Dati della ricerca: una introduzione

Paola Galimberti (Università degli Studi di Milano)

Agenda

- Perché i dati sono importanti
- Il tema della integrità della ricerca
- Le politiche dell’Europa
- EOSC
- Le politiche dell’Italia
- Cosa può fare un ateneo per supportare i propri ricercatori

Di cosa stiamo parlando

Dati, dati e dati

- In molte discipline quando si fa ricerca si producono enormi quantità di dati a sostegno di una determinata teoria

- I dati che non funzionano **in questo contesto** vengono eliminati, e non sono resi disponibili per altri scienziati. In questa interpretazione dei dati essi sono la prova della plausibilità di una ipotesi formulata ex ante, la certificazione di una nuova scoperta.

- Nel **nuovo approccio ai dati**, essi vengono posti come risultato di una ricerca con un proprio valore scientifico che va oltre la conferma di una determinata teoria e che ne permette interpretazioni diverse a seconda del background dello scienziato che li analizza (Sabina Leonelli ***La ricerca scientifica nell'era dei Big Data,*** 2018)

Article + approach (fino ad ora)

Pensato per le persone

L’articolo (a volte accessibile solo a pagamento) è al centro e allegati sono i link ai **supplementary materials.** Non è richiesto un formato specifico, non è definito il trattamento dei dati, né come sono stati prodotti

Sia i dataset che l’articolo sono difficilmente leggibili dalle macchine (per problemi di diritti e per i formati e i metadati descrittivi)

Data + approach (da adesso in poi)

Pensato per le macchine e le persone

Dati FAIR sono l'elemento centrale della ricerca a cui si accompagna un testo descrittivo

Both human-readable outputs and machine-readable outputs have their rightful place in modern scholarly communication. (Barend Mons, Data Stewardship for open science)

Perché i dati della ricerca ci interessano?

- Research integrity
- Corretta attribuzione delle responsabilità (chi ha fatto cosa e quando)
- Riproducibilità
- Riutilizzo

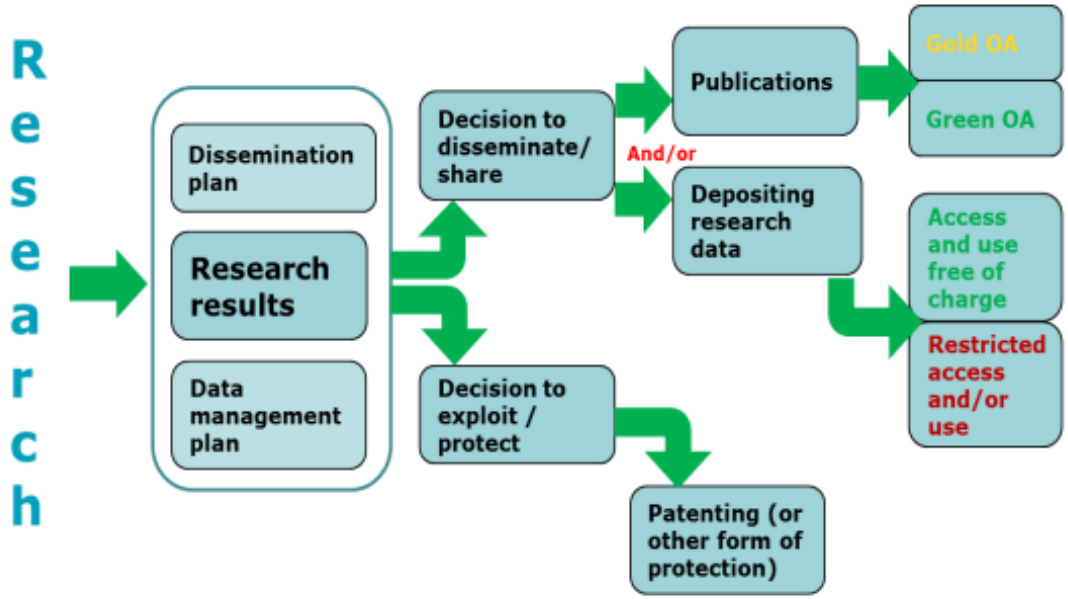
Perché i dati della ricerca ci interessano?

- A livello europeo è stato (da tempo) dato grande rilievo al tema dei dati della ricerca: **Open Governative Data**
- Let me underline one initiative that I am supporting **to make digital technology work for governance and transparency: by opening up public data.** In the digital age, data takes on a whole new value, and with new technology we can do great things with it. Opening it up is not just good for transparency, it also stimulates great web content, and provides the fuel for a future economy.
- That's why I say that **data is the new oil for the digital age**.

Neelie Kroes (Vice-President of the European Commission responsible for the Digital Agenda) Bratislava 5 maggio 2012

Perché i dati della ricerca ci interessano?

- A livello europeo è stato (da tempo) dato grande rilievo al tema dei dati della ricerca



Graph: Open access to scientific publication and research data in the wider context of dissemination and exploitation

I dati della ricerca: una definizione ampia

I dati della ricerca sono record fattuali (dati numerici, simboli, testi, immagini o suoni) utilizzati come fonti primarie della ricerca scientifica necessari per convalidare i risultati della ricerca (UNIMI)

electronic text documents, spreadsheets, laboratory notebooks, field notebooks and diaries, questionnaires, transcripts and codebooks, audiotapes and videotapes, photographs and films, examination results, specimens, samples, artefacts, slides, database schemas, database contents, models, algorithms and scripts, workflows, standard operating procedures and protocols, experimental results, metadata and other data files like e.g. literature review records and email archives.

I dati della ricerca: altre definizioni a livello europeo

- LERU Roadmap for Research Data (LERU Research Data Working Group, Advice Paper No. 14 – December 2014):
- “Research data, from the point of view of the institution with a responsibility for managing the data, includes: All data which is created by researchers in the course of their work, and for which the institution has a curational responsibility for at least as long as the code and relevant archives/record keeping acts require, and third-party data which have originated within the institution or come from elsewhere.”
- b) The Australian Griffith University:
- “Research data are factual records, which may take the form of numbers, symbols, text, images or sounds, which are used as primary sources for research, which are commonly accepted in the research community as necessary to validate research findings.”
- c) The University of Minnesota:
- “Research data are data in any format or medium that relate to or support research, scholarship, or artistic activity. They can be classified as: Raw or primary data: information recorded as notes, images, video footage, paper surveys, computer files, etc. Processed data: analyses, descriptions, and conclusions prepared as reports or papers Published data: information distributed to people beyond those involved in data acquisition and administration

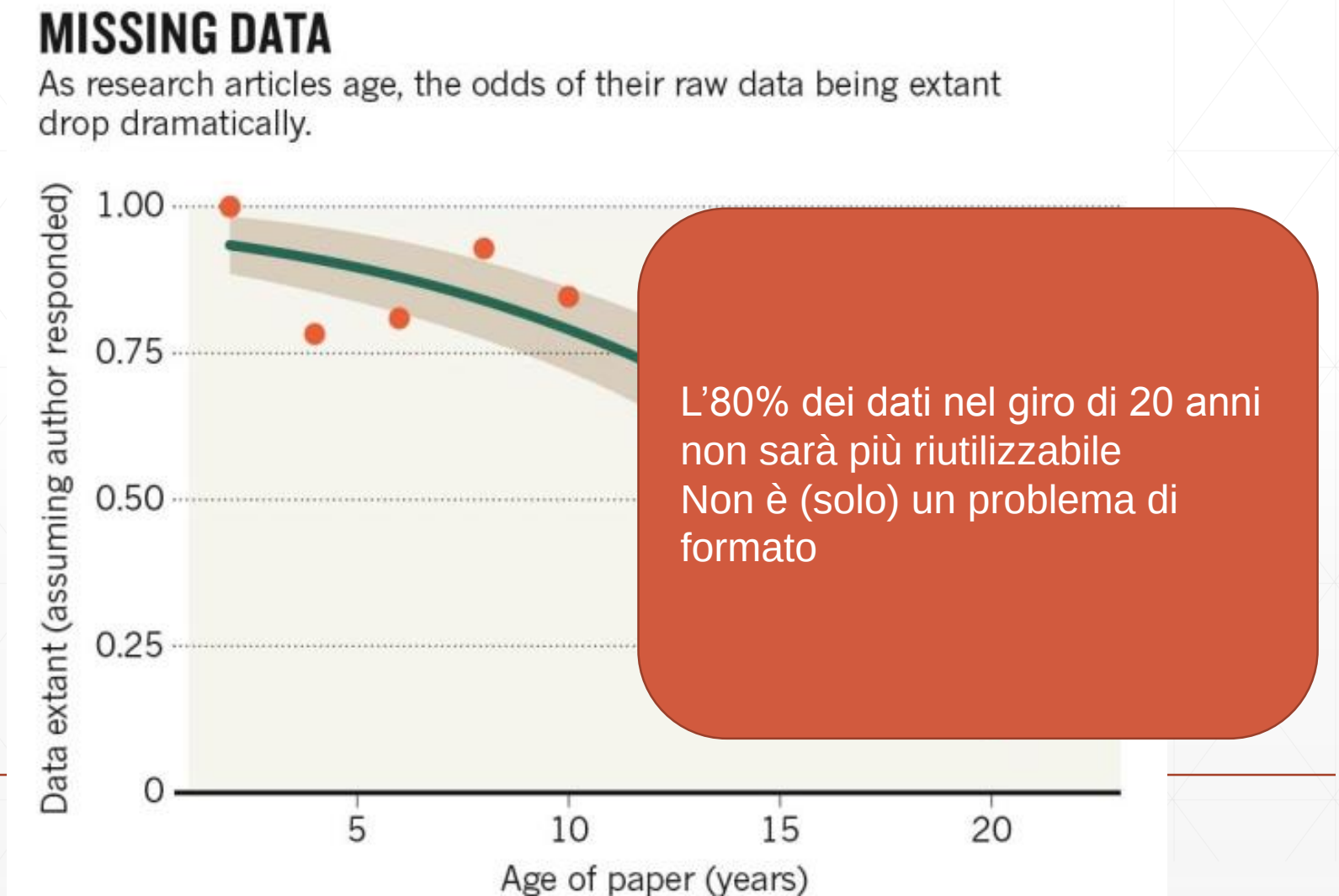
Concetti fondamentali

- Esiste una responsabilità precisa sui dati che devono essere gestiti e conservati secondo standard e regole condivisi
- I dati devono essere gestiti perché solo una corretta gestione ne permette il riutilizzo
- I dati correttamente gestiti permettono di validare le scoperte dei colleghi e rappresentano un valore inestimabile

Come si presentano i dati?

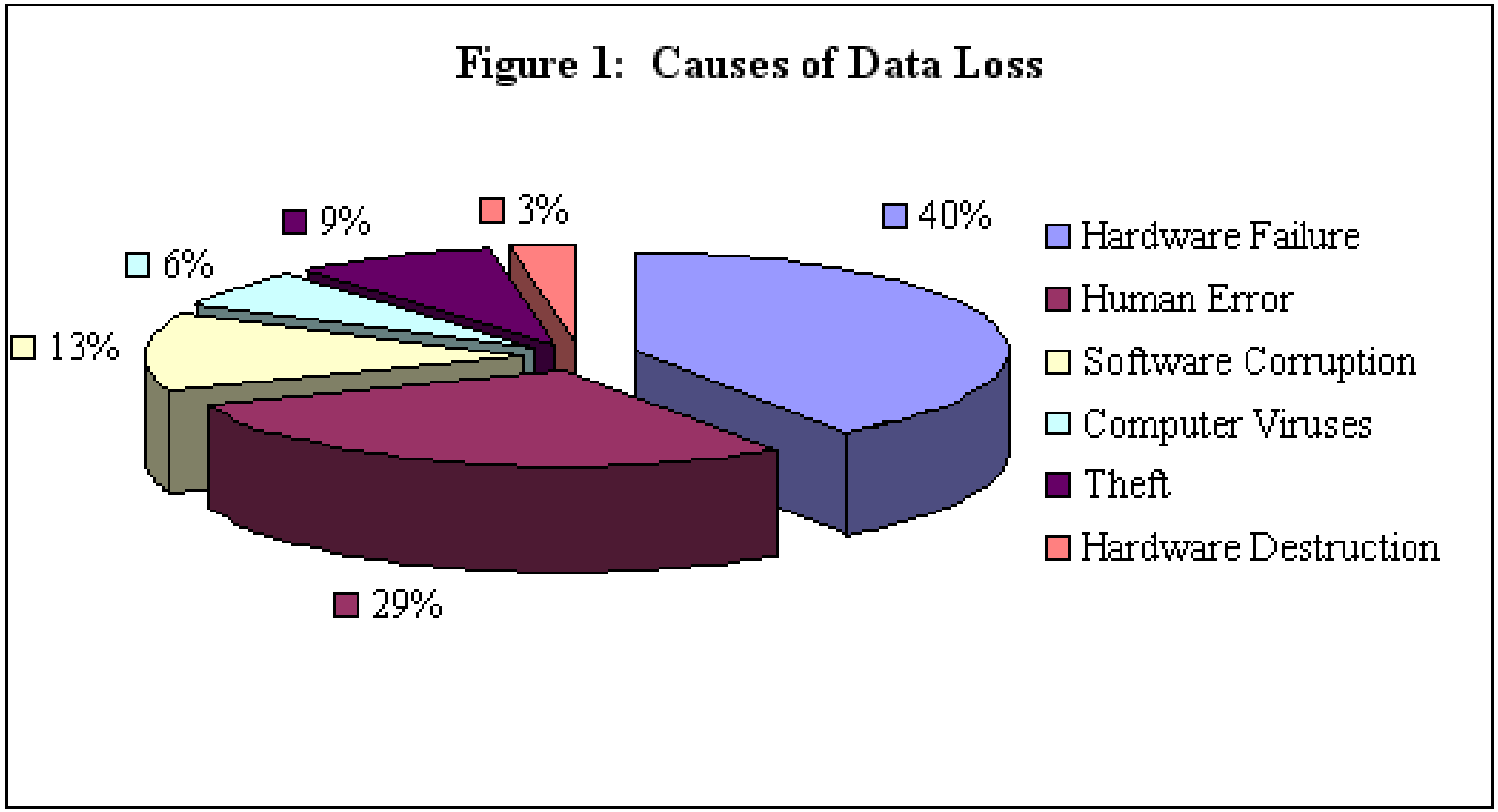
- I dati possono presentarsi sotto forme diverse:
  - Raw data
  - Elaborazioni dei raw data (è il caso più frequente)
  - Elaborazioni di dati già esistenti (con i problemi legati al loro riutilizzo)

Perché è importante gestire i dati?



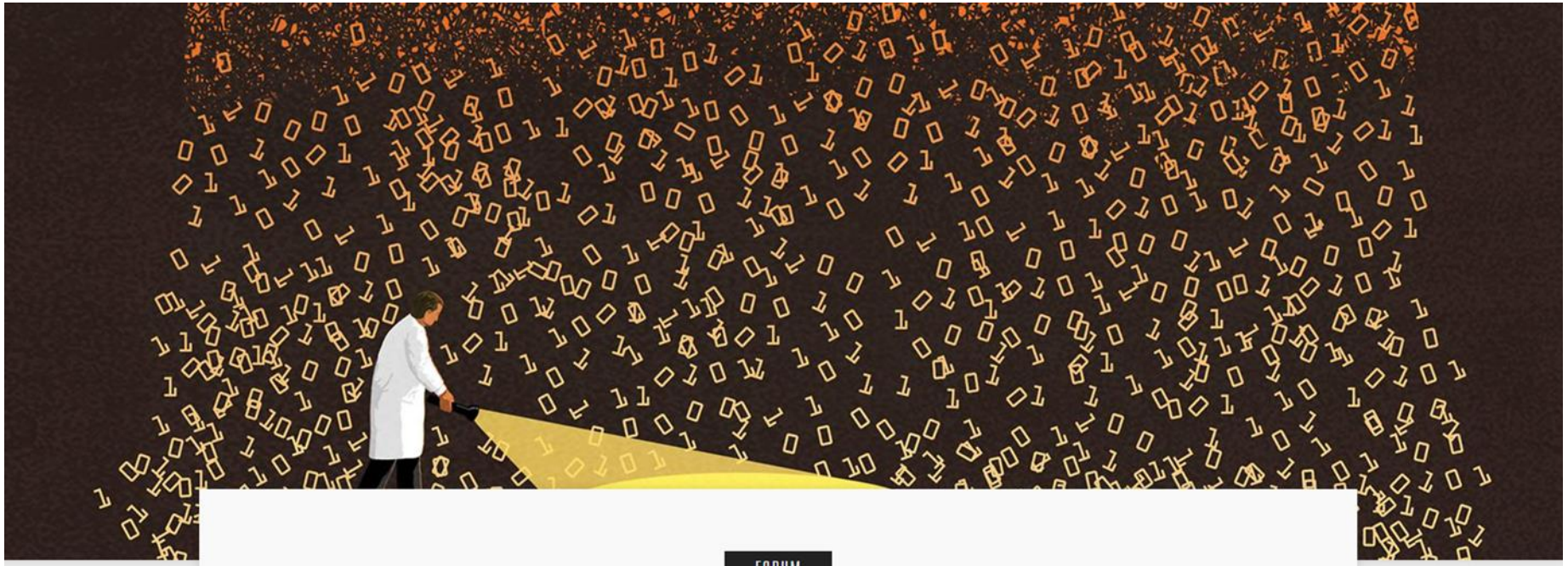
L'80% dei dati nel giro di 20 anni non sarà più riutilizzabile
Non è (solo) un problema di formato

Perdere i dati costa di più che conservarli



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.

Il valore dei dati (gestirli male ci fa perdere conoscenza e denaro)



FORUM

Lost Knowledge: Open Science is One Solution to Hidden Data

The progress of science depends on how we preserve and share what we know.

Il valore dei dati (gestirli male ci fa perdere conoscenza e denaro)

We lose knowledge when researchers design studies without first knowing about previous research into the question. We lose knowledge when studies or parts of studies (including negative or null results) are not reported. Knowledge may be lost if we report research findings in a way that makes them hard to find, such as in languages other than English (which may be the only language the searcher reads well) or in journals not indexed by Medline. And even when we publish in Medline-indexed journals, knowledge can be lost if the full publications are not generally available.

Responsabilità

To make your research as time-efficient, reproducible and safe as possible, it is important that your data management is well thought through, structured, and documented. A good data management strategy takes into account technical, organisational, structural, legal, ethical and sustainability aspects. The time invested in setting up a good data management strategy pays off when the time comes to reproduce your analysis and results. You will be able to easily find and understand your data, increase your data's reuse potential and comply with funder mandates at the same time.

Responsabilità

- Il tempo impiegato per pianificare il trattamento dei dati, per descriverli e per documentarli è tempo ben speso perché viene recuperato ex post quando si dovranno ritrovare i dati (nelle diverse versioni), individuare le responsabilità e cercare di riprodurre i risultati delle ricerche

C’è un problema di research integrity legato al tema del publish or perish

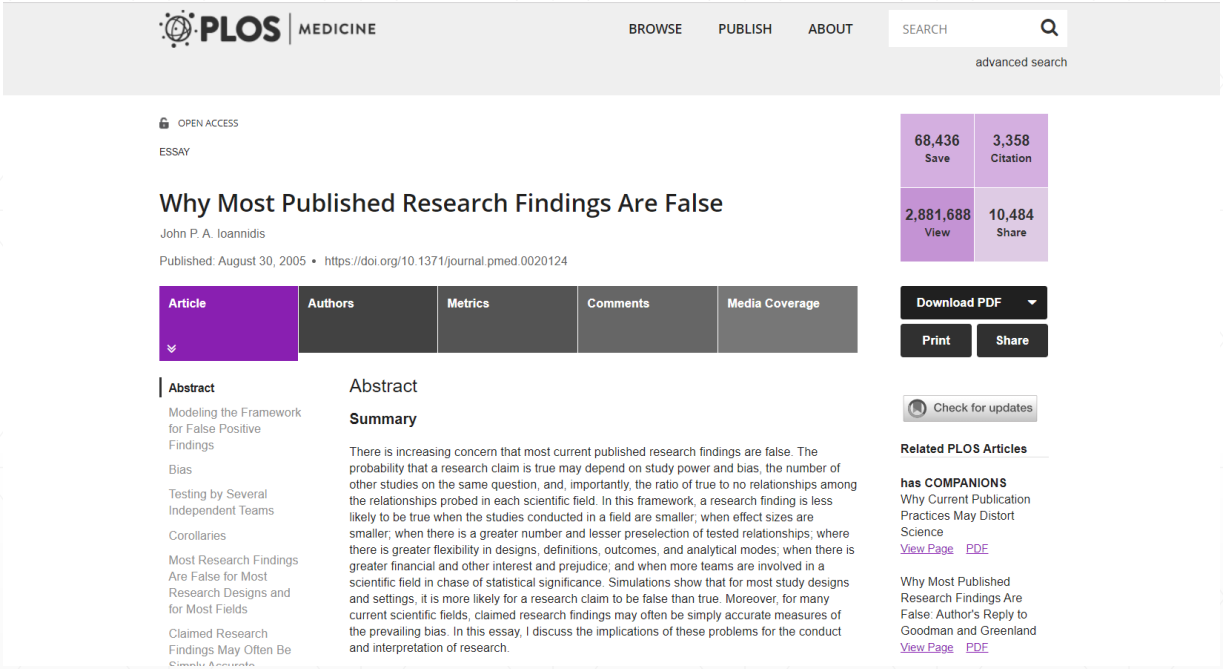
- *«The apparent endemicity of bad research behaviour is alarming … National assessment procedures, such as the Research Excellence Framework, incentivize bad practices»*

- The good news is that science is beginning to take some of its worst failings very seriously. The bad news is that nobody is ready to take the first step to clean up the system.

Research integrity

- La scienza ha oggi un forte problema rispetto alla riproducibilità delle ricerche.
- Alcune volte questa mancanza di riproducibilità si configura come **frode scientifica**
- Questo si traduce in **perdita di credibilità** rispetto a chi la ricerca la finanzia

Research integrity



PLOS MEDICINE

BROWSE PUBLISH ABOUT SEARCH

advanced search

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

Article | Authors | Metrics | Comments | Media Coverage

Abstract

Modeling the Framework for False Positive Findings

Bias

Testing by Several Independent Teams

Corollaries

Most Research Findings Are False for Most Research Designs and for Most Fields

Claimed Research Findings May Often Be Simply Accurate

Abstract

Summary

There is increasing concern that most current published research findings are false. The probability that a research claim is true may depend on study power and bias, the number of other studies on the same question, and, importantly, the ratio of true to no relationships among the relationships probed in each scientific field. In this framework, a research finding is less likely to be true when the studies conducted in a field are smaller; when effect sizes are smaller; when there is a greater number and lesser preselection of tested relationships; where there is greater flexibility in designs, definitions, outcomes, and analytical modes; when there is greater financial and other interest and prejudice; and when more teams are involved in a scientific field in chase of statistical significance. Simulations show that for most study designs and settings, it is more likely for a research claim to be false than true. Moreover, for many current scientific fields, claimed research findings may often be simply accurate measures of the prevailing bias. In this essay, I discuss the implications of these problems for the conduct and interpretation of research.

68,436 Save | 3,358 Citation

2,881,688 View | 10,484 Share

Download PDF | Print | Share

Check for updates

Related PLOS Articles

has COMPANIONS

Why Current Publication Practices May Distort Science

View Page PDF

Why Most Published Research Findings Are False: Author's Reply to Goodman and Greenland

View Page PDF

Research integrity

PLOS | MEDICINE

BROWSE PUBLISH ABOUT SEARCH

OPEN ACCESS

ESSAY

Why Most Published Research

John P. A. Ioannidis

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

Article | Authors | Metrics

Abstract

Modeling the Framework for False Positive Findings

Bias

Testing by Several Independent Teams

Corollaries

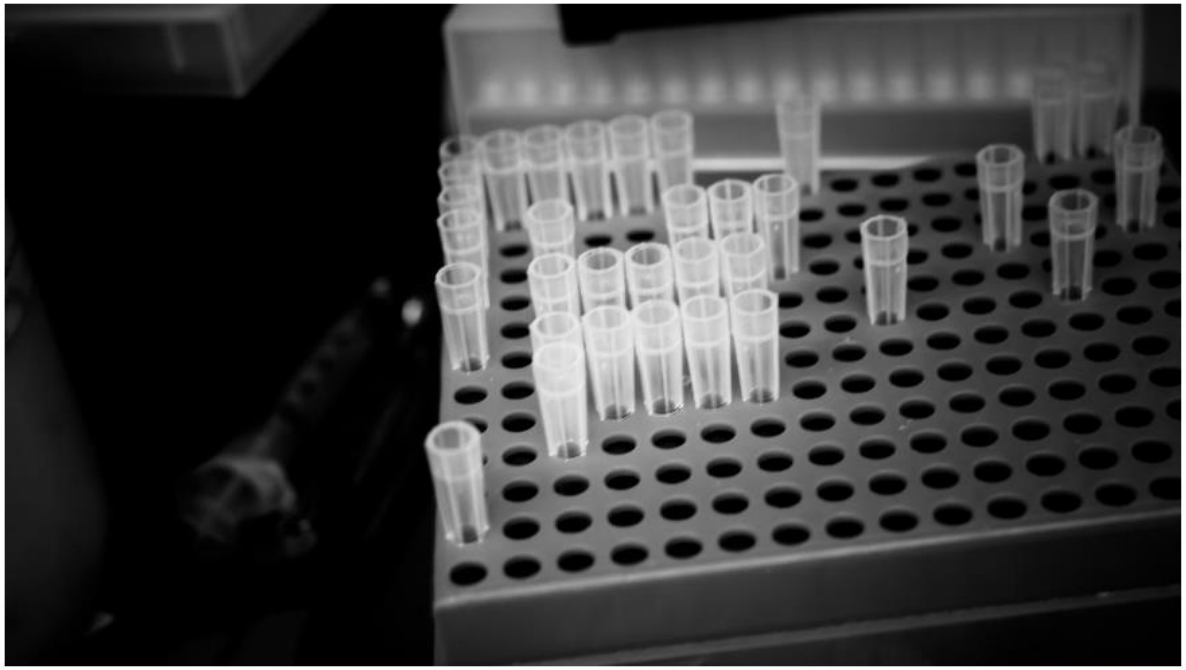
Most Research Findings Are False for Most Research Designs and for Most Fields

Claimed Research Findings May Often Be

Abstract

Summary

There is increasing concern t
probability that a research cla
other studies on the same qu
the relationships probed in ea
likely to be true when the stu
smaller; when there is a grea
there is greater flexibility in de
greater financial and other int
scientific field in chase of stat
and settings, it is more likely
current scientific fields, claim
the prevailing bias. In this ess
and interpretation of research



BILL DICKINSON/FLICKR (CC BY-NC-ND 2.0)

Study claims $28 billion a year spent on irreproducible biomedical research

By Jocelyn Kaiser | Jun. 9, 2015 , 1:30 PM

Esempio di un journal: Plos e la policy sui dati

- Data are any and all of the digital materials that are collected and analyzed in the pursuit of scientific advances. In line with its stance on providing Open Access to research articles themselves, PLOS strongly believes that, to best foster scientific progress, the underlying data should be made freely available for researchers to use, wherever this is legal and ethical. **Data availability allows validation, replication, reanalysis, new analysis, reinterpretation, or inclusion into meta-analyses, and facilitates reproducibility of research.** Making data available for all these uses provides a better “bang for the buck” out of scientific research, much of which is funded from public or nonprofit sources. Ultimately, our viewpoint is quite simple: Ensuring access to the underlying data should be an intrinsic part of the scientific publishing process.

Esempio di un journal: Plos e la policy sui dati

- PLOS journals require authors to make all data underlying the findings described in their manuscript fully available without restriction, with rare exception[1].
- When submitting a manuscript online, authors must provide a *Data Availability Statement* describing compliance with PLOS's policy. The data availability statement will be published with the article if accepted.
- Refusal to share data and related metadata and methods in accordance with this policy will be grounds for rejection. PLOS journal editors encourage researchers to contact them if they encounter difficulties in obtaining data from articles published in PLOS journals. If restrictions on access to data come to light after publication, we reserve the right to post a correction, to contact the authors' institutions and funders, or in extreme cases to retract the publication.
- […]

Esempio di un journal: Plos e la policy sui dati

- ***Data deposition (strongly recommended):*** All data and related metadata underlying the findings reported in a submitted manuscript should be deposited in an appropriate public repository[2], unless already provided as part of the submitted article. Repositories may be either subject-specific (where these exist) and accept specific types of structured data, or generalist repositories that accept multiple datatypes, such as Dryad [4]. Guidance on acceptable repositories is included below[2]. The *Data Availability Statement* must specify that data are deposited publicly and list the name(s) of repositories along with digital object identifiers or accession numbers for the relevant datasets. In some cases authors may not be able to obtain DOIs or accession numbers until the manuscript is accepted; in these cases, the authors must provide these numbers at acceptance. In all other cases, these numbers must be provided at submission.

Plos e la policy sui dati: scelta del repository

- PLOS requires that authors comply with field-specific standards for preparation and recording of data [6] and to select repositories appropriate to their field, for example deposition of microarray data in ArrayExpress or GEO; deposition of gene sequences in GenBank, EMBL or DDBJ; and deposition of ecological data in Dryad [7]. Authors are encouraged to select repositories that meet accepted criteria as trustworthy digital repositories, such as criteria of the Centre for Research Libraries [8] or Data Seal of Approval [9]. Large, international databases are more likely to persist than small, local ones. Copyright licensing for data held in repositories may be unclear. If authors use repositories with stated licensing policies; the policies should not be more restrictive than CC-BY.

Le richieste degli enti finanziatori: *as open as possible as closed as necessary*

Funder requirements

Templates for data management plans are based on the specific requirements listed in funder policy documents. The DCC maintains these templates, however, researchers should always consult the funder guidelines directly for authoritative information.

Search

Template Name	Download	Organisation Name	Last Updated	Funder Links	Sample Plans (if available)
BBSRC Template		Biotechnology and Biological Sciences Research Council (BBSRC)	18-10-2018		
NERC Template		Natural Environment Research Council (NERC)	22-05-2018		
ESRC Template		Economic and Social Research Council (ESRC)	18-10-2018		
Standard CRUK Template		Cancer Research UK (CRUK)	18-10-2018		
STFC Template		Science and Technology Facilities Council (STFC)	18-10-2018		
DCC Template		Digital Curation Centre	18-05-2018		
NSF - generic		National Science Foundation (USA)	18-10-2018		

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

I dati della ricerca vanno gestiti perché così richiedono gli enti finanziatori

The Commission is running a flexible pilot under Horizon 2020 called the Open Research Data Pilot (ORD pilot). The ORD pilot aims to improve and maximise access to and re-use of research data generated by Horizon 2020 projects and takes into account the need to balance openness and protection of scientific information, commercialisation and Intellectual Property Rights (IPR), privacy concerns, security as well as data management and preservation questions.

In the 2014-16 work programmes, the ORD pilot included only selected areas of Horizon 2020. Under the revised version of the 2017 work programme, the Open Research Data pilot has been extended to cover all the thematic areas of Horizon 2020.

While open access to research data thereby becomes applicable by default in Horizon 2020, the Commission also recognises that there are good reasons to keep some or even all research data generated in a project closed.

The Commission therefore provides robust *opt-out* possibilities at any stage, that is

- during the application phase
- during the grant agreement preparation (GAP) phase and
- after the signature of the grant agreement.

The ORD pilot applies primarily to the data needed to validate the results presented in scientific publications. Other data can also be provided by the beneficiaries on a voluntary basis, as stated in their Data Management Plans. Costs associated with open access to research data, can be claimed as eligible costs of any Horizon 2020 grant.

I dati della ricerca vanno gestiti perché così richiedono gli enti finanziatori

The Commission is running a flexible pilot under Horizon 2020 called the Open Research Data Pilot (ORD pilot). The ORD pilot aims to improve and maximise

- Open research data pilot 2014-16 su alcune aree
- Dal 2017 ORD su tutte le aree
- H2020 Open access ai dati della ricerca by default con possibilità di opt out in qualsiasi fase del progetto
- La presenza di un DMP per la gestione dei dati non prevede una valutazione migliore del progetto, ma una migliore gestione dei dati
- I costi di processamento e gestione dei dati sono imputabili al budget del progetto.

European Research Council
Executive Agency

Established by the European Commission

Guidelines on the Implementation of **Open Access to Scientific Publications** and **Research Data** in projects supported by the European Research Council under Horizon 2020

Version 1.1

07. April 2017

Research Data Management and Sharing

Concerning **research data**, the ERC Scientific Council's Open Access Guidelines further explain:

"The European Research Council supports the basic principle of Open Access to research data. It therefore recommends *to all its funded researchers that they follow best practice by retaining files of all the research data they have produced and used during the course of their work, and that they be prepared to share these data with other researchers whenever they are not bound by copyright restrictions, confidentiality requirements, or contractual clauses."*

Beneficiaries of ERC grants funded under the Work Programme 2016 may opt-in, on an individual and voluntary basis, **to the Horizon 2020 Pilot on Open Research Data** in order to facilitate access, re-use and preservation of research data generated during their research work. Beneficiaries choosing this option should carefully check the additional obligations that apply to projects that opt-in to the Pilot as described in Article 29.3 of the ERC Model Grant Agreement under Horizon 2020. As of the Work Programme 2017 the Pilot on Open Research Data is being extended to cover all thematic areas of Horizon 2020 and open access becomes the default setting for the research data generated. The beneficiaries may still opt out at any stage, freeing themselves from any obligations regarding the open access to digital research data generated in the action. Please also see ERC Data Management Plan template.

Le richieste a livello europeo

- Horizon 2020, il programma quadro di finanziamento europeo della ricerca, prevede per **TUTTI i progetti finanziati**:
- l'**obbligo** di rendere **disponibili in Open Access** i risultati della ricerca (**pubblicazioni**)
- **l'obbligo** di rendere **disponibili** insieme agli articoli anche **i dati della ricerca** su cui si basano (non i dati inediti, ovviamente). Il Progetto Pilota sugli Open Research Data, attivo inizialmente solo per 9 aree, è stato quindi esteso a tutti i progetti

Le richieste a livello europeo

- La gestione dei dati della ricerca è complessa in ragione
  - dell'eterogenità della loro natura
  - dell'eterogenità dei formati
  - di questioni legate alla integrità
  - di questioni legate all'accessibilità
- Per questo motivo, Horizon 2020 richiede che per ogni set di dati venga compilato un **Data Management Plan**, che indichi le misure prese per rendere fruibili i dati anche in futuro.
- I Data Management Plans non sono un carico burocratico ma uno **strumento di tutela** del ricercatore, perché **aiutano nella gestione, archiviazione, conservazione** dei set di dati.

EOSCpilot
The European Open Science Cloud for Research Pilot Project
www.eoscpilot.eu

What is EOSC?

"The objective of the EOSC is to give the Union a global lead in

research data management

and ensure that

European scientists reap the full benefits of data-driven science,

by offering 1.7 million European researchers and 70 million Professionals in science and technology

a virtual environment

with

free at the point of use, open and seamless services

for

storage, management, analysis and re-use of research data,

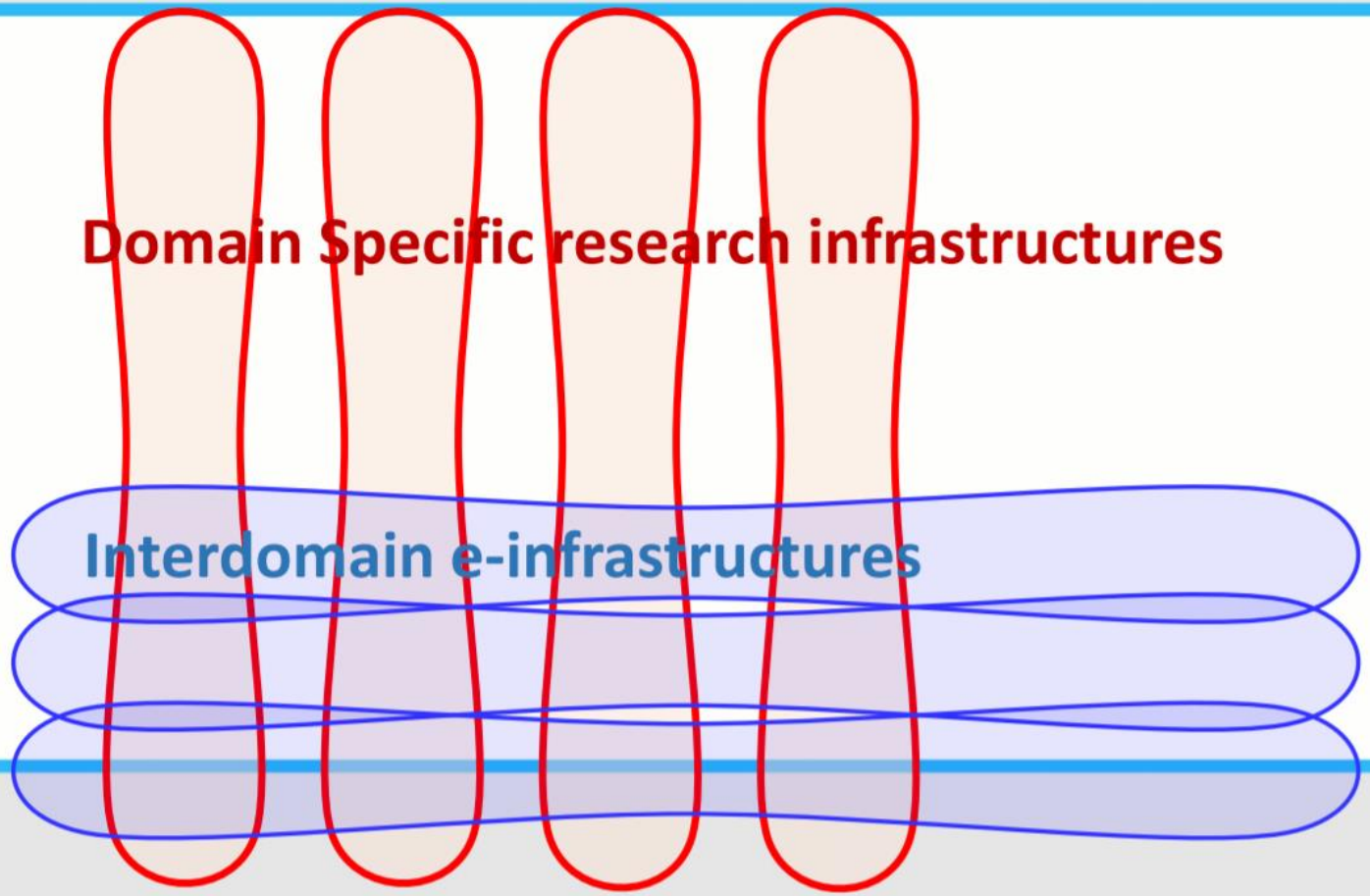
across borders and scientific disciplines."

EOSC Implementation Roadmap (March 2018)

Come funziona EOSC?

EOSC pilot
The European Open Science Cloud for Research Pilot Project
www.eoscpilot.eu

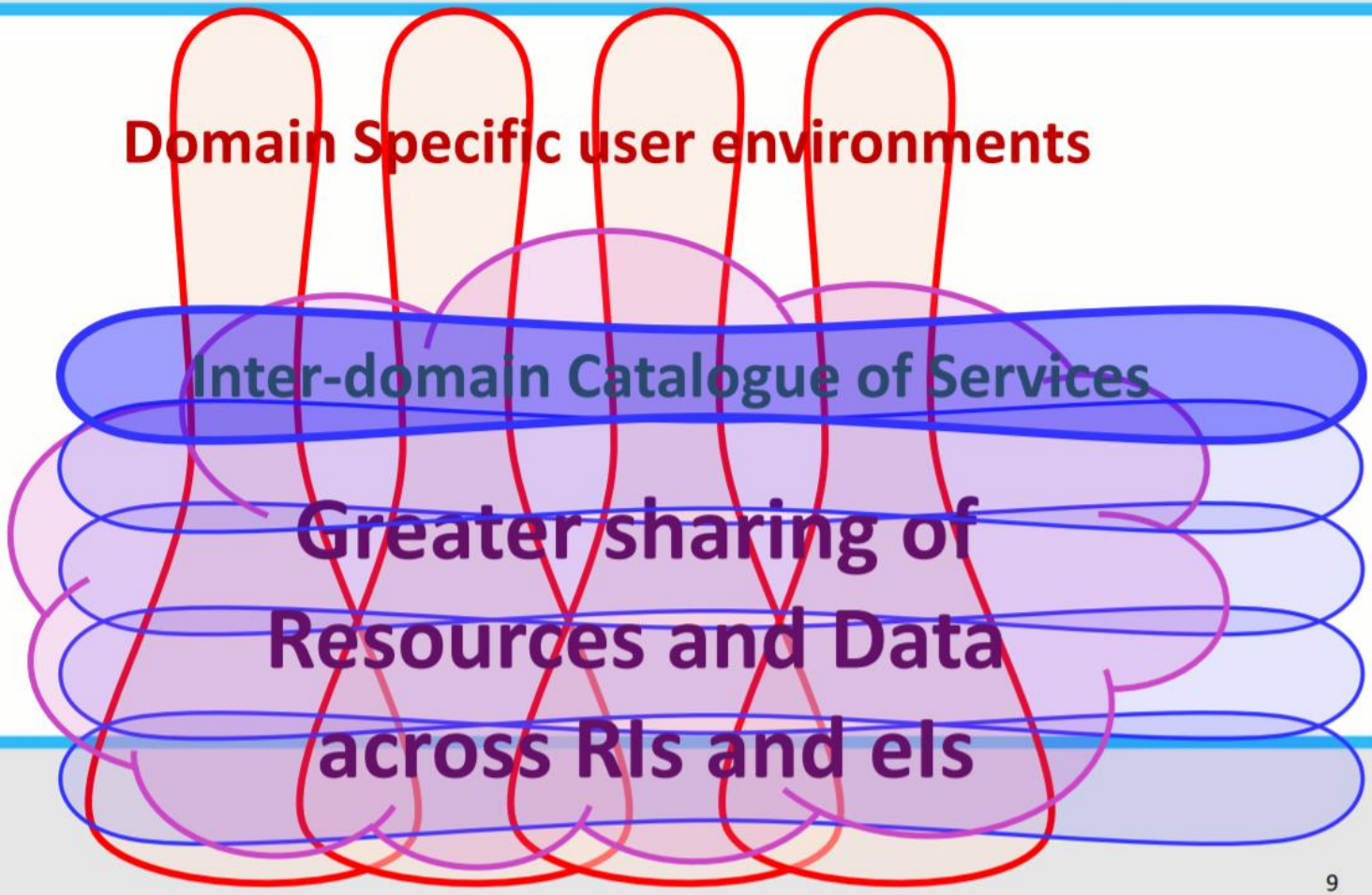
Existing
Horizontal and Vertical infrastructures



Credit: e-IRG, eInfrastructure Commons, 2013

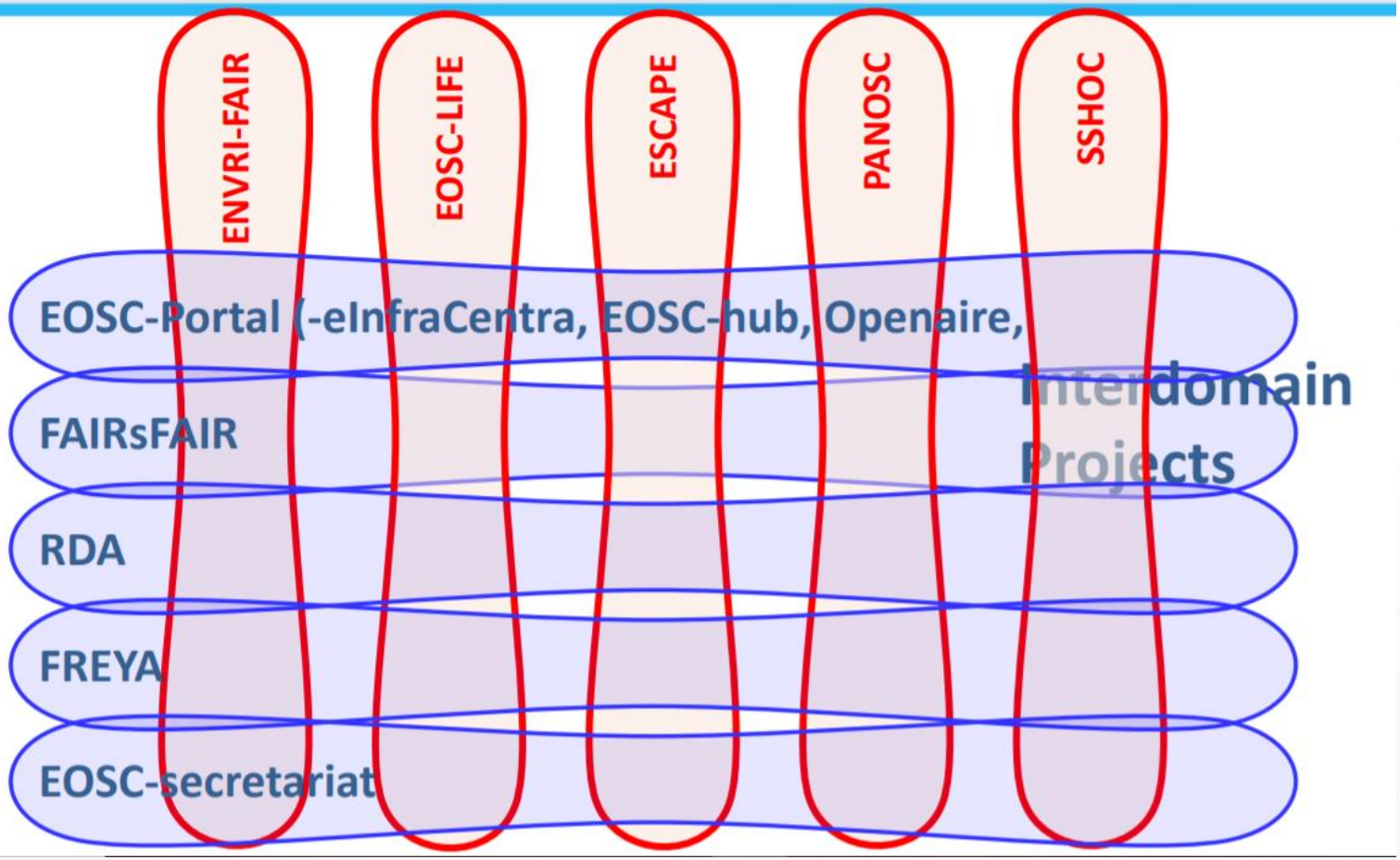
EOSCpilot
The European Open Science Cloud for Research Pilot Project
www.eoscpilot.eu

Merging
Horizontal and Vertical infrastructures



EOSC pilot
The European Open Science Cloud for Research Pilot Project
www.eoscpilot.eu

Developing
Horizontal and Vertical infrastructures



Horizon Europe, 2021-27, €100B

EOSCpilot
The European Open Science Cloud for Research Pilot Project
www.eoscpilot.eu

"**Open Science** will become the modus operandi of Horizon Europe."

The new programme will be implemented through three pillars:

- The **Open Science pillar** (€25.8 billion) supports frontier research projects defined and driven by researchers themselves through the **European Research Council** (€16.6 billion), funds fellowships and exchanges for researchers through **Marie Skłodowska-Curie Actions** (€6.8 billion), and invests in world-class research infrastructures.

[EU budget for the future, Horizon Europe,June 2018, endorsed by EP April 2019]]
https://ec.europa.eu/commission/sites/beta-political/files/budget-may2018-research-innovation_en.pdf

Europa

- Le indicazioni dell’Europa sono chiare.
- I dati e la scienza data driven avranno un ruolo fondamentale nel prossimo programma quadro.
- EOSC sarà la infrastruttura attraverso la quale passerà la ricerca finanziata dall’Europa.
- Vediamo alla luce di tutta questa attività cosa sta succedendo in Italia e quale ruolo hanno dati della ricerca e quale attenzione viene dedicata a questo tema.

In Italia

- Non ci sono a livello centrale regole precise riferite ai dati della ricerca
- Unico cenno al loro trattamento è in un documento del 2016 sui big data

I DATI DELLA RICERCA: OPEN SCIENCE

Alla seconda categoria di dati si può ricondurre l'accesso ai risultati della ricerca. La loro ampia circolazione e più efficace utilizzo sono considerati precondizioni necessarie per un adeguato sviluppo di una società innovativa e della conoscenza, posto che il loro impatto innovativo è chiaramente collegato alla condizione che essi siano accessibili al settore delle imprese e che la conoscenza circoli liberamente.

Il discorso sull'apertura dei risultati della ricerca è radicato nel tempo: partito da movimenti sulla condivisione aperta delle pubblicazioni scientifiche (Open Access), il dibattito si è evoluto verso la condivisione dei dati primari della ricerca (Open Science) e del sistema dell'innovazione (Open Innovation) [18].

Temi complessi e ibridi come l'Open Science investono ambiti, competenze e problemi assai vasti e articolati che coinvolgono direttamente la governance sia degli Enti di ricerca e delle Università sia del Ministero stesso. Coinvolgono le politiche della ricerca, le scelte strategiche nell'istruzione e formazione, la valutazione, i modelli e le politiche dell'informazione e comunicazione scientifica.

In Italia

- È necessario implementare un **sistema di incentivi per i ricercatori,** a partire dalle procedure di valutazione MIUR sulla produttività dei ricercatori (VQR)**, in modo che sia premiato il riutilizzo e la citazione dei dati da loro messi a disposizione.**
- È auspicabile l'emanazione di **bandi competitivi per il finanziamento di progetti di ricerca che abbiano i *Big Data* sia come metodo, sia come oggetto di studio.**
- Deve essere previsto su un piano di sensibilizzazione sul valore della condivisione dei dati da parte dei ricercatori con modalità che tutelino la proprietà intellettuale.
- Un'altra sfida attiene alla necessità di **sviluppare infrastrutture elettroniche di ricerca** che siano capaci di ospitare grandi moli di dati per la loro condivisione e conservazione nel tempo. Su questo punto, il MIUR deve promuovere un ruolo attivo dell'Italia per collaborare alla European Cloud Initiative cui tutte le nostre eccellenze del settore hanno immediatamente risposto partecipando da subito alle prime fasi del progetto.

In Italia

- Non ci sono a livello centrale r

- È necessario implementare un **sistema di incentivi per i ricercatori,** a partire dalle procedure di valutazione MIUR sulla produttività dei ricercatori (VQR)**, in modo che sia premiato il riutilizzo e la citazione dei dati da loro messi a disposizione.**
- È auspicabile l'emanazione di **bandi competitivi per il finanziamento di progetti di ricerca che abbiano i *Big Data* sia come metodo, sia come ogget-**

Da questo documento del 2016 SILENZIO

i modelli e le politiche dell'informazione e comunicazione scientifica.

Aspettando un archivio centrale per i dati che fare?

Più volte si sente dire «archivia in Zenodo, o archivia in Figshare o in Dryad» o in altri strumenti.

Di fatto sono gli stessi discorsi che si facevano quando non avevamo gli archivi istituzionali. C'era loginmiur, Pubmed, Scopus ecc.

Ma può una istituzione rispetto a un tema così delicato come quello dei dati essere responsabile di dati archiviati in strumenti che non governa? Che non gestisce?

Come rispondono le istituzioni alle richieste degli enti finanziatori?

- Possiamo supportare i nostri ricercatori?
- Sappiamo di cosa hanno bisogno?
- Abbiamo idea di quali sono le conoscenze dei nostri utenti rispetto ai requisiti richiesti e a come realizzarli?

I dati in un ateneo multidisciplinare: setting the scene

- Quali dati?
- Quanti dati?
- In quali formati?
- Sono ricercabili?
- Quale tipo di diffusione?
- Chi è il titolare dei dati?
- Dove vengono archiviati?
- E’ capitato che si siano persi dei dati?
- E’ prevista una conservazione a lungo termine?
- Dove? Come? A carico di chi?

I dati FAIR

- I dati, per poter essere **conservati, condivisi, riutilizzati**, devono essere trattati in maniera standard e avere determinate caratteristiche.

- Devono essere FAIR

Findable

- F1. Utilizzo di Identificativi univoci e persistenti
- F2. Uso di un set di metadati arricchito
- F3. I metadati comprendono anche l’identificativo univoco dei dati che descrivono
- F4. (meta)dati vengono registrati o indicizzati in una risorsa ricercabile (data repository e i data repository sono a loro volta registrati in cataloghi di repository es. re3data.org)

Accessible

- A1. I (meta)dati sono recuperabili attraverso identificatori utilizzando un protocollo di comunicazione standard
- A1.1.Il protocollo è aperto, libero e implementabile ovunque
- A1.2 Il protocollo permette la autenticazione e ove necessario procedure di autorizzazione
- A2**. I metadati sono sempre accessibili anche quando i dati non lo sono**

Interoperable

- **I1. I (meta)dati usano un linguaggio per la rappresentazione della conoscenza: formale, accessibile, condiviso e applicabile ad ampio spettro**
- I2 I(meta)dati usano vocabolari the seguono i principi FAIR
- I3. I (meta)dati comprendono il link qualificato ad altri metadati

Reusable

- R1. I (meta)dati sono descritti attraverso una **pluralità di attributi**
- R1.1. I (meta)dati sono rilasciati con una licenza di utilizzo chiara
- R1.2. Ai (meta)dati è associata la definizione chiara della **provenienza**
- R1.3. I (meta)dati seguono gli **standard rilevanti per la comunità di riferimento**

Findable

To aid automatic discovery of relevant datasets, (meta)data should be easy to find by both humans and machines and be assigned a persistent identifier.

Accesible

Limitations on the use of data, and protocols for querying or copying data are made explicit for both humans and machines.

Interoperable

(Meta)data should use standardised terms (controlled vocabularies), have references to other (meta)data and be machine actionable.

Reusable

(Meta)data are sufficiently well described for both humans and computers to be able to understand them and have a clear and accessible data usage license.

Findable

To aid automatic discovery of relevant datasets, (meta)data

ID persistenti

Accesible

Limitations on the use of data, and protocols for querying or

Criteri di accesso espliciti

Interoperable

(Meta)data should use standardised terms (controlled vocabularies).

Metadatazione standard

Reusable

(Meta)data are sufficiently well described for both humans and computers to be

Licenze esplicite

Documento della Swiss national science foundation

- Guida esplicativa sui principi FAIR applicati ai dati di ricerca:

http://www.snf.ch/SiteCollectionDocuments/FAIR_principles_translation_SNSF_logo.pdf

Sharing & Reusing Data

Ma è facile riutilizzare i dati?

- obvious: sharing is needed to enable new insights, but ...
- many are afraid of sharing beyond l&e aspects
  - it's about competition in science
  - sharing does not mean reusing
  - it's about having the skills & tools
- reusing data from others is hard – you need time, patience & skills
  - 80 % of time in data projects is wasted with "data wrangling"
  - data organisation and modelling is unspecified/different
  - metadata structure & element semantics not explicitly defined
  - data structure & element semantics not explicitly defined
  - much mapping/transformation/normalisation is required
  - who is able to develop the "wrangling" & analytics software????
  - who are the winners and who are the losers????

Da: Peter Wittenburg Max Planck Computing & Data Facility
(Presentazione presso ERC)

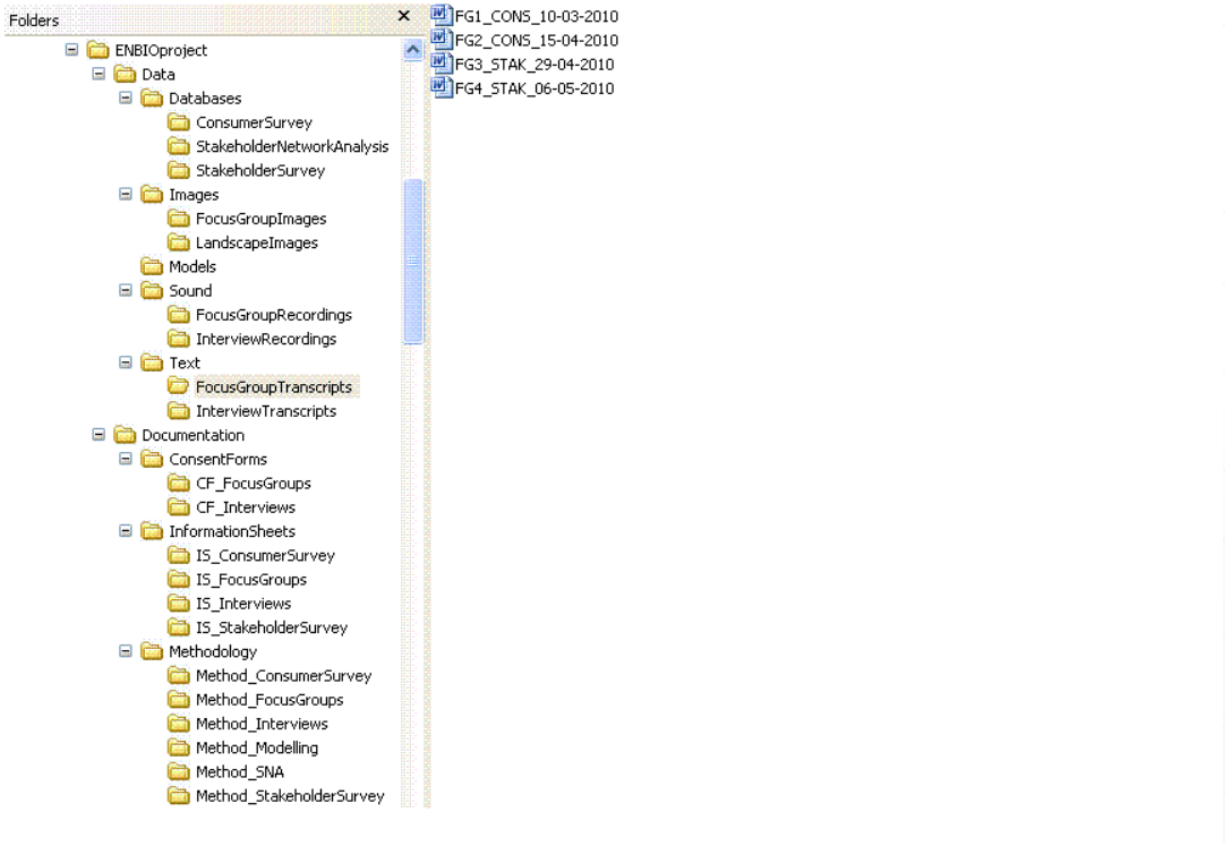
Punti di attenzione: formati

- Utilizzare formati open o standard per contrastare l'obsolescenza
- This typically means using open or standard formats (such as OpenDocument Format or ODF, ASCII, tab-delimited format, comma-separated values or XML) instead of proprietary ones. Some proprietary formats (such as Microsoft Rich Text Format, Microsoft Excel and SPSS) are widely used and likely to be accessible for a reasonable, but not unlimited, time. (UK Dataservice)
- I dati in formati proprietari per poter essere conservati dovrebbero essere convertiti in formati aperti o standard. Meglio se dai ricercatori stessi che ne possono garantire l'integrità

Formati raccomandati

- Da UK dataservice
- https://www.ukdataservice.ac.uk/manage-data/format

Punti di attenzione: struttura delle cartelle



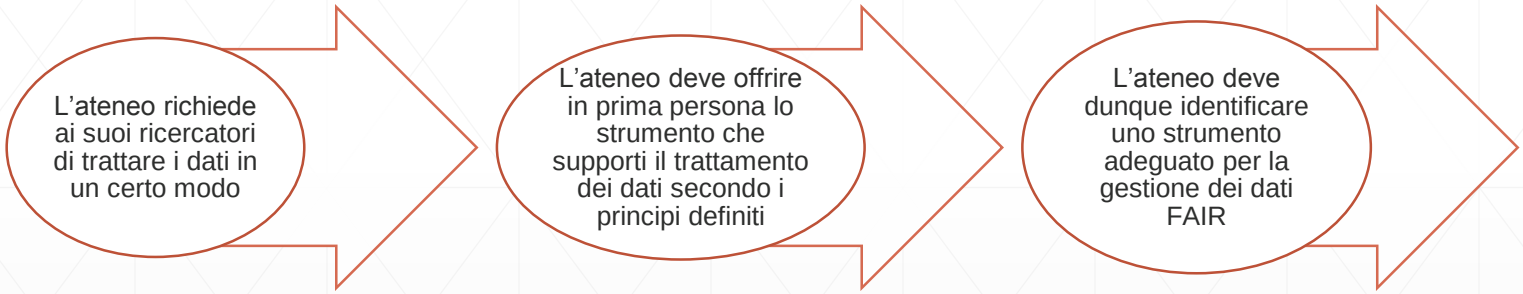
Punti di attenzione: nomi dei file

- create meaningful but brief names
- use file names to classify types of files
- avoid using spaces, dots and special characters (& or ? or !)
- use hyphens (-) or underscores (_) to separate elements in a file name
- avoid very long file names
- reserve the 3-letter file extension for application-specific codes of file format (e.g. .doc, .xls, .mov, .tif)
- include versioning within file names where appropriate

Punti di attenzione: versioning

file name	Changes to file
Interviewschedule_ 1.0	Original document
Interviewschedule_ 1.1	Minor revisions made
Interviewschedule_ 1.2	Further minor revisions
Interviewschedule_ 2.0	Substantive changes

La scelta dello strumento: repository



La scelta dello strumento

- Non ha senso attendere la definizione di un servizio a livello nazionale
- Può esserci un collettore centrale, ma la gestione dovrebbe essere fatta dall'Ateneo
- Zenodo : permette l'archiviazione di più versioni dei dataset, permette una corretta metadatazione e documentazione dei file, non permette di parametrizzare e controllare gli accessi
- Figshare : è uno strumento proprietario, se domani venisse chiuso non abbiamo la certezza di recuperare il controllo sui nostri dati

La scelta dello strumento

- Strumento Open Source
- Strumento controllabile direttamente dall’amministrazione
- Strumento che risponda ai requisiti FAIR