

# Research Data Management: Why?

*Emma Lazzeri*

ORCID iD: <https://orcid.org/0000-0003-0506-046X>

Istituto di Scienza e Tecnologie dell'Informazione  
Consiglio Nazionale delle Ricerche



Corso di formazione  
Praticare l'Open Science nelle scienze della Terra e dell'ambiente  
26 Novembre 2020 | Modulo 2



# Picture or...it didn't happen!



- How many pictures in you mobile?
- Would you consider your pictures as “your data”?
- Do you back them up regularly?

**A picture is worth a thousand words**





What is Research Data?

”


Data are  
facts, observations or experiences  
on which an argument or theory is  
constructed or tested.

Data may be numerical, descriptive, aural or  
visual. Data may be raw, abstracted or analysed,  
experimental or observational.

**Data or...it didn't happen!**

UCL Research Data Policy

<https://www.ucl.ac.uk/library/research-support/research-data-management>



What happens if  
you do not properly  
manage and share  
your research  
data?

# Data can get lost!

## JAMA journal retracts paper when author can't produce original data

In July 2017, a *JAMA* journal called for an investigation into a 2013 paper it had published after concluding that the article had “scientific and ethical concerns.” Now the journal, *JAMA Otolaryngology – Head & Neck Surgery*, is retracting the paper.

The article, “Dexamethasone for the prevention of recurrent laryngeal nerve palsy and other complications after thyroid surgery: a randomized double-blind placebo-controlled trial,” came from a group in Italy led by Mario Schietroma, of the Department of Surgery at the University of L'Aquila, in Abruzzo, Italy. Schietroma, who in December admitted to us that a retracted 2015 paper of his in the *Journal of the American College of Surgeons* suffered from “misinterpretation of the statistical data,” now has four retractions.



*“Neither [the original dataset and the approved protocol] have been provided by Dr Schietroma, and the university has informed us that “without those pieces of information the results of the papers under investigation cannot be validated.”*

<https://retractionwatch.com/2018/10/25/jama-journal-retracts-paper-when-author-cant-produce-original-data/>

# Depositing your data is of highest importance!!!

Carlisle has kept going. This year, he warned about dozens of anaesthesia studies by an Italian surgeon, Mario Schietroma at the University of L'Aquila in central Italy, saying that they were not a reliable basis for clinical practice<sup>6</sup>. Myles, who worked on the report with Carlisle, had raised the alarm last year after spotting suspicious similarities in the raw data for control and patient groups in five of Schietroma's papers.



Bottled oxygen, used by anaesthetists during surgery. Credit: Mark Thomas/Alamy

The challenges to Schietroma's claims have had an impact in hospitals around the globe. The World Health Organization (WHO) cited Schietroma's work when, in 2016, it issued a recommendation that anaesthetists should routinely boost the oxygen levels they deliver to patients during and after surgery, to help reduce infection. That was a controversial call: anaesthetists know that in some procedures, too much oxygen can be associated with an increased risk of complications – and the recommendations would have meant hospitals in poorer countries spending more of their budgets on expensive bottled oxygen, Myles says.

The five papers Myles warned about were quickly retracted, and the WHO revised its recommendation from 'strong' to 'conditional', meaning that clinicians have more freedom to make different choices for various patients. Schietroma says his calculations were assessed by an independent statistician and through peer review, and that he purposely selected similar groups of patients, so it's not surprising if the data closely match. He also says he lost raw data and documents related to the trials when L'Aquila was struck by an earthquake in 2009. A spokesperson for the university says it has left enquiries to "the competent investigating bodies", but did not explain which bodies those were or whether any investigations were under way.

# Data can be manipulated



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

# Research Integrity: we have a problem

58 articles published by Diederik Stapel were withdrawn because they were **based on invented data**.

His papers had been published in scientific journals considered prestigious (very high IFs!).

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

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

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

## REPORT

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

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

+ See all authors and affiliations

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

Article

Figures & Data

Info & Metrics

eLetters

PDF

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

## Abstract

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

<http://www.insidehighered.com/news/2011/11/28/scholars-analyze-case-massive-research-fraud>



# Data can be altered

The drug was widely used, also thanks to the fact that it had been widely studied.

However some of the data on which the studies were based had been altered and the benefits w.r.t. other similar drugs were not proved.

Retracted

## RETRACTED: Influence of a new hydroxyethylstarch preparation (HES 130/0.4) on coagulation in cardiac surgical patients

Article in [Journal of Cardiothoracic and Vascular Anesthesia](#) 15(3):316-21 · July 2001 with 22 Reads ⓘ

DOI: [10.1053/jcan.2001.23276](#) · Source: [PubMed](#)

[Cite this publication](#)



Günther Haisch



Joachim Boldt



Claudia Krebs

+ 2



Stephan Suttner

[Show more authors](#)

### Abstract

To compare volume therapy with HES 130/0.4, a new hydroxyethylstarch (HES) solution with a gelatin-based fluid replacement strategy. Prospective, randomized, safety study. Urban, university-affiliated hospital (single institution). Forty-two patients undergoing elective cardiac surgery. Patients were prospectively randomized into 2 groups: In group 1 (n = 21), gelatin was given perioperatively for volume support until the 1st postoperative day to keep the central venous pressure (CVP) between 10 and 14 mmHg; in group 2 (n = 21) HES 130/0.4 was administered using the same protocol as in group 1. Standard coagulation variables and modified thromboelastography (TEG) were used. Using different activators for extrinsic and intrinsic activation and heparin inactivation by heparinase, the onset of coagulation (coagulation time), kinetics of clot formation (clot formation time), and maximum clot firmness were measured. Measurements were performed after induction of anesthesia (T0), at the end of surgery (T1), 4 hours after surgery (T2), and on the morning of the 1st postoperative day (T3). A total of 3310 +/- 810 mL of gelatin and 3070 +/- 570 mL of HES 130/0.4 were used in the 2 groups during the study period. The 2 groups did not differ with regard to postoperative bleeding or in use of packed red blood cells or

# Data are used by policy makers

The make decisions also based on research evidence

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



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

## Covid-19 - Situazione in Italia



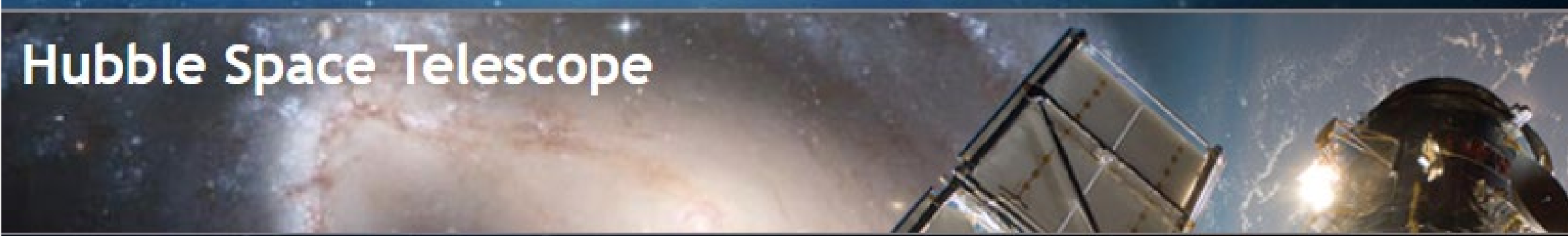
[Home](#) / [Emergencies](#) / [Diseases](#) / [Coronavirus disease \(COVID-19\)](#)

## Coronavirus disease (COVID-19) pandemic

[Advice for the public](#)

[Country and technical guidance](#)

# Data can be useful to others



## Hubble Space Telescope

### News

Text Size + -

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

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

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

**Exoplanet HR 8799 System**



”

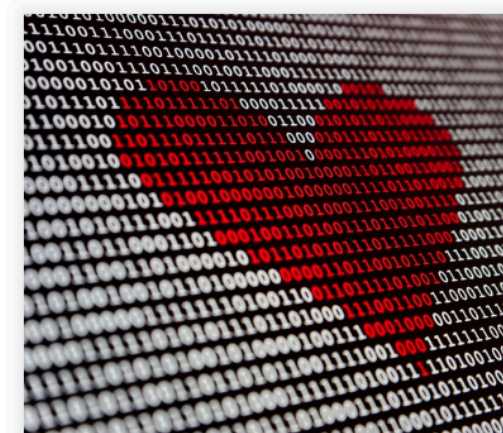
Data is not yours...that is  
why you should take care  
of it!

# Data is not yours!

- Data is **not** intellectual work, it is fact and information
- Copyright protection covers expressions and not ideas, procedures, operating methods or mathematical concepts as such.
- Protection is on databases and not on data. The data are protected only and especially when they are collected and organized in a database.
- The sui generis property right (only in Europe) covers not only the reproduction and dissemination of the database, but also the extraction and reuse of substantial parts of the database.

Read complete article [here](#).

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



## L'autore



### Simone Aliprandi

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

Sito e blog >

11 Dicembre 2019

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

di Simone Aliprandi

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

CONDIVIDI



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

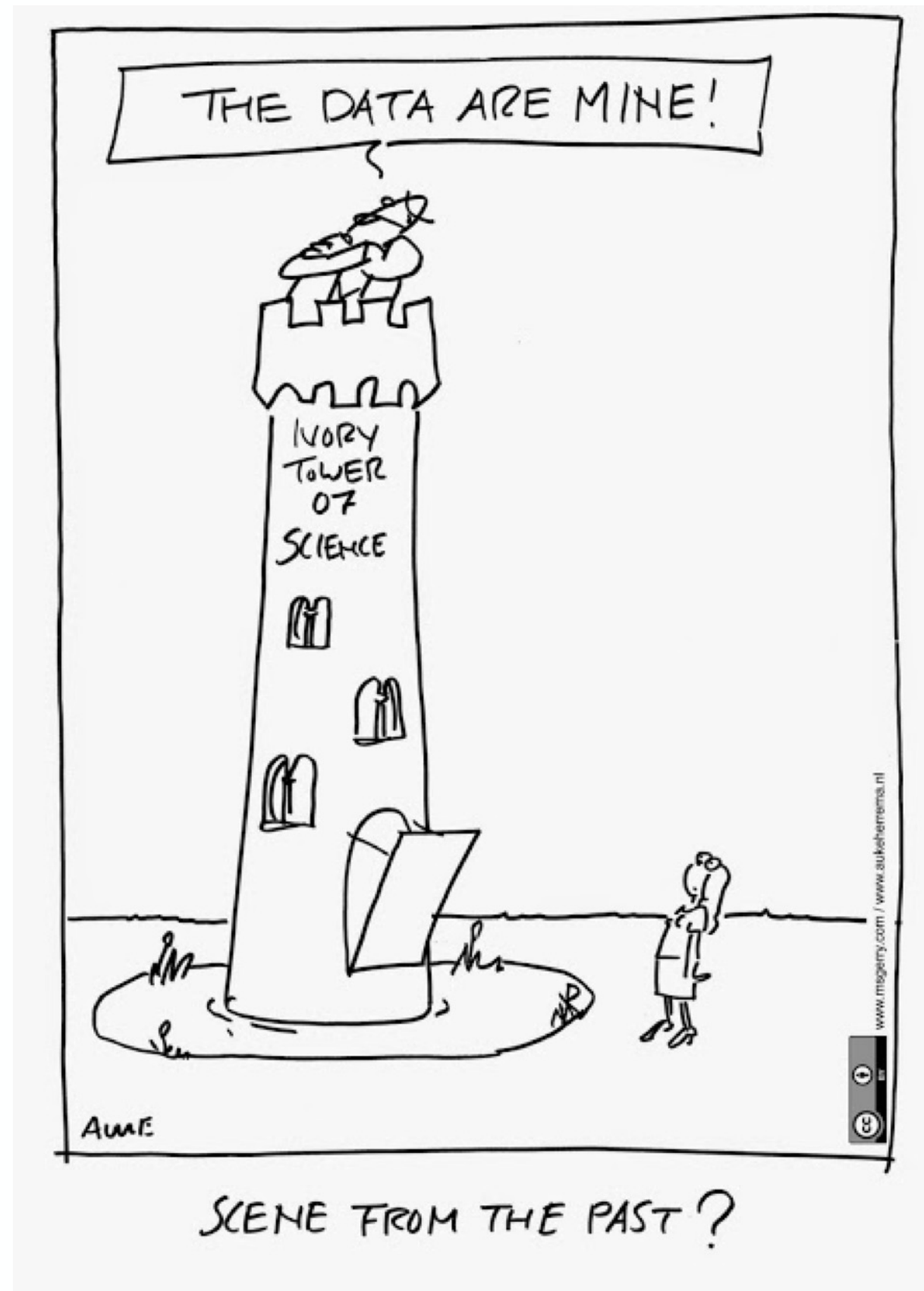
Si sente spesso parlare di *tutela del dato* o *titolarità del dato*, soprattutto in questi ultimi due/tre anni in cui temi come big data e open data sono diventati di pubblico dominio e in cui l'entrata in vigore del GDPR (il nuovo regolamento europeo sui dati personali) ha portato un'ondata – per certi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

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

Per inquadrare il tema correttamente e in modo completo, è tuttavia necessario tenere in debita considerazione anche il punto di vista della cosiddetta proprietà intellettuale, punto di vista che a me sta particolarmente a cuore. Anche perché quando acquisiamo, gestiamo, diffondiamo dei dati **non è detto che siano dati personali** e dunque non sempre le norme sulla privacy (GDPR e simili) entrano in gioco. Cerchiamo di capire meglio la questione.

## Nessun copyright su idee e dati

Innanzitutto: **non esiste alcun diritto di proprietà intellettuale sul dato in sé**. I dati nudi e crudi e le informazioni che da essi si deducono non sono oggetto di alcun tipo di proprietà intellettuale. Questo in virtù di uno dei principi cardine del diritto d'autore secondo cui il diritto tutela non l'informazione, bensì la specifica forma espressiva con cui l'informazione è presentata. Basti leggere il testo dell'articolo 9, numero 2 dell'Accordo TRIPS:



# Be aware...

---

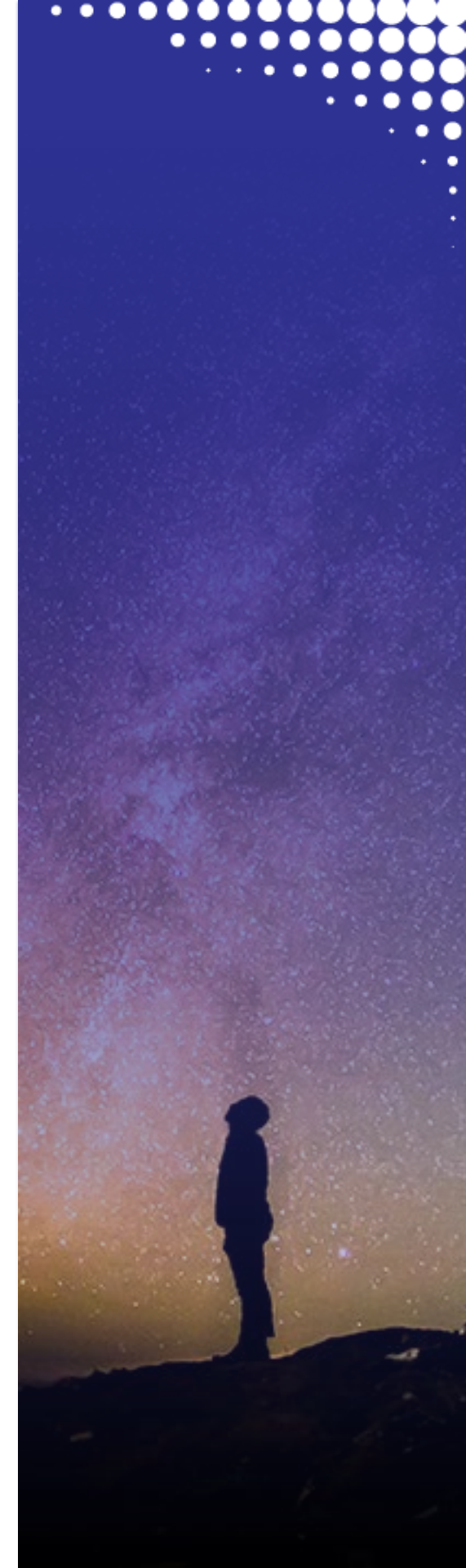
Managing data is not  
easy

No «one size fits all»

- 
- It takes a lot of time to manage data

- 
- So many aspects to be considered
    - Several tools to use

- 
- But, hey, BENEFITS are enormous!



# What is the cost of not managing data properly?

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

Published: 2019-01-16

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



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

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



What if  
**your data**  
gets lost forever?

# Backup is that thing that you should have done before you needed it

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



# Data is so fragile

## Therefore you need a Research Data Management!

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

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

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

NATURE | NEWS

### Scientists losing data at a rapid rate

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

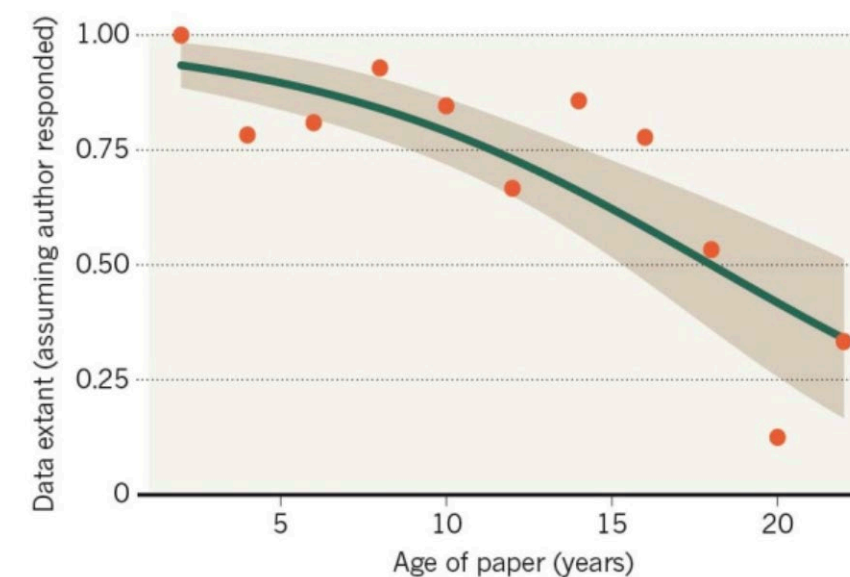
Elizabeth Gibney & Richard Van Noorden

19 December 2013

[Rights & Permissions](#)

#### MISSING DATA

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





Why should I care?

# Because...

- If you manage it, you probably will not **lose** it
- Organising your data will make your work more **efficient**
- Some data is **unique and not reproducible** (meteorology, observation from the field) so you should take care of it
- By correctly managing your data, you can improve **research integrity**
- By managing your data, you enable **validation and control**
- Someone else could use it in the future to **advance scientific progress**

# The Stick and the Carrot of RDM

## Requirements

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

## Benefits

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

So...

What is Research  
Data  
Management?

”

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

[How and why you should manage your research data: a guide for researchers](#)

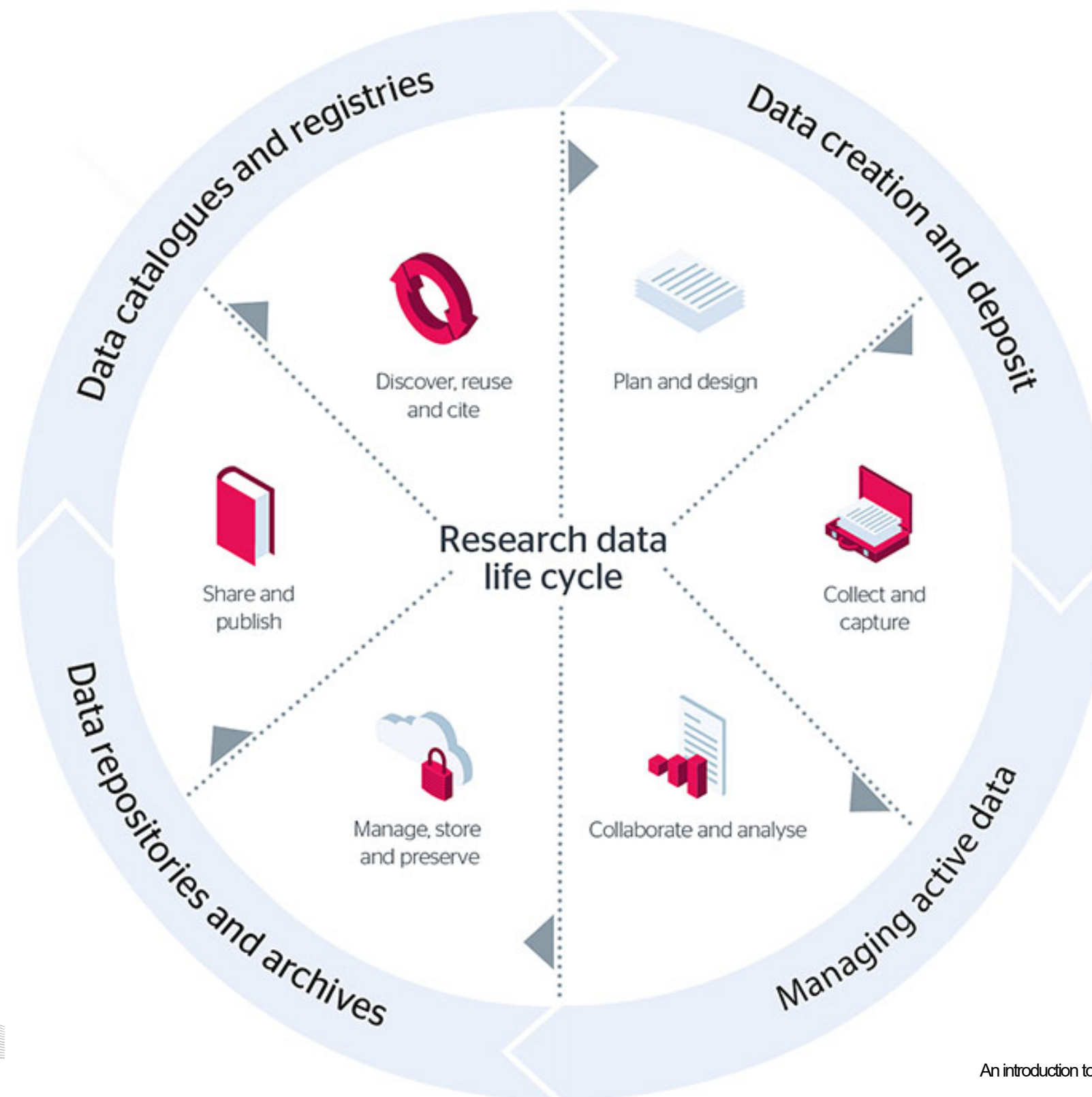
[An introduction to engaging with research data management processes.](#)

[Caroline Ingram](#), JISC Guides

# Do not harm...

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

# Research Data Lifecycle



# Data must be FAIR

## Findable

it should be clear where data are located

## Interoperable

(machine) readable with free software

## Accessible

for at least 10 years

## Reusable

usable for re-analysis or new research

***FAIR does not mean  
OPEN!***

# Open Access to Research Data

- Storage: data must be deposited in a «trusted» repository
  - **The best option is to deposit in disciplinary or thematic repositories, if available**
    - **Zenodo** it's the best solution because it's already connected with OpenAIRE
    - Disciplinary repositories can be found on **Re3data** ([www.re3data.org](http://www.re3data.org))
    - Provide **full metadata**, according to discipline standards.
- OPEN ACCESS: data must be opened (not all data can be granted **open** access)
  - An **embargo** period can be imposed, depending on the project
  - Choose the most open CC license possible to enable reuse
- DOCUMENTATION: all useful information about the gathering of the data must be provided, and possibly DEPOSIT the dataset.



# Fifty shades of no (to data sharing)

Too expensive

There's no business case

There's no commercial value

It's private

It's secret

It's our data

We have invested a lot of money in this

Link enough data and one will arrive at sensitive private information

It's not data, it's information

It will never work

We don't know how to do this

We don't have the right people to do this

We need the money

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

No idea what the quality of the data is

We don't know where to find it

It's not our job

It isn't in the right format

I am not authorised

Who is going to use this anyway

People are going to misuse it

Image damage for the minister

We are not ready for this

Image loss for Government

The data file is too big

Not enough bandwidth

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

We can't find it

We have no access

It is out of date / too old

We have no time

We don't know if it's worth it

Management says no

We never did this before

No value in it

No time / no resources

We will open up (but adapt 90%)

It's incorrect

Commercially sensitive

It is dangerous when linked

People are going to make the wrong conclusions

This is going to start a wrong discussion

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

We know the data is wrong, and people will tell us where it is wrong, then we'd waste resources inputting the corrections people send us

Our IT suppliers will charge us a fortune to do an ad hoc data extract

We have to be careful with existing contracts

Our website cannot hold files this large

It's not our data, it's someone else's

We don't have the right people to do this

We need the money

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

No idea what the quality of the data is

We don't know where to find it

It's not our job

It isn't in the right format

I am not authorised

Who is going to use this anyway

People are going to misuse it

Image damage for the minister

We are not ready for this

Image loss for Government

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

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

People would get upset

It's very sensitive information

We are not ready for this

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

# What can you do?

To correctly manage your research data



# What are your options?

## Your community is already part of a RI

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

## Your Institution could set up a data strategy

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

## Your community could be part of a RI

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

## None of the above

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

**Do not reinvent the wheel**

**Build on others' experience**

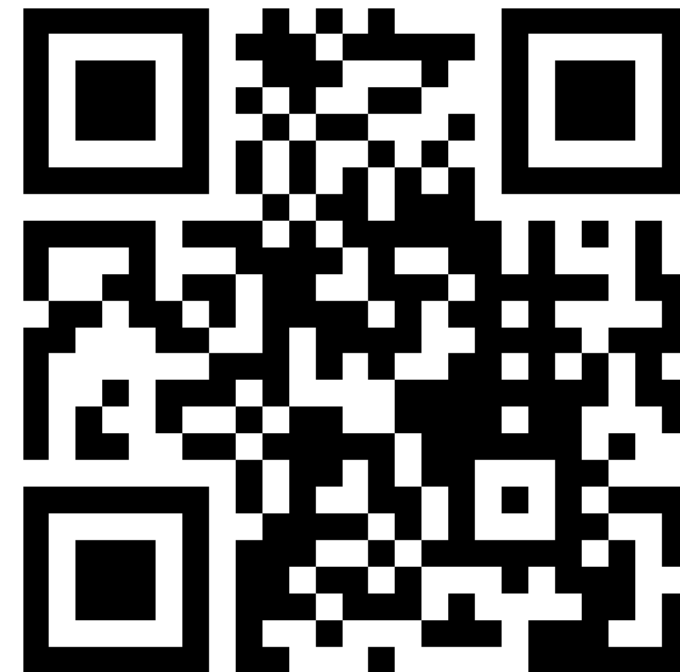
**Liaise with your and other communities**



# Mentimeter

You can **access** mentimeter from any device (mobile pc, tablet...)

- Go to [www.menti.com](https://www.menti.com) and enter code: **17 62 71**
- Click on the direct link: <https://www.menti.com/663okc6efu>
- Scan the QR code



# A good example: ISTI Open Portal

[menti.com](https://menti.com)

code:  
**17 62 71**



- The CNR Institute of Information Science and Technologies (ISTI) set up an **Open Access Working Group** (OA-WG) with all the involved stakeholders
- An **OA policy** was drafted by the WG and adopted by the Institute in 2018
- A **strategy** to develop and maintain an institutional «repository» was designed
- A **dedicated tool** (**ISTI Open Portal**) was developed to extract metadata and full text from CNR (closed) literature archive (PEOPLE)
- The OA WG now **runs and curate the ISTI Open Portal** and **support** the users on OA related issues
- Now the Open Portal is **registering in OpenAIRE** to become the «official» ISTI repository

**This solution is ready to be adopted by any other CNR institute!**



A gateway to the scientific production of the Institute of Information Science and Technologies (ISTI), an institute of the Italian National Research Council (CNR) (more in [About](#))

search for documents...



13,832

Research products

1,587

Authors

19

Laboratories

236

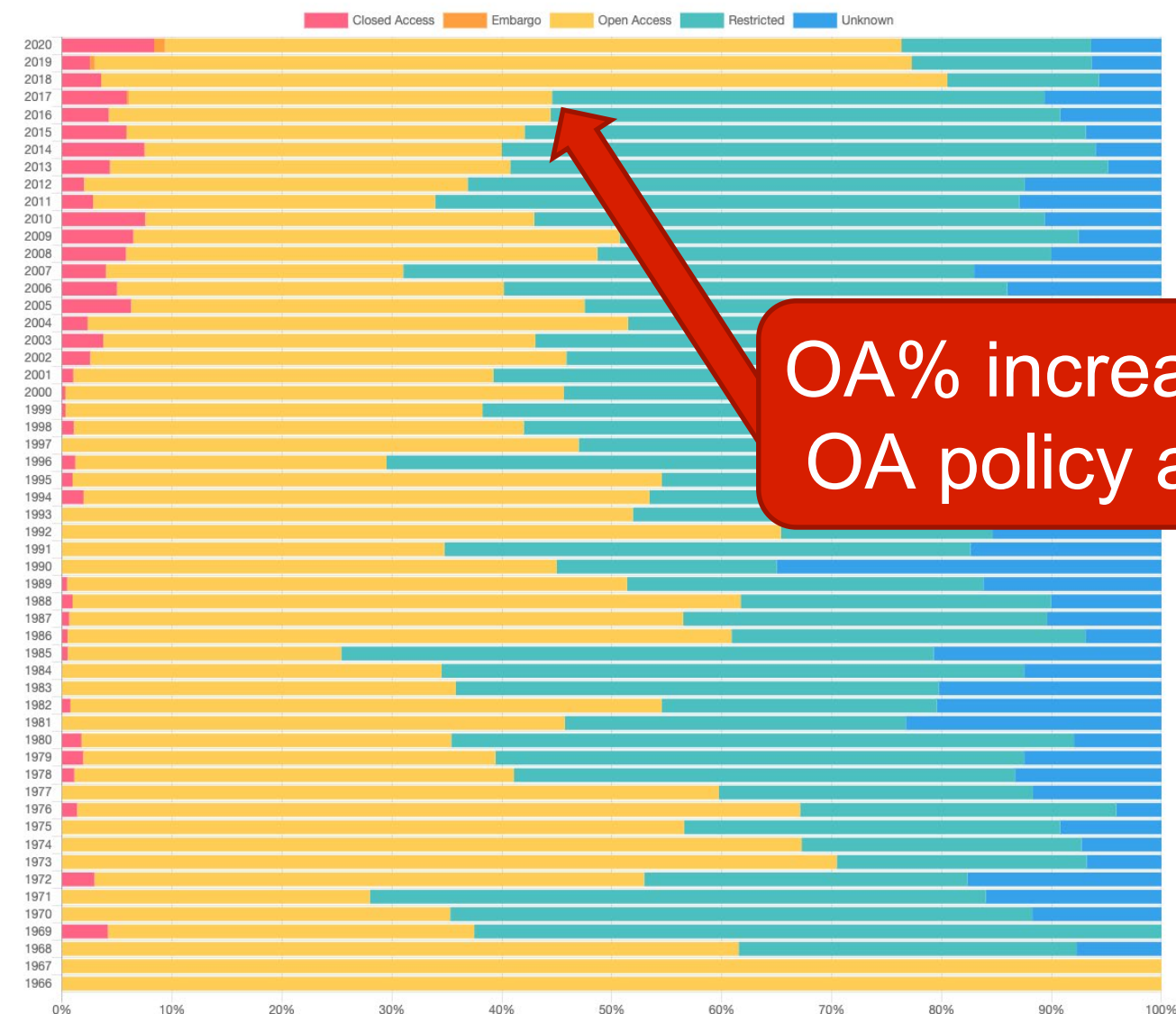
Projects

5,980

Open Access products

## Product Access Rights by Year

Graphs Table



OA% increase after OA policy adoption

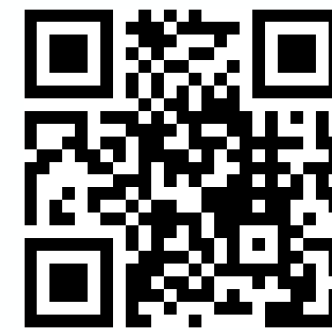
Content is enriched by OpenAIRE graph

Dedicated support!

menti.com

code:

17 62 71



203 result(s)

first « 1 2 3 ... » last

Page Size: 10, 20, 50  
Export: bibtex, xml, json, csv  
Order by: date

CNR Author operator: and / or

- ☐ Paolo Manghi (203)
- ☐ Alessia Bardi (61)
- ☐ Leonardo Candela (46)
- ☐ Donatella Castelli (42)
- ☐ Pasquale Pagano (39)
- ☐ Claudio Atzori (38)
- ☐ Marko Mikulicic (38)
- ☐ Michele Artini (37)
- ☐ Sandro Fabrizio La Bruzzo (36)
- ☐ Franco Zoppi (23)

more

2020 Conference object Open Access

### RepOSGate: Open Science Gateways for Institutional Repositories

Artini M., Candela L., Manghi P., Giannini S.

Most repository platforms used to operate Institutional Repositories fail at delivering a complete set of functionalities required by institutions and researchers to fully comply with Open Science publishing practices. This paper presents RepOSGate, a software that implements an overlay application ... [\[show more\]](#)

Source: ICRDL 2020 - 16th Italian Research Conference on Digital Libraries, pp. 151-162, Bari, Italy, 30-31 January 2020

DOI: 10.1007/978-3-030-39905-4\_15

DOI: 10.5281/zenodo.3819830

DOI: 10.5281/zenodo.3819829

Project(s): OpenAIRE-Advance

See at: [ISTI Repository](#) | [ZENODO](#) | [Unknown Repository](#) | [Unknown Repository](#) | [link.springer.com](#) | [Unknown Repository](#) | [CNR\\_OAPub](#)

2020 Article Open Access

### Entity deduplication in big data graphs for scholarly communication

Manghi P., Atzori C., De Bonis M., Bardi A.

See at: [ISTI Repository](#) | [CNR\\_OAPub](#) | [www.emerald.com](#) | [Data Technologies and Applications](#) | [Data](#)

Typology operator: and / or

- ☐ Report (63)
- ☐ Conference object (58)
- ☐ Article (47)
- ☐ Other (17)
- ☐ Part of book or chapter of book (11)
- ☐ Book (6)

## Contact us

Need help or have questions about ISTI Open Portal? Please [open a ticket](#)

Support topics include the following:

- ISTI Open Access Policy;
- Archiving policy;
- Use of Open ISTI Portal;
- Open Access - practice and legislation in force
- verification of the editorial policies related to the individual products
- relations with the CNR Information Systems Office concerning
  1. technical and functional aspects of PEOPLE
  2. queries about the bibliographic/management aspects of the products
- contacts with publishers

# Thank you!

**Emma Lazzeri**

[emma.lazzeri@isti.cnr.it](mailto:emma.lazzeri@isti.cnr.it)

This work was partially supported by European Union's Horizon 2020 under projects grant Agreement numbers 857650, and 777541 , and by **ICDI** (Italian Computing and Data Infrastructure)

