

Ciencia Abierta y Reproducibilidad Científica

Susana Sánchez Expósito, Lourdes Verdes-Montenegro, Julian Garrido, Laura Darriba, Javier Moldón, Manuel Parra, MªAngeles Mendoza

Instituto de Astrofísica de Andalucía (CSIC)

Marzo 2022



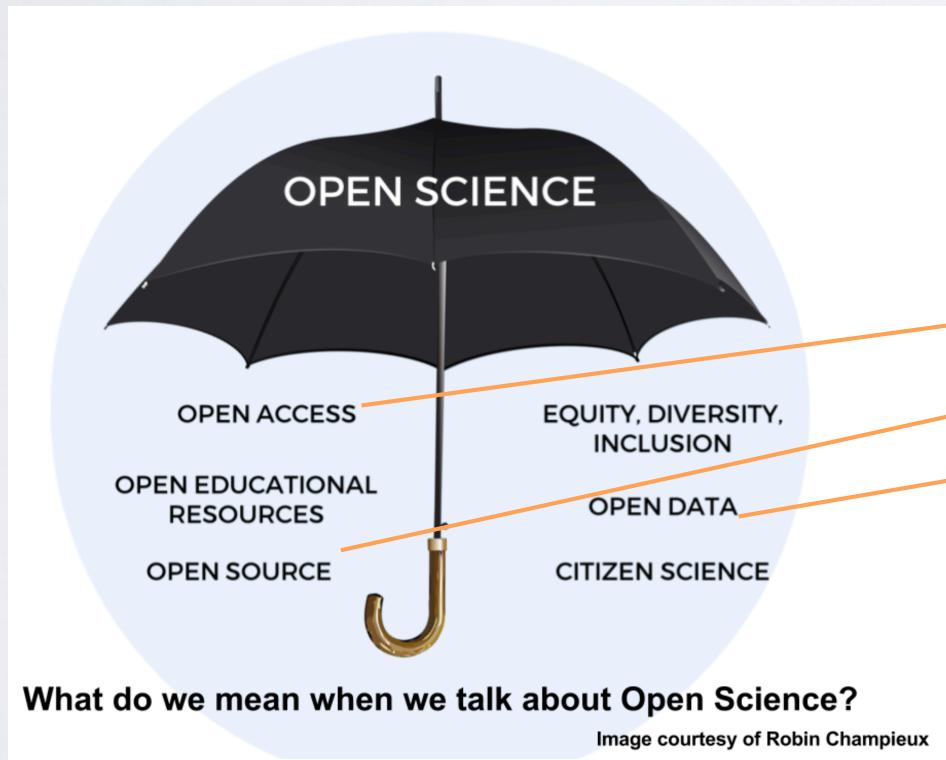
Instituto de Astrofísica de Andalucía, IAA-CSIC



Open Science Definition

Open Science represents an approach to research that is collaborative, transparent and accessible

[Open Science definition](#), European Commission, 2017, doi: 10.2777/75255



Improving the access to the elements involved in a scientific study:

- - Article
- - Method / software tools
- - Data
- - Etc.

Beyond the PDF initiative

Knowledge Burying in paper publication

Research Objects: Towards Exchange and Reuse of Digital Knowledge, S. Bechhofer et al. 2011

"New mechanisms are needed that will allow us to share, exchange and reuse digital knowledge"



Need of exposing the complete scientific record, not the story and in a way the experiment can be **discovered** and **understood**

MOVING FROM NARRATIVES (LAST 300 YRS)
TO THE ACTUAL OUTPUT OF RESEARCH



Open Science Definition

Open Science is the practice of science in such a way that **others can collaborate and contribute**, where research data, lab notes and other research processes are freely available, under terms that enable **reuse, redistribution and reproduction** of the research and its underlying data and methods.

Foster project, <https://www.fosteropenscience.eu/>



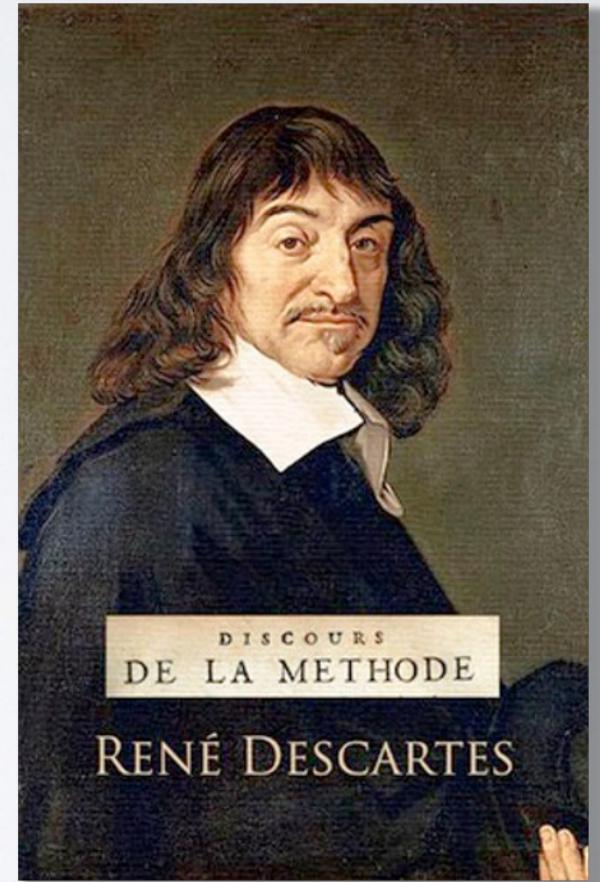
Open Science: a new concept?

Too many **adjectives** for science:

excellent, high quality, trustable, reproducible ... Open

Let's go back 383 years in time...

Scientific Reproducibility is a fundamental principle of the Scientific Method, a process established in the 17th century that marked the beginning of modern science and laid the foundations for the Philosophy of Science



Crisis of the scientific reproducibility



Questionnaire on reproducibility (1500 scientists)

- 70% of researchers have tried and failed to reproduce another scientist's experiments
- > 50% have failed to reproduce their own ones!
 - Chemistry: 90% (60%)
 - Biology: 80% (60%)
 - Physics and engineering: 70% (50%)
 - Medicine: 70% (60%)
 - Earth and environmental science: 60% (40%)

Baker (2016) <https://doi.org/10.1038/533452a>



Overly Honest Method
@OverlyHonestly



You can download our code from the URL supplied. Good luck downloading the only postdoc that can get it to run, though #OverlyHonestMethods



Reproduce an experiment is not so easy

- Original data are not publicly available
- They are available but not in an automatic way
- Processed data is only available in the published PDF
- There are some scripts for processing the data on a server somewhere, but no one remembers where
- The code is in a public repository, but good luck trying to install/execute it.

OPEN SCIENCE IS A SOLUTION FOR THOSE PROBLEMS



FAIR principles and Open Science

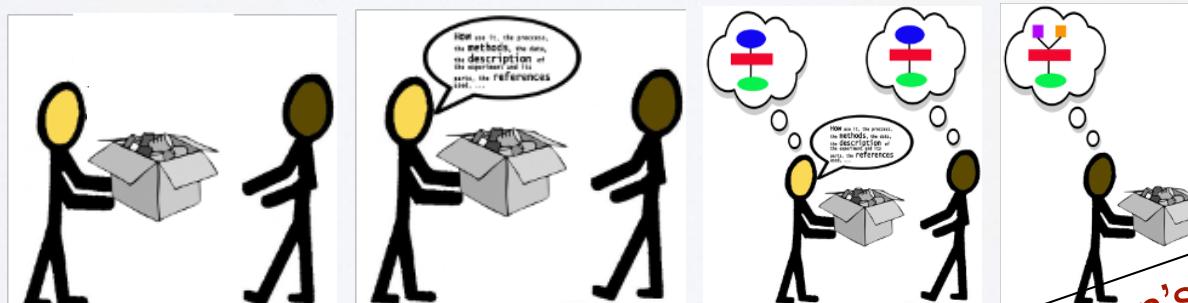
THE AVAILABILITY IS NOT ENOUGH → FAIR - ABILITY

Findable: data/methods can be discovered in an automatic way

Accesible: Available/stored in repositories/catalogues

Interoperable: Based on standards

Reusable: provenance information & licenses



See Julian's Talk in next session

As open as possible

- Embargo period compliant with Open Science
- Open Source is not Free Software
- Licenses for using scientific resources: a key tool in Open Science

The screenshot shows a Zenodo dataset page for "BIP4COVID19: Impact metrics and indicators for coronavirus related publications". The page includes a search bar, upload and community links, and log-in/signup buttons. Key statistics are displayed: 210,914 views and 30,125 downloads. The dataset is indexed in OpenAIRE. The "Dataset" tab is selected, and the "Open Access" button is visible. The main content area describes the dataset's purpose, contributors, and data sources. It details various impact measures like Influence, Popularity, and Popularity alternative, along with their descriptions and methods. A "Social Media Attention" section mentions a dataset of tweets. At the bottom, a "License (for files)" section specifies "Creative Commons Attribution 4.0 International".

zenodo

Search Log in Sign up

Upload Communities

March 19, 2022

Dataset Open Access

210,914 30,125

views downloads

See more details...

Indexed in

OpenAIRE

BIP4COVID19: Impact metrics and indicators for coronavirus related publications

Thanasis Vergoulis; Ilias Kanellos; Serafeim Chatzopoulos; Danae Pla Karidi; Theodore Dalamagas

This dataset contains impact metrics and indicators for a set of publications that are related to the COVID-19 infectious disease and the coronavirus that causes it. It is based on:

1. The CORD-19 dataset released by the team of Semantic Scholar¹ and
2. The curated data provided by the LitCovid hub².

These data have been cleaned and integrated with data from COVID-19-TweetIDs and from other sources (e.g., PMC). The result was a dataset of 503,162 unique articles along with relevant metadata (e.g., the underlying citation network). We utilized this dataset to produce, for each article, the values of the following impact measures:

- **Influence:** Citation-based measure reflecting the total impact of an article. This is based on the PageRank³ network analysis method. In the context of citation networks, it estimates the importance of each article based on its centrality in the whole network. This measure was calculated using the PaperRanking (<https://github.com/divis/PaperRanking>) library⁴.
- **Influence_alt:** Citation-based measure reflecting the total impact of an article. This is the Citation Count of each article, calculated based on the citation network between the articles contained in the BIP4COVID19 dataset.
- **Popularity:** Citation-based measure reflecting the current impact of an article. This is based on the AttrRank⁵ citation network analysis method. Methods like PageRank are biased against recently published articles (new articles need time to receive their first citations). AttrRank alleviates this problem incorporating an attention-based mechanism, akin to a time-restricted version of preferential attachment, to explicitly capture a researcher's preference to read papers which received a lot of attention recently. This is why it is more suitable to capture the current "hype" of an article.
- **Popularity alternative:** An alternative citation-based measure reflecting the current impact of an article (this was the basic popularity measured provided by BIP4COVID19 until version 26). This is based on the RAM⁶ citation network analysis method. Methods like PageRank are biased against recently published articles (new articles need time to receive their first citations). RAM alleviates this problem using an approach known as "time-awareness". This is why it is more suitable to capture the current "hype" of an article. This measure was calculated using the PaperRanking (<https://github.com/divis/PaperRanking>) library⁴.
- **Social Media Attention:** The number of tweets related to this article. Relevant data were collected from the COVID-19-TweetIDs dataset. In this version, tweets between 27/2/22-4/3/22 have been considered from the previous dataset.

We provide five CSV files, all containing the same information, however each having its entries ordered by a different impact measure. All CSV files are tab separated and have the same columns (PubMed_id, PMC_id, DOI, influence_score, popularity_alt_score, popularity_score, influence_alt_score, tweets count).

Publication date: March 19, 2022

DOI: DOI 10.5281/zenodo.6369427

Keyword(s): COVID-19, coronavirus, scientometrics, bibliometrics

Related identifiers: Cites <https://pages.semanticscholar.org/coronavirus-research> (Dataset) <https://github.com/divis/PaperRanking> (Software)

Supplement to www.biorxiv.org/content/10.1101/2020.04.11.037093v2 (Preprint)

Communities: Coronavirus Disease Research Community, COVID-19, Zenodo

License (for files): Creative Commons Attribution 4.0 International

Barriers to Open Science

Barriers to Data and Code Sharing in Computational Science

Survey of Machine Learning Community, NIPS (Stodden, 2010):

Code		Data
77%	Time to document and clean up	54%
52%	Dealing with questions from users	34%
44%	Not receiving attribution	42%
40%	Possibility of patents	-
34%	Legal Barriers (ie. copyright)	41%
-	Time to verify release with admin	38%
30%	Potential loss of future publications	35%
30%	Competitors may get an advantage	33%
20%	Web/disk space limitations	29%



Barriers to Open Science

Barriers to Open Science

- Lack of awareness and training
- Cultural inertia and misinformation
- Challenging the establishment
- Follow the status quo to succeed
- Perceived lack of reward
- Not considered for promotion
- Requires additional skills
- Takes time
- Publication bias towards novel findings



Fig: McKiernan <http://whyopenresearch.org>

Whitaker (2018) <https://doi.org/10.6084/m9.figshare.7140050.v2>

Dr. Rachael Ainsworth, JBCA



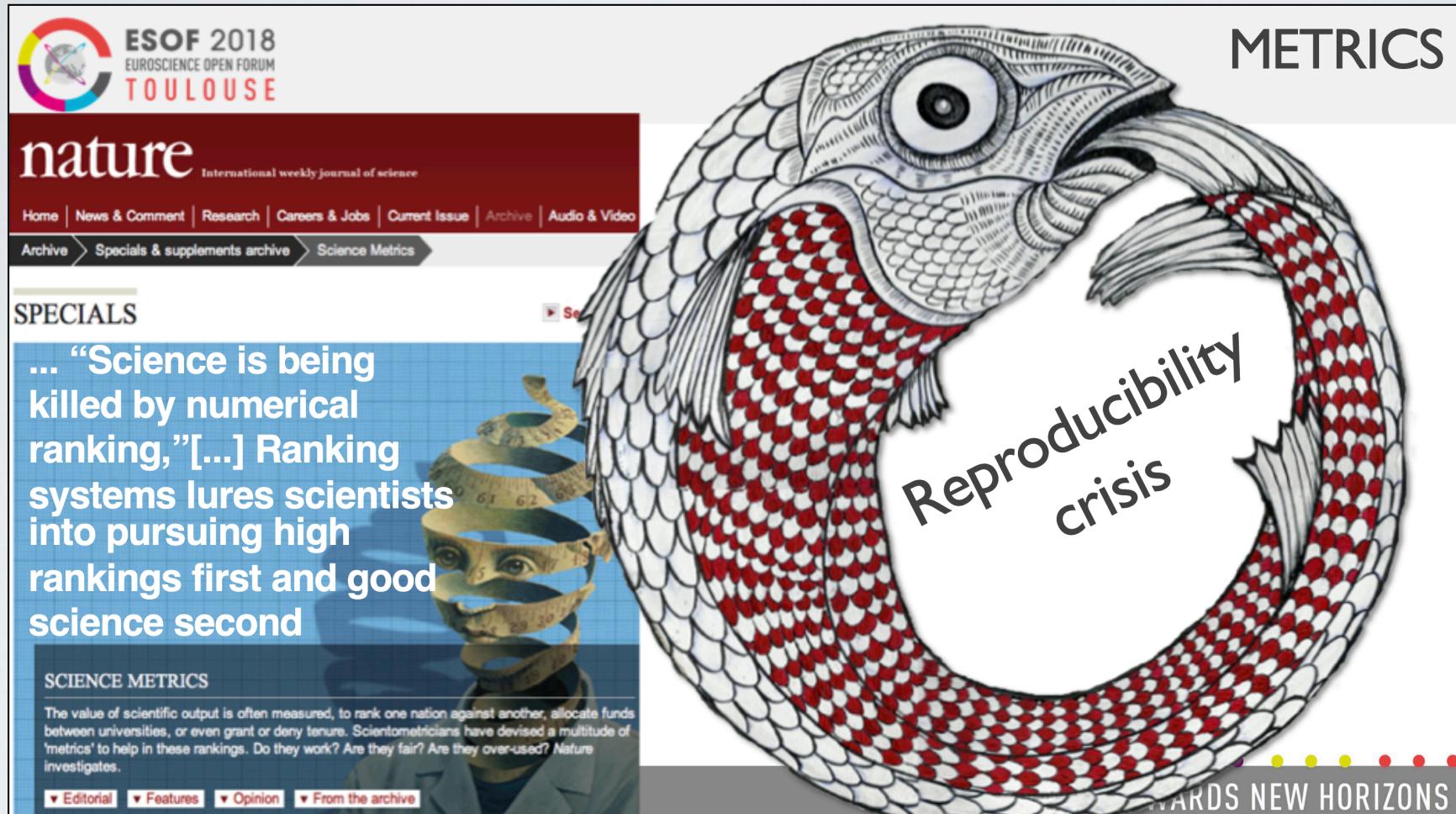
@rachaelevlyn #OpenScience #SKAscicon19

<https://doi.org/10.5281/zenodo.2631868>

R. Ainsworth, "Reproducibility & Open Science in the SKA Era". 2019 <https://doi.org/10.5281/zenodo.2631868>



Barriers to Open Science



L. Verdes-Montenegro, "Is the current measure of excellence perverting Science? A Data deluge is coming, it is time to act". ESOF 2018

Things are changing: DORA declaration. <https://sfdora.org>



Open Science



References

- **Love for science or “Academic Prostitution”?** L. Verdes-Montenegro.
<http://amiga.iaa.es/p/347-love-for-science-academic-prostitution.htm>
- **Reproducibility & Open Science in the SKA Era.** R. Ainsworth et al.
<https://doi.org/10.5281/zenodo.2631868>
- **Open Science for sustainability and inclusiveness: the SKA role model.** L. Verdes-Montenegro et al. <https://www.youtube.com/watch?v=ErNT9Va9vus>

Thanks!

