required?
- Who will be responsible for the data collection?
- During which period will the data be collected?
- Where will the data be collected?

Data organisation
- How will you organise your data?
- Will the data be organised in simple files or more complex databases?
- How will the data quality during the project be ensured?
- If data consists of many different file types (e.g. videos, text, photos), is it possible to structure the data in a logical way?

Data type and size
- What type(s) of data will be collected?
- What is the scope, quantity and format of the material?
- After the project: What is the total amount of data collected (in MB/GB)?

File format
- In what format will your data be?
- Does the format change from the original to the processed/final data?
- Will your (final) data be available in an open format?

Folder structure and names
- How will you structure and name your folders?

File structure and names
- How will you structure and name your files?

Documentation
- What documentation will be created during the different phases of the project?
- How will the documentation be structured?

Metadata
- What metadata will be provided with the collected/ generated/ reused data?
- How will metadata for each object be created?
- Is there any program that can be used to document the data?
- Can metadata be added directly into the files or will the metadata be produced in another program or document?

Metadata standard (if applicable)
- What metadata standard(s) will you use?

https://www.cessda.eu/content/download/4302/48656/file/TTT_DO_DMPExpertGuide_v1.2.pdf

DMP tools

ALLA FINE DI OGNI SEZIONE, TROVATE «ADAPT YOUR DMP» PER APPLICARE QUELLO CHE AVETE IMPARATO

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:

cessda TRAINING

⊕ 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:

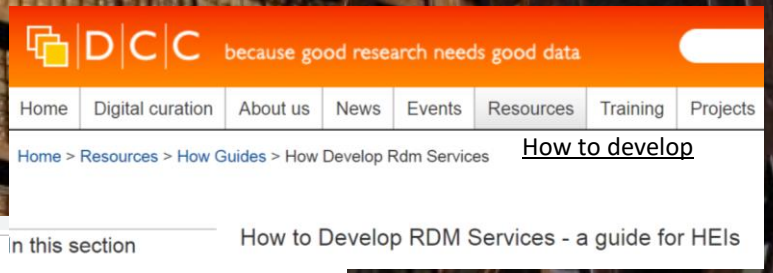
- Will established *terminologies/ontologies* (i.e. structured controlled vocabularies) be used in the project? If not, how does yours relate to established ones?
- Which *coding* is used (if any)? How does this relate to other research?

⊖ 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?

CESSDA Guide

Cosa conservare?



n this section

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 online

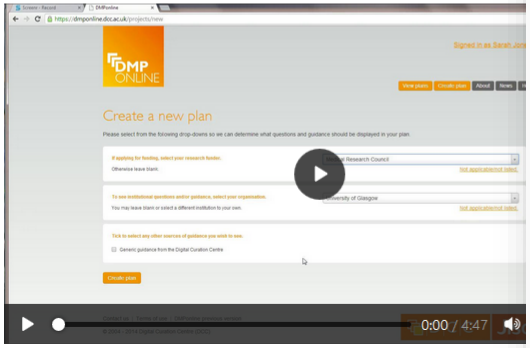
DMP ONLINE

Home | About | Future plans | Help | Change language

Welcome.

DMPonline helps you to create, review, and share data management
meet institutional and funder requirements. It has been jointly develo
Digital Curation Centre (DCC) and the University of California Cura
(UC3).

Screencast on how to use DMPonline



Sign in

Veteran tapes

Project Details | Plan overview | Write Plan | Share | Download

expand all | collapse all

13/13 answered

Data Collection (2 / 2)

What data will you collect or create?

The "Veteran tape " project will collect and generate different types of datasets:

Type of data	Volume	Format	Storage format
Video recordings	600 x 1Gb	.mkv	.mkv
Transcriptions	600 x 1500Kb	MS Word	.txt
Structured interview text	1 x 500Kb	MS word	.txt

For the video recordings the selected format is .mkv; the same .mkv format will be used for the long-term preservation .

Transcriptions will be written in MS Word and then stored as .txt files.

We checked the format compatibility against EASY File format https://dans.knaw.nl/en/deposit/information-about-depositing-data/before-depositing/file-formats

As the total volume of data is greater than 50Gb, DANS requires a fee for the storage. We are currently in touch with EASY to determine the costs of archiving.

Save

Guidance | Comments (1)

DCC | DCC guidance

Guidance
Questions to consider:

- What type, format and volume of data?
- Do your chosen formats and software enable sharing and long-term access to the data?
- Are there any existing data that you can reuse?

Guidance:

Give a brief description of the data, including any existing data or third-party sources that will be used, in each case noting its content, type and coverage. Outline and justify your choice of format and consider the implications of data format and data volumes in terms of storage, backup and access.

https://dmponline.dcc.ac.uk/

Citare i dati

Citing data is important in order to:

- Give the data producer appropriate credit
- Allow easier access to the data for repurposing or reuse
- Enable readers to verify your results

Citation Elements

A dataset should be cited formally in an article's reference list, not just informally in the text. Many data repositories and publishers provide explicit instructions for citing their contents. If no citation information is provided, you can still construct a citation following generally agreed-upon guidelines from sources such as the Force 11 Joint Declaration of Data Citation Principles and the current DataCite Metadata Schema.

Core elements

- There are 5 core elements usually included in a dataset citation, with additional elements added as appropriate.
  - **Creator(s)** – may be individuals or organizations
  - **Title**
  - **Publication year** when the dataset was released (may be different from the Access date)
  - **Publisher** – the data center, archive, or repository
  - **Identifier** – a unique public identifier (e.g., an ARK or DOI)
- Creator names in non-Roman scripts should be transliterated using the ALA-LC Romanization Tables.

Common additional elements

- Although the core elements are sufficient in the simplest case – citation to the entirety of a static dataset – additional elements may be needed if you wish to cite a dynamic dataset or a subset of a larger dataset.
  - **Version** of the dataset analyzed in the citing paper
  - **Access date** when the data was accessed for analysis in the citing paper
  - **Subset** of the dataset analyzed (e.g., a range of dates or record numbers, a list of variables)
  - **Verifier** that the dataset or subset accessed by a reader is identical to the one analyzed by the author (e.g., a Checksum)
  - **Location** of the dataset on the internet, needed if the identifier is not "actionable" (convertable to a web address)

Example citations

- Kumar, Sujai (2012): 20 Nematode Proteomes. figshare. https://doi.org/10.6084/m9.figshare.96035.v2 (Accessed 2016-09-06).
- Morran LT, Parrish II RC, Gelarden IA, Lively CM (2012) Data from: Temporal dynamics of outcrossing and host mortality rates in host-pathogen experimental coevolution. Dryad Digital Repository. https://doi.org/10.5061/dryad.c3gh6
- Donna Strahan. "08-B-1 from Jordan/Petra Great Temple/Upper Temenos/Trench 94/Locus 41". (2009) In Petra Great Temple Excavations. Martha Sharp Joukowsky (Ed.) Releases: 2009-10-26. Open Context. https://opencontext.org/subjects/30C3F340-5D14-497A-B9D0-7A0DA2C019F1 ARK (Archive): http://n2t.net/ark:/28722/k2125xk7p
- OECD (2008), Social Expenditures aggregates, OECD Social Expenditure Statistics (database). https://doi.org/10.1787/000530172303 (Accessed on 2008-12-02).
- Denhard, Michael (2009): dphase_mpeps: MicroPEPS LAF-Ensemble run by DWD for the MAP D-PHASE project. World Data Center for Climate. https://doi.org/10.1594/WDCC/dphase_mpeps
- Manoug, J L (1882): Useful data on the rise of the Nile. Alexandria : Printing-Office V Penasson. http://n2t.net/ark:/13960/t44q88124

https://dmptool.org/general_guidance#metadata-data-documentation

…ora tocca a voi!

Grazie