














Pathos

Open Science Impact Pathways

Deliverable 2.1 and 2.2

Open Science Indicator Handbook

Deliverable Number and Name	Deliverable 2.1 and 2.2 – Open Science Indicator Handbook
Due Date	31 August 2023
Delivery Date	31 August 2023
Work Package	WP2
Type	Handbook
Author	I. Grypari  (ARC) I. Karasz  (TGB) T. Klebel  (KNOW) E. Kormann  (KNOW) N. Manola  (ARC) H. Papageorgiou  (ARC) P. Stavropoulos  (ARC) L. Stoy  (TGB) T. van Leeuwen  (ULEI) T. Venturini  (CNRS) L. Waltman  (ULEI) T. Willemse  (ULEI) V.A. Traag  (ULEI)
Reviewers	Ioanna Grypari (ARC) Thomas Klebel (KNOW) Tommaso Venturini (CNRS) Vincent Traag (CWTS)
Approved by	Ioanna Grypari
Dissemination Level	PU - Public
Version	1.0

Number of Pages

243

The information in this document reflects only the author's views and the European Commission is not liable for any use that may be made of the information contained therein. The information in this document is provided "as is" without guarantee or warranty of any kind, express or implied, including but not limited to the fitness of the information for a particular purpose. The user thereof uses the information at his/ her sole risk and liability.



This project has received funding from the European Union's Horizon Europe framework programme under grant agreement No. 101058728. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Research Executive Agency. Neither the European Union nor the European Research Executive Agency can be held responsible for them.

Revision History

VERSION	DATE	REASON	REVISED BY
0.1	21-08-2023	First draft generated from source files	
1.0	31-08-2023	Reviewed and revised draft	All

Table 1: Document Revision History

Disclaimer

This document contains description of the PathOS project findings, work and products. Certain parts of it might be under partner Intellectual Property Right (IPR) rules so, prior to using its content please contact the consortium head for approval.

In case you believe that this document harms in any way IPR held by you as a person or as a representative of an entity, please do notify us immediately.

The authors of this document have taken any available measure in order to ensure that its content is accurate, consistent and lawful. However, neither the project consortium as a whole nor the individual partners that implicitly or explicitly participated in the creation and publication of this document hold any sort of responsibility that might occur as a result of using its content.

This publication has been produced with the assistance of the European Union. The content of this publication is the sole responsibility of the PathOS consortium and can in no way be taken as a reflection of the views of the European Union.

PathOS is a project funded by the European Union (Grant Agreement No 101058728).



OPEN SCIENCE INDICATOR HANDBOOK

Table of contents

Disclaimer.....	3
Table of contents.....	5
Introduction.....	8
Part I: Open Science Indicators.....	12
APC Costs.....	12
Availability of data repositories	18
Availability of preprint repositories	22
Availability of publication repositories.....	26
Citizen Science Indicators.....	30
Deposition of Open Metadata	39
Evaluation of Open Science in research assessment	42
Distribution of Open Access journal models.....	49
Open Access uptake.....	53
Open Science training facilities.....	59
Prevalence of national Open Science policies.....	63
Prevalence of open/FAIR data practices.....	70
Prevalence of Open Method practices.....	75
Prevalence of Open Peer Review.....	77
Prevalence of Open Science funding policies	81
Prevalence of Open Science support.....	90
Prevalence of preprinting.....	93
Prevalence of replication studies.....	98
Transformative publishing agreements.....	101
Part II: Academic Impact	105
Academic readership	106
Citation impact.....	108
Collaboration intensity	112
Diversity	114
Industry collaboration	117

Interdisciplinarity	119
Novelty	121
Societal collaboration	123
Topic trends.....	125
Use of code in research.....	127
Use of data in research.....	130
Use of methods in research.....	132
Use of patents in research	134
Part III: Societal Impact	136
Uptake in citizen science	137
Uptake in education.....	141
Science literacy.....	143
Uptake by policy makers	145
Uptake in the legal sector	147
Uptake in the public debate.....	149
Uptake in education.....	152
Uptake by patient groups	154
Uptake by civil society	156
Uptake by the general public	158
Uptake by policy makers	160
Effect on democracy.....	162
Effect on ethnic inequality	164
Effect on gender inequality	166
Effect on SDGs.....	168
Part IV: Economic Impact.....	170
Science-industry collaboration.....	171
Innovation output.....	174
Uptake of research result by industry.....	177
Socially relevant products and processes	180
Economic growth of companies.....	183

Labour market impacts	188
Cost savings.....	191
Part VI: Reproducibility	193
Consistency in reported numbers.....	195
Impact of Open Code in research	200
Impact of Open Data in research	207
Inclusion in systematic reviews or meta-analyses.....	219
Levels of replication found	223
Polarity of publications	229
Reuse of code in research	233
Reuse of data in research.....	239

Abbreviations

Abbreviation	Description	Explanation
APC	Article Processing Charge	Fee sometimes charged for publishing articles.
API	Application programming interface	Interface allowing to make use of certain service or program
DMP	Data Management Plan	Formalized document delineating the systematic approach to handling, storing, and disseminating research data both during the course of a research project and upon its conclusion.
DOI	Digital Object Identifier	Digital identifier of an object — physical, digital, or abstract. Core element of academic infrastructure. A type of persistent identifier (PID) (A long-lasting reference to a resource).
EOSC	European Open Science Cloud	An initiative aimed at developing an infrastructure to provide users with services promoting Open Science practices.
EU	European Union	Supranational political union consisting of member states that are located primarily in Europe
FAIR	Findability, Accessibility, Interoperability and Reusability	Principles for academic data management (https://www.go-fair.org/fair-principles/)
NCI	Normalised Citation Impact	Citation impact that is normalised for field differences
IPR	Intellectual Property Rights	Rights to intangible creations of the human mind.
NLP	Natural Language Processing	Computer techniques and algorithm to automatically process and analyse natural language
OA	Open Access	Principles and practices around distributing research outputs, free from any barriers.
OS	Open Science	Practices and principles for making science more open in various ways
OSF	Open Science Framework	Platform to support and enable collaboration
REST API	Representational state transfer API	Type of API often used for web services
UNESCO	United Nations Educational, Scientific and Cultural Organization	Specialized agency of the United Nations (UN) aimed at promoting world peace and security through international cooperation in education, sciences, culture, communication and information.

Executive Summary

This document outlines the primary components and objectives of the Open Science Indicators Handbook.

Objective and Scope: Deliverable D2.2 represents the inaugural release of the Open Science Indicators Handbook. This compilation encompasses definitive Open Science Indicators along with a preliminary framework for indicators on Academic, Economic, and Societal Impact and Reproducibility.

Integration of deliverable D2.1: The content and methodology from Deliverable D2.1, titled *A Data-Driven Methodology for Reproducibility Indicators*, has been incorporated into a dedicated section of this handbook on reproducibility indicators. The rationale behind this integration is to provide a singular, consolidated resource that enhances the clarity and coherence of information.

Accessibility and Continuous Improvement: To promote widespread accessibility and to foster ongoing refinement by the broader research community, the handbook will be hosted online on <https://handbook.pathos-project.eu>. It will be presented in a curated format, enabling sustained access and permitting iterative community-driven enhancements.

Introduction

This is the work-in-progress Open Science Indicator Handbook by PathOS. In this handbook we cover various indicators measuring various aspects around Open Science itself, their academic, societal and economic impacts, and reproducibility. The handbook will be made available openly as an easily navigable website on <https://handbook.pathos-project.eu>, in addition to being posted as a PDF on Zenodo.

The ultimate objective of the PathOS project is to develop a causal perspective on studying Open Science. This necessitates making a distinction between impact itself, and the effect of Open Science on impact. For instance, we could very well see an Open Source research tool being used frequently by industry. In that sense, the Open Source research tool can be said to have a type of economic impact. However, it could very well be that the research tool would have been similarly used by industry had it been released as closed software under a commercial licence. We are interested in the difference between its actual impact under the Open Science principles and its counterfactual impact under a closed principle. That is, we are interested in singling out the *causal* effect of Open Science within the more general impact of science.

Causal inference, however, is not straightforward. We will include an introductory chapter to explain some of the challenges around causal inference in Open Science. At the same time, this chapter also divulges the possibilities for inferring causality from observational data. Sometimes, we will learn that it will be impossible to correctly identify a certain causal effect. Although this limits our possible conclusions, we believe it is better to be clear about the impossibility of identifying a causal effect in some cases than to pretend we did identify some causal effect.

The challenge of causal inference also clarifies that we cannot provide straightforward indicators of effects of Open Science. There are many aspects of Open Science, including Open Access, Open Data and Open Code, but also elements such as Citizen Science, Open Science infrastructure, policies and training. In addition, there are many different types of impacts in each domain of academia, society and economy. This leads to a combinatorial number of possible effects of Open Science on any type of impact. The investigation of each of these effects necessitates a careful reflection about its causal inference, and what other factors should be controlled for, or should not be controlled for. This is a daunting task, which goes well beyond our capacities. Still in this handbook we can provide guidance on the first (and possibly the most important) step in this work of causal inference: the operationalisation of various indicators that can be used as steppingstones in studies on the effect of Open Science. We hope this is useful to the research community, and that we together face the challenges of causal inference in Open Science studies.

Not all indicators will be equally well-developed. Some indicators, like citation impact, are already long established and studied in scientometrics. Other indicators, such as on data usage or reproducibility are much more recent. Such indicators may be under active development, or may actually not be researched at all yet. We still include these understudied indicators in this handbook

when we believe them to be vital for studying the causal effect of Open Science. This indicator handbook should therefore not only be seen as an inventory of what is possible today, but also of what we believe is necessary tomorrow.

We hope this handbook will be a central hub to keep track of Open Science related indicators. At the moment, the handbook contains only a first draft of Open Science indicators and Reproducibility indicators. It also contains an outline of the various impact indicators that we intend to cover. Within the PathOS project the indicator handbook will remain in continuous development until January 2025. We hope to be able to contribute to keeping the indicator handbook up-to-date also afterwards. The handbook is open to community contributions. Together we can create a central resource that is useful to all.

Acknowledgements

Part I: Open Science Indicators

In the ever-evolving landscape of scientific research, the principles of Open Science have emerged as fundamental pillars supporting transparency, collaboration, and accessibility. Central to this paradigm is the need for tangible metrics to gauge and guide the adherence to these principles. This chapter on "Open Science Indicators" delves into the metrics and benchmarks designed to measure the breadth and depth of Open Science practices. Through these indicators, we aim to offer researchers, institutions, and policymakers a coherent framework to evaluate and enhance their commitment to making science more open, inclusive, and accountable.

APC Costs

History

Version	Revision date	Revision	Author
1.1	2023-08-28	Draft for initial publication	I. Grypari
1.0	2023-05-09	First draft	I. Grypari, N. Manola, H. Papageorgiou, P. Stavropoulos

Description

Article Processing Charges (APCs) represent the price that publishers demand from authors to pay in order to publish their articles and books under an open access license. They capture the affordability and accessibility of Open Access publishing for different types of stakeholders, such as researchers, institutions, and funding agencies. It is also relevant for policy-makers seeking to optimize Open Science policies.

Tracking and comparing APCs could also be used to encourage publishers to adopt more transparent and equitable pricing policies and support the development of sustainable Open Access publishing models accessible to all researchers regardless of their financial resources.

APCs have both benefits and drawbacks. In a strict sense, they do not remove the economic barriers between the writing and the reading of scientific results but shift these costs from the readers to the authors. In countries where funder reimbursement of APC costs or transformative agreements do not cover these costs, APCs can create a financial barrier that limits access to Open Access journals, often generating asymmetries between richer and poorer countries and academic institutions. On the other hand, APCs also incentivize publishers to offer Open Access publishing, which promotes Open Science.

Metrics

Number/Share (%) of publications with an APC cost

These metrics measure the number or share (in percentage - %) of publications in journals and have incurred an APC. The share provides a more nuanced understanding of the affordability and accessibility of Open Access publishing than the absolute number and it can be used to compare the

affordability and accessibility of Open Access publishing across different journals, publishers, and regions in a more meaningful way.

Limitation:

- Not knowing who incurred that APC (funder, institution, author, etc.) limits the usefulness of this indicator.
- The share of publications with an APC could be more useful when put together with % Diamond OA publications, as opposed to stand alone.

MEASUREMENT.

DATASOURCES

[OPENAIRE GRAPH](#)

METHODOLOGY

The [OpenAIRE Graph](#) dump currently does not include OA color classifications, though they are already implemented in the OpenAIRE MONITOR and are expected to be integrated into the graph dump in Q1 2024.

1. Retrieve all Gold OA publications from OpenAIRE, which refers to those published in entirely OA journals. Exclude those classified as Diamond OA (meaning they don't have an APC). The publications left in this group are Gold OA articles with associated APCs.
2. Incorporate Hybrid and Bronze OA publications to this set, as these also come with APCs.
3. Cross-reference the APCs listed in OpenAIRE, sourced from OpenAPC ([openapc.net](#)), to ensure no additional articles with APCs are overlooked.
4. Using this refined set of OA publications with APCs, determine the number or share based on your specific area of interest (e.g., country).

Limitations:

- This methodology is a workaround chosen as it provides better coverage than any APC dataset we are aware of, however it is not a direct source of whether a publication has incurred an APC or not.

Average APC

Description:

The "average APC" metric measures the average cost of Article Processing Charges (APCs) across a defined level of interest (per year, country, organization, etc.).

Usefulness:

The “average APC” metric can help assess the affordability and accessibility of Open Access publishing. It provides a broad understanding of the cost of Open Access publishing and identifies trends and changes in APC pricing over time. This metric can help researchers, institutions, and funding agencies to compare the cost-effectiveness of different Open Access publishing models and identify affordable publishing options.

Limitations:

- The cost of APCs can vary widely depending on the field of research, the region, and the specific publisher or journal, therefore taking an average may be misleading.
- Not knowing who incurred that APC (funder, institution, author, etc.) limits the usefulness of this metric.

MEASUREMENT.

DATASOURCES

OPENAIRE GRAPH

The OpenAPC APC dataset, which is integrated in the [OpenAIRE Graph](#).

Limitations:

- Incomplete data: Publishers do not generally provide data on their APC fees, OpenAPC (openapc.net) has a growing collection but it is not complete.
- In an (organization, publication, APC cost) triplet of OpenAPC, to the best of our knowledge, it is not possible to distinguish if the APC cost is the entire cost of the publication or just the what the organization paid.

METHODOLOGY

Via OpenAIRE MONITOR (monitor.openaire.eu)

1. Identify the unit of analysis (e.g. average APCs for an institution)
2. Examine the coverage of APCs for the relevant publications (see metric Number of OA publications with APC).
3. Examine if the coverage is adequate and the distribution of costs meaningful for taking an average.
4. Take the average APC for that level of analysis.

Limitations:

- Averages have the benefit of summarizing and normalizing information, however depending on the underlying distribution of costs, they may be misleading (e.g. via outliers)

Total APC

Description:

The "Total APC" metric measures the sum of APCs paid for all the articles published for a defined level of interest (a year, country, organization, etc.)

Usefulness:

- The "total APC" metric can help assess the affordability and accessibility of Open Access publishing. It provides a broad understanding of the cost of Open Access publishing and identifies trends and changes in APC pricing over time. By summing the APCs this metric measures the total financial burden of OA publishing for the unit of analysis and can be compared to other aggregate measures.

Limitations:

- Not knowing who incurred that APC (funder, institution, author, etc.) limits the usefulness of this metric, more than the previous ones.
- It does not contain information of the distribution of APCs across a subdomain, e.g. Total cost does not give info on how it is distributed across scientific domains.

MEASUREMENT.

DATASOURCES

OPENAIRE GRAPH

The OpenAPC APC dataset, which is integrated in the [OpenAIRE Graph](#).

Limitations:

- Incomplete data: Publishers do not generally provide data on their APC fees, OpenAPC (openapc.net) has a growing collection but it is not complete.
- In an (organization, publication, APC cost) triplet of OpenAPC, to the best of our knowledge, it is not possible to distinguish if the APC cost is the entire cost of the publication or just the what the organization paid.

METHODOLOGY

Via OpenAIRE MONITOR

1. Identify the unit of analysis (e.g. total APCs for an institution)

2. Examine the coverage of APCs for the relevant publications (see metric Number of OA publications with APC).
3. Examine if the coverage is adequate.
4. Sum the APCs for that level of analysis.

Limitations:

- Totals have the benefit of giving a bird's eye view, however depending on the underlying distribution of costs, they can have different implications.

Known correlates

Via: <https://direct.mit.edu/qss/article/1/1/6/15582/Article-processing-charges-Mirroring-the-citation>

- Year
- Publisher
- Hybrid vs. Gold OA
- SNIP

Notes

An APC extrapolation exercise was conducted for the purposes of the following report.

https://research-and-innovation.ec.europa.eu/knowledge-publications-tools-and-data/publications/all-publications/monitoring-open-access-policy-horizon-2020_en

(page 107)

References

Schönfelder, Nina. "Article processing charges: Mirroring the citation impact or legacy of the subscription-based model?." *Quantitative Science Studies* 1.1 (2020): 6-27.

European Commission, Directorate-General for Research and Innovation, *Monitoring the open access policy of Horizon 2020 – Final report*, Publications Office, 2021, <https://data.europa.eu/doi/10.2777/268348>

Availability of data repositories

History

Version	Revision date	Revision	Author
1.1	2023-07-20	Edited & revised	T. Willemse
1.0	2023-04-13	First draft	T. Willemse

Description

In the context of Open Science, the availability of research data is an important topic. Open Data is often, but not exclusively, made available through data repositories, which can store and archive data for long-term preservation. More and more Open Data repositories are initiated and efforts to establish the needed infrastructure are undertaken. However, these repositories differ vastly in their nature and accessibility. It is therefore important to get an overview of the accessibility of these different data sources for the assessment and practice of Open Science.

Governments and governmental agencies, individual universities and research communities are types of organisations involved in setting up data sources in various fields (Goben & Sandusky, 2020). There are also publisher-driven data repositories that stimulate cooperation and can be openly accessible to a certain extent. Lastly, there exist non-institution affiliated data repositories, ranging from field specific (e.g. gene databanks, such as International Nucleotide Sequence Database Collaboration) to more general repositories including multiple topics (e.g. Zenodo, Dryad, figshare).

The wide variety of data repositories out there present a number of opportunities and challenges (Goben & Sandusky, 2020). A clear opportunity is the large increase in accessible data by an increasing number of repositories. However, the wide variety of repositories and infrastructure also presents a challenge in finding the right data repository or dataset that one is looking for. The increase in variety also leads to a risk of data misinterpretation or misuse and can lead to data loss.

Given the potential for Open Data repositories it can be very helpful to get an indication of the accessibility of these resources and how they link up with research. It must be noted however that this indicator is not meant to be solely used to rank data repositories or scientific entities. To do this, other indicators and measures should be taken into account, as well as relevant contextual factors that are difficult to capture in quantitative data.

Metrics

Number of data repositories

The number of data repositories in a given area of interest can provide an indication of the availability of data repositories in this area. The main benefit of this metric being that it can serve as a quick indication of the availability. However, it is limited by the fact that it does not take into account the nature and dimensions of each repository. Since data repositories can differ vastly in their size and openness this metric can give a skewed representation of the available data in the area of interest.

When looking for an indication of how widespread the availability of data repositories is, the percentage of territories, organisations etc., that provide a data repository service could be considered. If sufficient data can be found, this can serve as an indication without much calculation. Again however, this measure is limited, as it does not consider the characteristics of the data repositories.

MEASUREMENT.

The number of data repositories can be represented as a simple count measure. However, in addition to the previously mentioned limitations it might be difficult to obtain data on all existing data repositories, given the large number and variety. The number of accessible data repositories is not a widely adopted metric yet, so data on the (the number of) data repositories is not available on all mainstream platforms. Nevertheless, there are some sources that could help to find information on the number of data repositories. If the information sources allow it, count per field, organisation etc. or percentages can be calculated from these sources as well. This can be done by either limiting the count to the area of interest or in the case of percentages dividing the number of identified data repositories by the total number of units included.

DATASOURCES

RE3DATA

To delve a bit deeper in the characteristics of the data repository, [Re3data](#) maintains a [database](#) of data repositories that is associated with DataCite. It provides many integrated filters on the data and data repositories, like Open Database access and repository type. The website also has an integrated function to filter on subject type, content, and country. This source is therefore useful if characteristics of the data repositories are of importance.

OPENDOAR ([HTTPS://V2.SHERPA.AC.UK/OPENDOAR/](https://v2.sherpa.ac.uk/opendoar/))

OpenDOAR is a global directory of Open Access repositories. It has the functionality to filter on location, type of material and software among others. In addition to providing an overview of other Open Access repositories, it also provides an overview of Open Access repositories that host datasets. One can filter on country and repository name on the website. For each data repository in the database information is available like content types, subjects and identifiers for instance.

DATA CITE

[DataCite](#) is a non-profit organisation that provides DOI's for research data and research output. Within their services DataCite also produces an [overview](#) of data repositories. These include a wide variety of data repositories that are associated with the data that is documented by DataCite. Although, many data repositories and associated datasets are documented here, the catalogue is somewhat limited in filtering for the openness of the data repositories themselves. It can therefore mainly serve as an information source on what type of data repositories are out there. An overview on the analytical possibilities of DataCite can be found in (Robinson-Garcia et al., 2017).

FAIRSHARING.ORG ([HTTPS://FAIRSHARING.ORG/](https://fairsharing.org/))

FAIRsharing.org provides a database on Open Access repositories and in addition provides information on data standards and links to policy documents. The data can be accessed and filtered via an API

OPENAIRE

To determine the availability of data repositories in a specific country using the OpenAIRE Graph, users can access the Graph dump through Zenodo. It is important to note that although the OpenAIRE Graph integrates over 129,000 data sources, the results will encompass data repositories integrated within the OpenAIRE Graph and are contingent upon the quality of information provided by those sources.

Quality of data repositories

Apart from looking at the sole number of data repositories one can also opt to assess the quality of data repositories to indicate its availability. Quality in this context could be seen as the openness of the data repository, cleanliness and completeness of data and metadata and the presence of data curation procedures for instance.

It can be difficult to obtain data on all existing data repositories, given the large number, variety and lack of metadata curation. There is not yet widely available information on the large scientific platforms, but there are some efforts that provide information on the topic listed below.

CORE TRUST SEAL

Core Trust Seal is a non-profit organisation that labels data sources with their seal if data sources adhere to the FAIR principles. On the website a [list](#) is maintained with all the data sources that the seal has been assigned to. Data stored in these sources can thus be considered to be produced in accordance with the FAIR principles. When performing research related to the availability of data repositories, one can consider repositories that have received the CoreTrustSeal, the Nestor Seal DIN31644, the ISO16363 certification, or similar, to be automatically trusted (Jahn et al., 2023).

References

Goben, A., & Sandusky, R. J. (2020). *Open data repositories: Current risks and opportunities* | Goben | *College & Research Libraries News*. <https://doi.org/10.5860/crln.81.1.62>

Jahn, N., Laakso, M., Lazzeri, E., & McQuilton, P. (2023). *Study on the readiness of research data and literature repositories to facilitate compliance with the Open Science Horizon Europe MGA requirements*. Zenodo. <https://zenodo.org/record/7728016>

Robinson-Garcia, N., Mongeon, P., Jeng, W., & Costas, R. (2017). DataCite as a novel bibliometric source: Coverage, strengths and limitations. *Journal of Informetrics*, 11(3), 841–854. <https://doi.org/10.1016/j.joi.2017.07.003>

Availability of preprint repositories

History

Version	Revision date	Revision	Author
1.1	2023-08-28	Draft for publication	I. Grypari
1.0	2023-05-09	First draft	I. Grypari, N. Manola, P. Stavropoulos, H. Papageorgiou

Description

Description:

- The “availability of preprint repositories” indicator measures the availability and accessibility of preprint repositories that are available to researchers in a particular level of interest such as field, country or organization.

Usefulness:

- The “availability of preprint repositories” indicator provides insight into the infrastructure and resources available to researchers for sharing their work as preprints.
- It can help to identify areas where preprint repositories are lacking or where access to these repositories is limited, potentially hindering the adoption of preprinting practices.
- This indicator can also be used to assess the impact of preprint repositories on scholarly communication and to identify areas where additional resources and support may be needed.

Limitations:

- Not all fields or research areas may have a culture of preprinting, which can affect the applicability of the indicator in different contexts.
- There may be good preprint practices in general repositories (such as Zenodo) that would not be identifiable via this indicator.

Metrics

Number of preprint repositories

Description:

- The “number of preprint repositories” indicator measures the quantity of preprint repositories available to researchers in a particular field, country, organization, etc.

Usefulness:

- Provides a straightforward measure of repository availability.

Limitations:

- Does not account for the quality or size of repositories .

MEASUREMENT.

Count

DATASOURCES

ASAPBIO

ASAPbio maintains a list of preprint repositories, available from <https://asapbio.org/preprint-servers>.

Quality of preprint repositories

Evaluating pre-print repositories based on predetermined standards encompassing metadata consistency, user interface, search functionalities, etc.

Usefulness:

- Differentiates repositories by the reliability and efficiency of their content and platform.

Limitation:

- Quality standards can be subjective; high quality in one domain might not be viewed as such in another.

MEASUREMENT.

Use established repository ranking or rating systems, if available. Alternatively, develop a criteria checklist and review each repository against it.

DATASOURCES

METHODOLOGIES

Size of pre-print repositories

Assessing the volume of content or number of publications each repository holds.

Usefulness:

- Differentiates repositories by the reliability and efficiency of their content and platform.

Limitation:

- Size does always correlate with relevance or quality of content, for example in terms of document types.

MEASUREMENT.

Directly query the repository (if it offers such statistics) or query data bases that aggregate repositories.

DATASOURCES

OpenAIRE Graph

METHODOLOGIES

To determine the size or volume of content within specific pre-print repositories in a field, country, or organization using the OpenAIRE Graph, researchers can analyze the Graph dump available on Zenodo or browse OpenAIRE EXPLORE (explore.openaire.eu). It's crucial to understand that while the OpenAIRE Graph has integrated over 129,000 data sources, the data derived will strictly represent repositories integrated into the graph, subject to the quality of information those sources provide (<https://graph.openaire.eu/docs/> can be consulted)

Accessibility of pre-print repositories

Evaluating the access model of publication repositories, specifically categorizing them as open access, subscription-based, or limited to specific users.

Usefulness:

- Provides clarity on whether a wider audience can readily access the content of a repository or if there are restrictions.

Limitation:

- Some repositories might switch access models over time, or have hybrid models combining elements of open and subscription-based access.

MEASUREMENT.

Examine each repository's documented access model. This can be achieved by querying databases that list repositories.

DATASOURCES

OpenAIRE Graph and Directory of Open Access Repositories (DOAR) (integrated in the OpenAIRE Graph)

METHODOLOGIES

All repositories in both data sources are Open Access.

Known correlates

The availability of preprint repositories **may** correlate with specific fields of study, given that certain scientific disciplines are more inclined to use preprints than others. Furthermore, regional or national open science policies, funding opportunities, and research culture can also influence the presence or absence of preprint repositories in a particular country or region.

References

Availability of publication repositories

History

Version	Revision date	Revision	Author
1.1	2023-08-28	Draft for publication	I. Grypari
1.0	2023-02-12	First draft	I. Grypari, N. Manola, H. Papageorgiou, P. Stavropoulos

Description

The availability of publication repositories refers to the presence and accessibility of platforms where academic and research publications are stored and can be accessed by others within specific levels of interest, such as a particular field, country, or organization. These repositories are pivotal in disseminating knowledge, fostering open access, and ensuring that research outputs are preserved and accessible for relevant stakeholders within those defined areas of interest.

Usefulness:

- Proxy for promotion of Open Access: A higher availability of publication repositories often correlates with a stronger emphasis on open access within a region or discipline.
- Proxy for research accessibility: Allows researchers, policymakers, and the general public to access research outputs without barriers, promoting transparency and knowledge-sharing.
- Benchmarking and Strategy: Institutions and countries can use this indicator to gauge their standing in terms of open access and to formulate strategies to enhance their research dissemination.

Limitations of Indicator:

- The number of repositories might not provide insights into the quality or relevance of the content stored within.
- Varied Repository Standards: All repositories are not created equal; there can be variations in terms of indexing, metadata standards, and user-friendliness.
- Temporal Limitations: Some repositories might be more updated than others, leading to discrepancies in the latest research availability.

Metrics

Number of publications repositories

- Usefulness: Provides a straightforward numerical representation of repository availability, allowing for easy comparisons and benchmarks across different domains or regions.

Limitation:

- A mere count does not provide insight into the quality, relevance, or depth of the content within each repository.

MEASUREMENT.

Tallying the number of available publication repositories within a specified level of interest (e.g., field, country, organization).

[DATASOURCES](#)

[OPENAIRE GRAPH, AND OTHERS](#)

[METHODOLOGIES](#)

Query specific databases to obtain the number of repositories for the specified area of interest.

Quality of publications repositories

Evaluating repositories based on predetermined standards encompassing metadata consistency, user interface, search functionalities, etc.

Usefulness:

- Differentiates repositories by the reliability and efficiency of their content and platform.

Limitation:

- Quality standards can be subjective; high quality in one domain might not be viewed as such in another.

MEASUREMENT.

Use established repository ranking or rating systems, if available. Alternatively, develop a criteria checklist and review each repository against it.

DATASOURCES

METHODOLOGIES

Size of publications repositories

Assessing the volume of content or number of publications each repository holds.

Usefulness:

- Differentiates repositories by the reliability and efficiency of their content and platform.

Limitation:

- Size does always correlate with relevance or quality of content, for example in terms of document types.

MEASUREMENT.

Directly query the repository (if it offers such statistics) or query data bases that aggregate repositories.

DATASOURCES

OpenAIRE Graph

METHODOLOGIES

To determine the size or volume of content within publication repositories in a specific field, country, or organization using the OpenAIRE Graph, researchers can analyze the Graph dump available on Zenodo or use OpenAIRE Explore (explore.openaire.eu). It's crucial to understand that while the OpenAIRE Graph has integrated over 129,000 data sources, the data derived will strictly represent repositories integrated into the graph, subject to the quality of information those sources provide (<https://graph.openaire.eu/docs/> can be consulted)

Accessibility of publications repositories

Evaluating the access model of publication repositories, specifically categorizing them as open access, subscription-based, or limited to specific users.

Usefulness:

- Provides clarity on whether a wider audience can readily access the content of a repository or if there are restrictions.

Limitation:

- Some repositories might switch access models over time, or have hybrid models combining elements of open and subscription-based access.

MEASUREMENT.

Examine each repository's documented access model. This can be achieved by querying databases that list repositories.

DATASOURCES

OpenAIRE Graph and Directory of Open Access Repositories (DOAR) (integrated in the OpenAIRE Graph)

METHODOLOGIES

All repositories in both data sources are Open Access.

Known correlates

Various factors **may** influence the availability of publication repositories in specific fields, countries, or organizations such as: research output, research funding, technological infrastructure, Open Access mandates, research collaboration, overall research infrastructure.

Notes

[Add text here]

References

[Add Zotero bibliography here]

Citizen Science Indicators

History

Version	Revision date	Revision	Author
1.0	2023-03-23	First draft	T. Venturini

Description

While there are many competing definitions of citizen science (also called participatory, community, civic, crowd-sourced volunteer science), the notion is generally used to refer to scientific knowledge production with the active and genuine participation of the public (i.e., lay people or non-experts, who are not professionally affiliated with academic or industrial research initiatives).

The European Citizen Science Association ECSA (Robinson et al., 2018) published ten principles describing the citizen science approach and the first five constitute an excellent definition of this approach:

1. Citizen science projects actively involve citizens in scientific endeavour that generates new knowledge or understanding. Citizens may act as contributors, collaborators or as project leaders and have a meaningful role in the project.
2. Citizen science projects have a genuine science outcome. For example, answering a research question or informing conservation action, management decisions or environmental policy.
3. Both the professional scientists and the citizen scientists benefit from taking part.
4. Citizen scientists may, if they wish, participate in multiple stages of the scientific process.
5. Citizen scientists receive feedback from the project.

Notably principle 7 also claims that “citizen science project data and metadata are made publicly available and, where possible, results are published in an open-access format”. In line with this principle, citizen science plays a key role in the movement towards Open Science by opening up the means of knowledge production to the participation of societal actors, across the entire research cycle. As claimed in principle 4 above, in citizen science projects, the public can contribute to scientific efforts in different ways, namely by taking part in

- the definition of the objectives or the research [1. design];
- the development of hypotheses, research questions and methods [2. development];
- the collection of records or knowledge [3. Data collection];
- the cleaning and preparation of the datasets [4. processing];
- the analysis of the data [5. analysis];

- the interpretation of the results [6. interpretation];
- the dissemination of the findings/conclusions [7. dissemination];
- the conservation and sharing of the resources generated by the project [8. ownership]
- the recognition as authors/protagonists of the research [9. credit]

This document describes a series of metrics to quantify (but also qualify) each of these nine different forms of citizen participation in science, as well as a tenth indicator accounting for the capacity of a citizen sciences project to span across multiple forms of participation [10. span].

Existing datasources

Data on citizen science projects can be derived from five different sources:

SCIENTIFIC PROJECT PORTALS

While they tend to have a distinctive approach, many citizen science projects still consider themselves as research projects and are generally funded as such. This means that these projects will be listed in national and international directories, particularly those kept by research funders (e.g., <https://data.jrc.ec.europa.eu/collection/CITSCI>) . Information extracted from these portals can therefore be used to know more about the subjects, the institutions and the finance of citizen science (cf. for example <https://op.europa.eu/en/publication-detail/-/publication/770d9270-cbc7-11ea-adf7-01aa75ed71a1>) as well as to compare these projects to the rest of the scientific project financed in the same years or addressing the same topics.

Yet, it is important to note that the project collections by portals overseen by formal research organizations may focus on citizen science projects initiated by researchers and may overlook projects that have a less academic and more activist nature. It is therefore important not to limit data collection to this source alone.

INDIVIDUAL PROJECT WEBSITES

Because they need to recruit citizens willing to contribute to their research effort, many citizen science projects have developed websites that describe the objectives, activities and results (see a growing portal of this type of projects compiled by the MICS platform: <https://mics.tools/>). These websites can be harvested to collect information about the projects and calculate the metrics described below. Examples include:

- Globe at Night - <https://www.globeatnight.org/>
- Project FeederWatch - <https://feederwatch.org/>
- eBird - <https://ebird.org/home>
- Foldit - <https://fold.it/>
- EyeWire - <https://eyewire.org/>

- MilkyWay@Home - <https://milkyway.cs.rpi.edu/milkyway/>
- Phylo - <http://phylo.cs.mcgill.ca/>
- SETI@home - <https://setiathome.berkeley.edu/>
- BOINC - <https://boinc.berkeley.edu/>
- CoCoRaHS - <https://www.cocorahs.org/>

This type of research has been carried out notably by the project CStrack (<https://cstrack.eu>), which extracted information about almost 5000 citizen science projects extracted for more than 59 websites (<https://zenodo.org/record/7356627>)

CITIZEN SCIENCE WEB PORTALS

The problem with collecting information from individual websites is that their content and architecture may vary significantly from one another and thus require considerable efforts for manual collection and standardisation of data. Indeed global efforts exist to make project descriptions interoperable via an agreed upon metadata schema and vocabulary - see <https://core.citizenscience.org/docs/history>

Alternatively, an increasing number of citizen science projects tend to rely on specialized portals that facilitate some of their activities (e.g., the recruitment of volunteers; their training; the tracking of their contributions; the support of the interaction between volunteers and with the project organisers; etc.). Examples of Citizen science web portals includes

- CitizenScience.gov - <https://www.citizenscience.gov/>
- Zooniverse - <https://www.zooniverse.org/>
- EU-Citizen.Science - <https://eu-citizen.science/>
- Citizen.Science.Asia - <https://citizenscience.asia/>
- SciStarter - <https://scistarter.org/>
- BOINC - <https://boinc.berkeley.edu/>

There is also another layer of portals curated by national citizen science associations, for example: <https://www.citizen-science.at/en/>, <https://www.schweizforscht.ch/>, <https://www.buergerschaffenwissen.de/>, and <https://www.iedereenwetenschapper.be/>.

BIBLIOGRAPHIC DATABASE

As for all research projects, an important output of citizen sciences projects consists of scientific publications (particularly projects that are initiated by research, less so for 'activists-initiated' projects which tend to focus more on data, policy recommendations and social innovation actions etc.). These publications are stored in bibliographic databases and are often available as open access publications (because of the obvious affinity between this type of publication and the approach of civic science). Most of these publications will mention the fact that their results are based on a participatory initiative, cite one of the main citizen science portals or be signed with a collective name (for a couple

of examples of how this can be done see Hunter & Hsu, 2015 and Ozolinčiūtė et al, 2022). All these signs facilitate the identification of publications from citizen science initiatives, allowing analyzing their publication results and to compare them with the rest of the scientific literature.

At the same time, not unlike what noted in relation to the 1st source of data, relying on bibliographic databases will miss all the “academically invisible” citizen science projects that never publish in academic journals, which are in fact in a very large number. This is why none of the sources described here should be used in isolation.

DATA PORTALS

Besides their publication, many citizen science projects also tend to openly publish their dataset in an effort to give back to the public the information collected through its cooperation. Some of these datasets will be released through the individual websites of each project (and sometime in formats that do not necessarily facilitate the reuse), but some others may be published through general data portals, making it possible to collect information that is standardized and comparable with non-participatory projects.

Metrics

Many citizen science projects, particularly when they are initiated by the researchers, tend to concentrate on the central steps of the research process (the collection, processing and analysis of data) as these steps can be externalized (or crowd-sourced) without losing control of the research. These are also the steps on which more information is available since, by dealing with data, these activities are also the easiest to datafy. It is however crucial to gauge the participatory nature of all the stages of a research because real openness tends to be better achieved if all or most of these stages are truly receptive to public input (for a complete typology of citizen science models see Shirk et al., 2012)

Citizen science design

This metric is meant to assess to what extent citizens have been involved in the decisions surrounding the design of the research approach, as well as the nature of their implication: Does the project address concerns that have been surfaced by the community that participate to the project? Have the research questions been defined in collaboration with the public? Is the project led by academic publication/career objectives or is it also guided by civic preoccupations? Open science practitioners have developed a standard vocabulary to talk about these aspects (see <https://core.citizenscience.org/>)

Assessing whether projects truly support co-creation or co-design is a particularly difficult task and can only be assessed by qualitative analysis. To assess citizen science design, researchers can

investigate the history of the projects and discuss with their protagonists, or they can closely read the project documentations to detect if people and concerns from outside the academia are considered and highlighted.

Citizen science development

In researcher-led citizen science project, this step is often the one that is the least often open to citizen participation. Some scientists (particularly those who follow a “deficit model” thinking) would indeed argue that this step should be kept under the control of the expert to assure that the development of the research protocol and methodology remain strictly adherent to scientific best practices, thus guaranteeing the value of the data as well as their comparability (but see Downs et al., 2021). Proponents of a more widely participatory approach, however, will argue (not without reason) that this is the key step of any research project and that if this stage is not open to the public, than citizens cannot truly be the protagonists of the research (and will instead be relegated to the role of useful, but powerless helpers).

As the previous one (and maybe even more than it), this metric can only be assessed through careful qualitative inspection, considering the description of the research protocol and the way in which it has been put together.

Citizen science data collection

This is one of the research steps that has been more traditionally crowdsourced to lay people. Since the Renaissance amateur naturalists have participated in the effort of logging and cataloging different species of plants and animals together with their counterparts in academia. This tradition continues today as biodiversity loss demands to observe and count the movement of different species of insects, birds and amphibians.

Because this step concerns the harvesting of data it is easy to imagine metrics related to the quantity and quality of information collected by citizens, for example:

- Percentage of data collected by citizens of the total amount of data harvested by the project.
- Number of sites or phenomena that are observed exclusively or predominantly by citizens.
- Importance of the crowdsourced data versus other data sources.

For another example of this type of measuring see Fraisl et al., 2020 as well as the Global Biodiversity Information Facility (<https://www.gbif.org>).

Citizen science processing

If data collection is the most classic of crowdsourced scientific activities, the cleaning and (pre)processing of data is the step that is most often crowd-sourced through micro-labor platforms.

A classic hurdle of all current research is an overabundance of poor quality. In the last decades, sensors and other digital technologies have multiplied the number of records collected and stored by scientific projects but have also increased the noise associated with them. Duplicates, errors, impossible outliers need to be detected manually and carefully removed before moving on with the analysis.

The role played by citizens in this work of data processing can be measure through

- Absolute or relative number of errors corrected by citizens.
- Number of hours (or days) invested in manual data cleaning.
- Increase in the quality of data (how such quality is measured depends of course on the specific project)

Citizen science analysis

This step is very close to the previous one and in some cases overlaps with it. Yet, the distinction points at the difference between the relatively low-level work of detecting and removing errors, and the more high-level effort of detecting meaningful patterns and trends in the datasets. Despite the stunning progress of artificial intelligence and other computational techniques, human being remains crucial in the process of pattern recognition and unreplaceable in the constitution of qualified datasets that can be used for machine learning training.

Possible metrics includes:

- Absolute or relative size of the data analyses by citizens (See for example the 'meta' publications of the Zooniverse: <https://www.zooniverse.org/about/publications#meta>).
- Number of hours (or days) invested in the analysis.
- Absolute or relative numbers of pattern detected by citizens (as compared to expert or automatic detection).

Citizen science interpretation

While the two previous steps (processing and analysis) can be simplified (and sometime gamified) to the point of being accessible to anyone – and for this reason represent the standard crowdsourced activities – this step is less often assigned to citizens as it typically involves a different kind of data interface (for example, some advanced statistical software) requiring greater training or technical skills. However, the more the citizens are associated with the work of data interpretation (which is the step where the greatest scientific value is produced) and the more they have agency in it, the more the research can be said to be truly open and participatory.

To assess the role of lay experts in the interpretation of data and generation of findings, one can assess

- Complexity and significance of crowdsourced research tasks.
- Possibility for citizen participants to complete findings/results autonomously (as opposed to intervening only in low level activities, but not being able to achieve the results).
- Participation of citizens in the writing up of the research conclusions.

Citizen science societal impact and participant learning

Many observers and organizers of citizen science projects have argued that, even when it fails to produce new data or findings, one of the main advantages of this approach is that it sensitizes the public to the work of research and helps build science literacy (Roche et al., 2020). Because it involves people outside academia (sometimes in large numbers) citizen science has built-in dissemination effects.

The significance of these effects can be measure by

- Number of regular VS occasional contributors.
- Increase of the number or quality of contributions over time.
- Diversification of the projects that citizens contribute to (when using the same account to participate in different projects within the same portal, see Jackson, 2016).

Citizen science ownership

Because, in participatory science projects, citizens provide an important part of the work, bring in insights and contextual information that greatly improves the quality of the research and its impacts, and sometimes fill major spatial and temporal data gaps, it is crucial that results are also shared with them – be them datasets, scientific findings, policy briefs, intervention recommendations, governance decisions, individual and collective action, social innovation and possibly their intellectual or commercial offshoots. The last of 10 ECSA Principle 10 explicitly states that “the leaders of citizen science projects take into consideration legal and ethical issues surrounding copyright, intellectual property, data-sharing agreements, confidentiality, attribution and the environmental impact of any activities”.

To assess how ownership is shared among all the actors who participated to a citizen science initiative, researchers can look for:

- Legal mechanisms assuring the public ownership of the data or results of the project (e.g., open licenses or collective patents).
- Organizational mechanisms assuring that members/representatives of the public are associated with all decisions related to the research and all the benefits generated by it.

- Political, economic or civil society initiatives deriving from the project and the way in which they are carried out by the same people VS a subset of the people who contributed to the research.

Citizen science credit

Crediting the people who have contributed to the production of science can be as important as granting them the actual ownership of the data or of the results of the research. Sometimes, crediting (in the form of signing or otherwise authoring the projects results of the project) is actually more important than ownership as the primary source of recognition and can provide a stronger form of participants motivation (cf. Land-Zandstra et al., 2021 and Levontin et al., 2022)

Crediting can be assessed by:

- Number of documents (scientific publications, policy briefs, legal interventions, recommendations, governance decisions, technical blueprints, etc.) that mention the use of a citizen science approach.
- Number of documents that mention the name of all the individuals or of the citizen organizations that have contributed to the research.

Citizen science span

The metrics described above refer to specific stages of the research process, but (as explained in section I) the opening of multiple steps is even more important.

- The most straightforward way of measuring this is by the simple count of the number of steps in which citizens have been given a chance to be active.
- A less binary option is to define a scale of "citizen agency/participation" for each step and then compute the average value across the whole research protocol (see for instance the model proposed by Gharesifard et al., 2017).

References

- Downs, Robert R., Hampapuram K. Ramapriyan, Ge Peng, and Yaxing Wei. 2021. "Perspectives on Citizen Science Data Quality." *Frontiers in Climate* 3 (April). <https://doi.org/10.3389/fclim.2021.615032>.
- Fraisl, Dilek, Jillian Campbell, Linda See, Uta Wehn, Jessica Wardlaw, Margaret Gold, Inian Moorthy, et al. 2020. "Mapping Citizen Science Contributions to the UN Sustainable Development Goals." *Sustainability Science* 15 (6): 1735–51. <https://doi.org/10.1007/s11625-020-00833-7>.

Gabrys, J., Pritchard, H., & Barratt, B. (2016). Just good enough data: Figuring data citizenships through air pollution sensing and data stories. *Big Data & Society*, 3(2), 205395171667967. <https://doi.org/10.1177/2053951716679677>

Gharesifard, Mohammad, Uta Wehn, and Pieter van der Zaag. 2017. "Towards Benchmarking Citizen Observatories: Features and Functioning of Online Amateur Weather Networks." *Journal of Environmental Management* 193 (May): 381–93. <https://doi.org/10.1016/j.jenvman.2017.02.003>

Hunter, Jane, and Chih-Hsiang Hsu. 2015. "Formal Acknowledgement of Citizen Scientists' Contributions via Dynamic Data Citations." In , 64–75. https://doi.org/10.1007/978-3-319-27974-9_7.

Jackson, Corey, Carsten Østerlund, Veronica Maidel, Kevin Crowston, and Gabriel Mugar. 2016. "Which Way Did They Go?" In *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing*, 624–35. New York, NY, USA: ACM. <https://doi.org/10.1145/2818048.2835197>.

Lämmerhirt, D., Gray, J., Venturini, T., & Meunier, A. (2019). Advancing Sustainability Together? Citizen-Generated Data and the Sustainable Development Goals. *SSRN Electronic Journal*. <https://doi.org/10.2139/ssrn.3320467>

Robinson, Lucy Danielle, Jade Lauren Cawthray, Sarah Elizabeth West, Aletta Bonn, and Janice Ansine. 2018. *Citizen Science*. UCL Press. <https://doi.org/10.14324/111.9781787352339>.

Roche, Joseph, Laura Bell, Cecília Galvão, Yaela N. Golumbic, Laure Kloetzer, Nieke Knobens, Mari Laakso, et al. 2020. "Citizen Science, Education, and Learning: Challenges and Opportunities." *Frontiers in Sociology* 5 (December). <https://doi.org/10.3389/fsoc.2020.613814>.

Ozolinčiūtė, Eglė, William Bülow, Sonja Bjelobaba, Inga Gaižauskaitė, Veronika Krásničan, Dita Dlabolová, and Julija Umbrasaitė. 2022. "Guidelines for Research Ethics and Research Integrity in Citizen Science." *Research Ideas and Outcomes* 8 (November). <https://doi.org/10.3897/rio.8.e97122>.

Shirk, J. L., Ballard, H. L., Wilderman, C. C., Phillips, T., Wiggins, A., Jordan, R., ... Bonney, R. (2012). Public participation in scientific research: A framework for deliberate design. *Ecology and Society*, 17(2). <https://doi.org/10.5751/ES-04705-170229>

Strasser, B. J., Baudry, J., Mahr, D., Sanchez, G., & Tancoigne, E. (2017). "Citizen Science"? Rethinking Science and Public Participation. *Science & Technology Studies*, *Forthcoming*.

Deposition of Open Metadata

History

Version	Revision date	Revision	Author
1.0	2023-05-21	Initial draft	L. Waltman

Description

Scientific publications have metadata that describes important properties of the publication, such as the title, the authors, the publication date, and the references. This metadata is used to help researchers and others find relevant literature. Metadata of publications is also often used in bibliometric analyses to support research evaluation and research management.

Traditionally the metadata of scientific publications is made available in proprietary commercial databases. However, the importance of openness of this metadata is increasingly recognized, as shown for instance by the [Initiative for Open Citations](#) and the [Initiative for Open Abstracts](#).

Our focus is on openness of the metadata of research articles. We acknowledge the importance of openness of other types of publications, such as books, book chapters, and policy reports. Nevertheless, in most scientific fields, research articles are seen as the most important publication type, and we therefore restrict ourselves to this publication type.

Metrics

Journal articles with Open References

This metric provides the number or the percentage of journal articles for which the references are openly available.

MEASUREMENT.

Crossref is the datasource for this metric. Crossref makes the references of articles openly available, but it can do this only for articles for which the publisher submitted the references to Crossref. The metric is obtained by determining for each article whether it has a Crossref DOI and whether the metadata for this DOI includes references.

If the metadata of an article in Crossref does not include references, this means that either the publisher did not submit the references to Crossref or the article does not have any references. There is no straightforward way to distinguish between these two possibilities.

The [Crossref Participation Reports](#) provide this metric at the level of publishers and journals. The metric is also used in [this article on open metadata in Crossref](#). To calculate the metric yourself, you may use the Crossref API, documented at <https://api.crossref.org/swagger-ui/index.html>. The 'references-count' field in the API output indicates whether the metadata of an article includes references.

Journal articles with Open Abstracts

This metric provides the number or the percentage of journal articles for which the abstract is openly available.

MEASUREMENT.

Crossref is the datasource for this metric. Crossref makes the abstracts of articles openly available, but it can do this only for articles for which the publisher submitted the abstract to Crossref. The metric is obtained by determining for each article whether it has a Crossref DOI and whether the metadata for this DOI includes an abstract.

If the metadata of an article in Crossref does not include an abstract, this means that either the publisher did not submit the abstract to Crossref or the article does not have an abstract. There is no straightforward way to distinguish between these two possibilities.

The [Crossref Participation Reports](#) provide this metric at the level of publishers and journals. The metric is also used in [this article on open metadata in Crossref](#). To calculate the metric yourself, you may use the Crossref API, documented at <https://api.crossref.org/swagger-ui/index.html>. If the metadata of an article includes an abstract, it can be found in the 'abstract' field in the API output.

Journal articles with Open Author Affiliations

This metric provides the number or the percentage of journal articles for which author affiliations are openly available.

MEASUREMENT.

Crossref is the datasource for this metric. Crossref makes the affiliations of the authors of articles openly available, but it can do this only for articles for which the publisher submitted the author affiliations to Crossref. The metric is obtained by determining for each article whether it has a Crossref DOI and whether the metadata for this DOI includes author affiliations.

If the metadata of an article in Crossref does not include author affiliations, this means that either the publisher did not submit the author affiliations to Crossref or the article does not include any author affiliations. There is no straightforward way to distinguish between these two possibilities.

This metric is used in [this article on open metadata in Crossref](#). To calculate the metric yourself, you may use the Crossref API, documented at <https://api.crossref.org/swagger-ui/index.html>. If the metadata of an article includes author affiliations, these can be found in the 'affiliation' field in the API output.

Journal articles with Open Funding Information

This metric provides the number or the percentage of journal articles for which funding information is openly available.

MEASUREMENT.

Crossref is the datasource for this metric. Crossref makes funding information of articles openly available, but it can do this only for articles for which the publisher submitted funding information to Crossref. The metric is obtained by determining for each article whether it has a Crossref DOI and whether the metadata for this DOI includes funding information.

If the metadata of an article in Crossref does not include funding information, this means that either the publisher did not submit funding information to Crossref or the article does not include any funding information. There is no straightforward way to distinguish between these two possibilities.

The Crossref Participation Reports provide this metric at the level of publishers and journals. The metric is also used in [this article on open metadata in Crossref](#). To calculate the metric yourself, you may use the Crossref API, documented at <https://api.crossref.org/swagger-ui/index.html>. If the metadata of an article includes funding information, this information can be found in the 'funder' field in the API output.

Evaluation of Open Science in research assessment

History

Version	Revision date	Revision	Author
1.1	2023-07-20	Edited & revised	V.A. Traag
1.0	2023-07-12	Initial draft	T. van Leeuwen

Description

Research assessment is organised differently across the global science system. We here discern four points of research assessment, national assessment exercises or protocols, research funding policies, institutional hiring policies and finally journal peer review.

On the country level, research assessment can be organised in national initiatives (e.g., Italy), rely on protocols (e.g., the Strategy Evaluation Protocol (SEP) in the Netherlands), or by having performance-based funding systems (e.g., UK, Italy, Australia, Norway). However, many countries lack clearly defined research assessment procedures on the national level (e.g., in Europe France or Germany, or the USA). Some countries have a national system that is very tightly organised and allows for very little changes in the system, as the evaluation procedures are embedded in the national laws on higher education (e.g., Italy). In some of the countries mentioned, proof of Open Science practices is considered in the assessment. For example, in the UK REF, only Open Access publications were assessed, while in the Netherlands Open Science is one of the aspects that are evaluated in the SEP.

When it comes to funding of research, funders in various countries ask for different things. In most cases, OA publishing, in various forms, is encouraged as part of results dissemination. COALition S, an international consortium of research funders aimed at OA publishing, and as a next target, on the openness of the resulting research data. Similarly, the research funded by the European Commission should also be published in OA format, and data should be as open as possible. All of this consists of ex ante requirements, and such mandates have varying effectiveness (Larivière & Sugimoto, 2018). It is difficult to get an impression on how this develops, or how this is monitored.

When it comes to hiring and/or promotion procedures, these are often, if not always, organised on the institutional level. Increasingly, the degree of Open Science practices in prior positions in a career track is taken into consideration, but how this is organised in various institutions is not immediately clear.

So, we see that across different levels of organization, the uptake of Open Science practices in the primary knowledge creation process is in general not very systematically integrated into research evaluation practices. Often the publishing part is considered in terms of available data to create valid and trustworthy indicators (see the indicator on Open Access publishing, as well as ROARMAP (<https://roarmap.eprints.org/>), an international registry on OA publishing mandates). On most other dimensions of the primary knowledge creation process (e.g., logbooks, data sharing, peer review, etc.) such information is not available or available only in a partial and fragmented way (e.g., for research data, DataCite is such a source, that is a good entrance on data storage, sharing and impact, but hardly comprehensive).

Given the systemic differences and the requirements in various levels of organization mentioned above, indicators have to be relatively simple (e.g., straight counts of occurrences), embedded in a narrative to explain the situation concerning such an indicator. A general problem in this domain is the absence of systematic data sources to support the creation of substantive and robust generic metric indicators, so one has to compromise by aiming at simple indicators. A recent EU project, <https://graspos.eu/GraspOS> is particularly focused on the role of Open Science in research assessment, also with the aim to collect more systematic data.

Metrics

Number/% hiring policies that reward OS

This metric can be constructed by creating a perspective on the ways that, within a national setting, hiring of new staff is enriched by aiming at the uptake of Open Sciences practices by potential candidates. The national-institutional level is probably most successful, as internationally one might get into issues due to various national and /or funding agencies requirements (see above), apart from the efforts due in collecting such information on an international scale.

The metric consists of comparing the number or share of institutions that have policies in place that positively assess the uptake of Open Science practices by candidates, versus the total number of institutions involved.

Potential issues in designing this indicator are:

- what variety of Open Science practices does one take into consideration ?
- How open are institutions in sharing such information on their hiring policies ?

This indicator could be aligned with other potential indicators on the uptake of Open Science practices, e.g., on the uptake of Open Access publishing, Preprinting, Data Sharing, Open Peer Review, Open Logbooks, but also Registered Reports, Preregistration, etc.

MEASUREMENT.

One could create this metric best by aiming at the national level, and on that national level at institutions (universities and other publicly funded research organizations). There one might expect a certain alignment to the national policies on Open Science and the way it should be rewarded. The international level does not supply such an alignment (except perhaps for internationally operating funding agencies).

Potential issues in creating this metric, and measuring the number/% hiring policies that reward OS are:

- Within an organization, differences might exist on faculty level, regarding the positive reward of the uptake of Open Science practices (since not all scholarly disciplines are equally aligned when it comes to the uptake of Open Science practices, the inclusion of the scholarly domain would be a good suggestion, which complicates this metric substantially)
- This metric is time-consuming, since no systematic data sources are available.

This information needs to be collected by a qualitative approach, as automation is still not possible at the moment.

Institutions should be approached asking for input on their hiring policies, and the position of the uptake of Open Science practices in these hiring policies. The result of such a data collection procedure is that one has the number or share of the institutions on national level that positively assess the uptake of Open Science practices by candidates for a job opening, compared to the total number of institutions on the national level.

Alternatively, one might study job advertisements to see if Open Science aspect are mentioned (Khan et al., 2022). However, this is again a manual process, for which no automated procedures are available (yet). Additionally, it is possible that there might be policies for considering Open Science elements in hiring decisions, but that this is not reflected in job advertisements.

Number/% grant evaluation policies that reward OS

This metric can be constructed by creating a perspective on the ways funding agencies include Open Science practices in their assessment procedures. As with the previous metric, all kinds of complexities play a role here, as we distinguish supra-national from national funding agencies, and public funders from private funders and charities. As no systematic overview exists of all the various requirements on the uptake of Open Science practices in the assessment of grant proposals, this has to be collected separately. Given that here the supra-national and the national are hardly separable, Science Europe might play a relevant role, next to national funding agencies and private funders and charities.

The metric could exist of the number or share of policies that reward Open Science practices in assessing research grant proposals, compared to the total number of agencies (of different kinds).

Potential issues in designing this indicator are:

- what variety of Open Science practices does one take into consideration ?
- How open are funding agencies in sharing such information on their grant evaluation policies ? Probably the public funders are more transparent, but are the private funders and charities equally transparent on their grant evaluation policies?
- Missions might be different, charities have a more urgent pressure to fund societally relevant research, given that charities are dependent on their donors, so how is that aligned with more general trends towards Open Science?

MEASUREMENT.

One could create this metric best by approaching funding agencies (Science Europe might play a role here, given its supra-national character) to inquire on their grant evaluation procedures, inquiring what elements are mentioned in funding calls that need a more open and transparent approach by the potential grantees. The outcome of such an inquiry delivers the total number of agencies that were approached, and the ones that consider Open Science practices in their assessment of grant proposals.

Potential issues in creating this metric, and measuring the number/% hiring policies that reward OS are:

- Different funding agencies might require different aspects of the uptake of Open Science practices, how does one take that into consideration?
- Since not all scholarly disciplines are equally aligned when it comes to the uptake of Open Science practices, the inclusion of the scholarly domain would be a good suggestion, which complicates this metric substantially.
- Given the varied missions of funding agencies (the differences that might occur between public versus private funders and charities) might lead to differences in what is required regarding openness of research and its results.
- This metric is time-consuming, since no systematic data sources are available.

This information needs to be collected by a qualitative approach, as automation is still not possible at the moment.

Number/% journal peer review policies that incentivise OS

There is a large variety of journals, including also gold and diamond Open Access journals, for which we can expect different Open Science incentives. For journals there are more resources developed than for institutional policies. In particular, for Open Access policies, there are quite well maintained

resources available. There are also various initiatives to track other aspects around journal policies, including peer review and data sharing policies.

Potential issues in designing this metric are:

- Not all scholarly disciplines are equally aligned when it comes to the uptake of Open Science practices. Such field differences might need to be considered when constructing this indicator.
- The varied scopes of journals might lead to differences in what is required regarding openness of research and its results.

MEASUREMENT.

Based on various data sources, we can construct metrics about the number of journals that implement a particular policy. In principle, one could combine various such policies to get at an overall metric of journals that implement one or more policies that incentivises Open Science. To get at a percentage, we also need the total number of journals, which can be more challenging to describe. This could be based on the data source used to measure a particular policy, but this might come with limitations. Various bibliometric databases cover various journals, but few databases are truly comprehensive. One of the most comprehensive journal list is Ulrich's Periodical Directory.

DATASOURCES

SHERPA ROMEO

Sherpa Romeo provides a graphical user interface at <https://www.sherpa.ac.uk/romeo/>. This can be browsed to collect information about various journals. This includes information about various conditions around Open Access publishing. They also have an API available, see <https://v2.sherpa.ac.uk/api/> for more information.

Not all journals and publishers are necessarily included on Sherpa Romeo, and is limited to journals that have at least one Open Access option.

TRANSPARENCY AND OPENNESS PROMOTION GUIDELINES

The Transparency and Openness Promotion Guidelines (Nosek et al., 2015) are available from <https://www.cos.io/initiatives/top-guidelines>, while various metrics around Open Science practices of journals are available from the related website <https://www.topfactor.org/>. They provide an overall TOP factor, which is a compound metric based on the various individual Open Science aspects that journals adhere to. The TOP factor and the underlying scores on the individual Open Science aspects are available in a graphical user interface at <https://www.topfactor.org/>. An overview of the various policies covered and scores is available from <https://www.cos.io/initiatives/top-guidelines>. The underlying data is available for download from <https://osf.io/qatkz>.

At the moment, there are over 2600 journals included in the TOP factor. Although this is quite extensive, it is a relatively small proportion of the total number of journals, with a relatively higher representation of journals in psychology, economics and education in addition to more general science outlets.

PLATFORM FOR RESPONSIBLE EDITORIAL POLICIES

The Platform for Responsible Editorial Policies is available from <https://www.responsiblejournals.org/>. This platform is focused in particular on (open) peer review policies. It covers various aspects about the timing, the openness, the specialization, and the technical infrastructure of peer review, with details provided on <https://www.responsiblejournals.org/information/peerreviewpolicies>. All information can be browsed through a graphical user interface, but is also available for download from <https://www.responsiblejournals.org/database/download>. Coverage is limited to about 500 journals at the moment.

TRANSPOSE

TRANsparency in Scholarly Publishing for Open Scholarship Evolution (Transpose) maintain information about (open) peer review policies, and is available from <https://transpose-publishing.github.io>. It also covers various aspects around peer review, but also about preprinting, with details provided on <https://transpose-publishing.github.io/#/more-information>. It contains a graphical user interface, but data can also be downloaded in full.

JOURNAL OBSERVATORY

The Journal Observatory integrates information from various sources, including some of the aforementioned sources, and is available from <https://www.journalobservatory.org/>. It provides an integrated view of these various sources, and provides a framework for describing journals. A prototype of the framework provides also an integrated view, which is accessible through an API and as a SPARQL end-point from <https://www.journalobservatory.org/prototype/>. It also provides a graphical user interface where results can be browsed at <https://app.journalobservatory.org/>.

EXISTING METHODOLOGIES

Additional information needs to be collected by a qualitative approach, automated detection of journal policies is not yet possible at the moment.

References

Khan, H., Almoli, E., Franco, M. C., & Moher, D. (2022). Open Science failed to penetrate academic hiring practices: A cross-sectional study. *Journal of Clinical Epidemiology*, 144, 136–143. <https://doi.org/10.1016/j.jclinepi.2021.12.003>

Larivière, V., & Sugimoto, C. R. (2018). Do authors comply when funders enforce Open Access to research? *Nature*, *562*(7728), 483–486. <https://doi.org/10.1038/d41586-018-07101-w>

Nosek, B. A., Alter, G., Banks, G. C., Borsboom, D., Bowman, S. D., Breckler, S. J., Buck, S., Chambers, C. D., Chin, G., Christensen, G., Contestabile, M., Dafoe, A., Eich, E., Freese, J., Glennerster, R., Goroff, D., Green, D. P., Hesse, B., Humphreys, M., ... Yarkoni, T. (2015). Promoting an open research culture. *Science*, *348*(6242), 1422–1425. <https://doi.org/10.1126/science.aab2374>

Distribution of Open Access journal models

History

Version	Revision date	Revision	Author
1.1	2023-05-09	Draft for publication	I. Grypari
1.0	2023-05-09	First Draft	N. Manola, I. Grypari, H. Papageorgiou, P. Stavropoulos

Description

Description:

- This indicator offers a comprehensive view of the diversity and prevalence of different open access publishing models in a specified research domain, whether regional, national, or subject-specific. It differentiates journal business model along two dimension.

Access rights of Articles:

- Fully OA Journal: All articles are OA
- Hybrid Journal: Some articles are OA some not.
- Subscription Journal (remaining)

APC costs of OA Articles:

- Diamond OA Journal: Fully OA journal without APCs.
- Fully OA Journal with APCs (remaining)

Usefulness:

- Provides a comprehensive view of the open access publishing ecosystem, showcasing how traditional and modern publishing models coexist.
- Enables stakeholders to gauge the extent of OA adoption and the diversity of financial models supporting it.

Can inform policy decisions, grant funding requirements, and authors' publication choices by showcasing the prominence and availability of various OA models

- Solely quantitative; does not address the qualitative aspects of journals, such as their reputation, impact factor, or the quality of peer review.
- Not all fields or research areas may have a culture of OA publishing, which can affect the comparison of the indicator across different contexts.
- This indicator also does not account for other forms of OA publishing, such as OA monographs or book chapters, which can also play a role in the dissemination of research findings.

Metrics

Number/Share (%) of Fully OA Journals

Description:

Number/Share (%) of OA journals in a specific area of interest (e.g., country, subject, year) Usefulness:

- Allows for comparison of Open publishing prevalence across different regions or fields of study.
- Provides insights into the growth or decline of OA journals over time.
- Proxy for research visibility as areas with a higher percentage of OA journals likely offer greater visibility and accessibility to research findings.
- Can be used by institutions and funders to gauge the prevalence of OA platforms available to researchers in specific areas and adjust funding or publishing mandates accordingly.

Limitations:

- A sheer count or percentage doesn't give insights into the quality of the OA journals.

NOT ALL FIELDS OR RESEARCH AREAS HAVE A CULTURE OF OA PUBLISHING, WHICH CAN SKEW COMPARISONS.MEASUREMENT.

Count or share.

[DATASOURCES](#)

[DOAJ \(INCLUDED IN THE OPENAIRE GRAPH\)](#)

Limitations:

- While DOAJ is comprehensive, not all OA journals may be listed, especially if they are newer or have not met the DOAJ's criteria.

OpenAIRE Graph

METHODOLOGY

This methodology is taken from [OpenAIRE MONITOR](#) and uses the OpenAIRE Graph as the input database.

A journal is defined as fully Open Access if one or more of the following occur:

- It is in the Directory of Open Access Journals (DOAJ)
- It has a known fully OA Publisher (curated list that will be included in the OpenAIRE Graph in Q1 2024).
- It only publishes OA articles.

Number/Share (%) of Hybrid OA Journals

A hybrid OA journal is a subscription journal where some of its articles are open access.

Usefulness

- Indicates the prevalence of journals that provide open access options without being fully open access.
- Provides an understanding of how many journals offer a middle-ground approach to OA.

Limitations:

- The metric may not capture the nuances of each hybrid journal's open access policies.
- The prevalence of hybrid OA doesn't necessarily indicate the volume of OA content..

MEASUREMENT.

DATASOURCES

OPENAIRE GRAPH

Count or take the share (%) of journals in OpenAIRE with open access articles that are not fully OA journals as defined in the previous metric.

Number/Share (%) of Diamond OA Journals

A diamond OA journal is a fully OA journal that does not charge article processing charges (APCs). In other words the diamond OA journals are a subset of the fully Open Access journals (described in 10.3.1).

Usefulness

- Provides insights into journals that are promoting open access without transferring the cost to authors. Indicates the prevalence of journals that provide open access options without being fully open access.
- Can indicate a commitment to equitable knowledge dissemination in the academic publishing landscape.

Limitations:

- Relying on APC data might not capture other potential costs or financial barriers associated with publishing in the journal.

MEASUREMENT.

DATASOURCES

DOAJ

OPENAIRE GRAPH

Use APC data from DOAJ (integrated in the OpenAIRE Graph or using DOAJ's Public Data Dump - an exportable version of the journal metadata). Use it to determine whether a particular fully OA journal charges APCs.

References

[Add Zotero bibliography here]

Prevalence of Open Access publishing

History

Version	Revision date	Revision	Author
1.1	2023-07-12	Revised based on review	V.A. Traag
1.0	2023-02-08	Added initial description	V.A. Traag

Description

Open Access publishing is one of the pillars of Open Science. Whereas traditional academic publishing charges a fee for reading a publication, Open Access publications are free to read, not only by scientists, but also by the general public. In addition to being free to read, most definitions of Open Access also require publications to be reusable (Suber, 2012).

There are various models of Open Access publishing. Hybrid Open Access journals publish both Toll Access, where readers have to pay a fee, and Open Access articles. Some journals publish all articles Open Access, this is called Gold Open Access. If no fee is charged for publishing in such an only Open Access journal, this is called Diamond Open Access. If a publication in a journal is also deposited in a publication repository, for instance a university repository, and is made available openly, this is called Green Open Access. Finally, some articles are freely available from the website of the publisher, but have no licence attached to them, so it is not clear whether the article will remain freely available in the future and under what conditions it can be used. This is sometimes called Bronze Open Access. For an overview of the typology see (Robinson-Garcia et al., 2020).

In summary, we have the following types of Open Access:

- Green
 - Article that is deposited in an Open Access repository.
- Gold
 - Open Access article in a journal that publishes only Open Access articles.
- Diamond
 - Open Access article in a journal that publishes only Open Access articles.
 - No publishing fee, i.e. no APC
- Hybrid
 - Open Access article in a journal that also publishes Closed Access articles.
 - Typically charges an APC
- Bronze

- Article that is accessible without paywall from a publisher.
- No licence attached, unclear whether article will remain freely available in the future and under what conditions it can be used.

Note that the different Open Access statuses are not necessarily mutually exclusive. In particular, any article can be both Green Open Access and any other type of Open Access. Similarly, Diamond Open Access is a subtype of Gold Open Access, so that every Diamond Open Access article is also Gold Open Access.

We are here interested in the extent to which articles are published Open Access. Tracking such an indicator over time could provide some idea of whether how the uptake of Open Access publishing develops. We can split out the overall uptake of Open Access publishing per type of Open Access to understand the developments in more detail.

Books and other material can also be published Open Access, but Open Access publishing has not yet been taken up frequently in book publishing (Grimme et al., 2019). The book publishing landscape is also quite scattered and challenging. Books, especially monographs, play an important role in various fields in the humanities and social sciences. Also including books in studies of Open Science is therefore particularly relevant for these fields. However, we do not (yet) include a metric for Open Access books here.

There are some recurrent questions around Open Access publishing. Some questions concern effects of Open Access publications, for instance whether Open Access increases the impact of publications. Other questions focus on effect on Open Access, for instance whether certain Open Access mandates increasing Open Access publishing. By measuring the overall uptake of Open Access publishing, this indicator can help answer such questions.

Metrics

Number/% of Open Access journal publications (per OA type)

By counting the overall number of Open Access publications, we get a first impression of the overall uptake of Open Access publishing. If Open Access publishing becomes more widespread, we expect to see more Open Access publications, and hence, by counting the number of Open Access publications, we may get some idea of the overall uptake of Open Access publishing.

This metric is expected to scale with the overall number of publications, and we therefore call it a size-dependent metric.

One limitation of simply counting the number of Open Access publications is that it may partly reflect the general trends of publishing. For instance, suppose we observe an increase in the number of Open Access publications. We could then conclude that Open Access publishing has become more widespread. In terms of the sheer amount of scientific knowledge that has become openly available, this observation seems to be quite right. However, suppose that the increase in Open Access publications is simply proportional to the overall increase in total number of publications. This may then indicate that overall, the chance of a paper being published Open Access has remained the same. This may indicate that, for instance, researchers have not become more likely to publish Open Access. What viewpoint is more relevant depends on the context. For this reason, we therefore also describe a metric of the % of Open Access publications.

We can count the number of Open Access publications based on different selections. For instance, we may want to select only publications within a particular country, institution, field of science or publication year. We do not describe here how to make such selections, we here merely clarify how to count the Open Access publications, given a particular set of publications. As explained in the Description, we can distinguish between different types of Open Access publications: Hybrid, Gold, Diamond, Green and Bronze. We will clarify how to measure these aspects of Open Access publishing.

MEASUREMENT.

For each publication we need to measure whether it is Open Access at all, and if so, what specific type(s) of Open Access it is. In order to be able to do this, we need to know whether the publication is deposited in one or more publication repositories (for Green OA) and know whether the article is directly available from the publisher. In the latter case, we also need to be able to know whether the journal itself is fully Open Access or whether it is a hybrid journal. In addition, if it is a fully Open Access journal, we need to determine whether the journal charges any publishing fees (APC), if it does not, we can classify it additionally as Diamond Open Access.

Hence, this requires (1) data on publication repositories; (2) data on published articles; and (3) data on journals. Fortunately, much, but not all, of this information is integrally available from some existing datasources.

By looking at the percentage of Open Access publications, we get an idea of the uptake of Open Access publishing as a practice. That is, it provides an indicator of the likelihood that a publication is published Open Access. If this percentage increases, it suggests that scholars are more often publishing their work Open Access. Unlike the Number of Open Access publications, this metric does not scale with the overall number of publications, and is therefore called a size-independent metric.

We can study the percentage of Open Access publications based on different selections. For instance, we may want to select only publications within a particular country, institution, field of science or publication year. We do not describe here how to make such selections, we here merely clarify how to calculate the percentage of Open Access publications, given a particular set of publications.

DATASOURCES

UNPAYWALL

[Unpaywall](#) is a database that contains information about online locations of scholarly publications and the Open Access status of these locations (Piwowar et al., 2018). Unpaywall is organised on the basis of DOIs and is limited to Crossref DOIs. For each DOI, information about online locations and Open Access status is provided based on scraping journal websites and publication repositories. Sometimes, information from Unpaywall may be incorrect (Piwowar et al., 2018), so there is room for improvement.

There are some particular limitations regarding Green Open Access. First, Unpaywall may not be able to track all publication repositories. Hence, it might be that some publications are in reality Green Open Access, but Unpaywall may have missed the repository from which it is available. Second, even if Unpaywall has located a repository from which the publication is available, it is difficult to ascertain whether it is the same version as the published version, or whether there are still some differences.

Unfortunately, Unpaywall does not record whether a journal charges an APC or not, and so we cannot determine whether an article is Diamond Open Access. Please see the indicator on Journal Open Access journals.

Unpaywall can easily be used through an API, as illustrated below.

```
import requests
doi = '10.7717/peerj.4375'
email = 'unpaywall_01@example.com'
url = f'https://api.unpaywall.org/v2/{doi}?email={email}'
response = requests.get(url)
doi_info = response.json()
```

Please note that the online API is rate-limited, and for bulk access, it is highly preferable to download the complete dataset and process it locally. Unpaywall provides a field `is_oa`, which indicates whether the article is Open Access. In addition, it provides the field `oa_status`, which clarifies the type of Open Access (limited to `gold`, `hybrid`, `bronze`, `green` and `closed`). However, this assigns a publication to only one Open Access type, while an Open Access publication can in principle be both Green Open Access and any other type of Open Access.

Instead of using the Unpaywall provided Open Access status, we can also define Open Access status explicitly. This in particular allows us to define multiple Open Access statuses for a single paper. We can do that as illustrated below.

```
oa_green = False
oa_gold = False
oa_hybrid = False
oa_bronze = False
for location in doi_info['oa_locations']:
```



```

if location['host_type'] == 'repository':
    oa_green = True
elif location['host_type'] == 'publisher':
    if doi_info['journal_is_oa']:
        oa_gold = True
    else:
        if location['licence']:
            oa_hybrid = True
        else:
            oa_bronze = True

```

To facilitate interaction with the Unpaywall API there are also supporting packages available in Python (<https://pypi.org/project/unpywall/>) and R (<https://cran.r-project.org/package=roadoi>).

Once we know the Open Access status of each publication, we can easily calculate the percentage, simply as the number of publications out of the total that are Open Access. Additionally, we can do this for each separate Open Access type. We could for example refer to the % Green OA or % Gold OA. Note that percentages may not add up to 100%, but the total may also even exceed 100%, because publications can have both a Green Open Access status and another Open Access status. For that reason, you should not report percentages of Open Access statuses in a cumulative fashion (e.g. not in a stacked bar chart), in particular not when reporting on Green Open Access.

Known correlates

There is a large ongoing debate whether Open Access publishing increases the citation impact of publications, known as the so-called Open Access citation advantage. A recent systematic review on the topic suggests that the evidence is inconclusive (Langham-Putrow et al., 2021).

References

Grimme, S., Taylor, M., Elliott, M. A., Holland, C., Potter, P., & Watkinson, C. (2019). *The State of Open Monographs* [Report]. Digital Science. <https://doi.org/10.6084/m9.figshare.8197625.v4>

Langham-Putrow, A., Bakker, C., & Riegelman, A. (2021). Is the Open Access citation advantage real? A systematic review of the citation of Open Access and subscription-based articles. *PLOS ONE*, *16*(6), e0253129. <https://doi.org/10.1371/journal.pone.0253129>

Piwohar, H., Priem, J., Larivière, V., Alperin, J. P., Matthias, L., Norlander, B., Farley, A., West, J., & Haustein, S. (2018). The state of OA: A large-scale analysis of the prevalence and impact of Open Access articles. *PeerJ*, *6*, e4375. <https://doi.org/10.7717/peerj.4375>

Robinson-Garcia, N., Costas, R., & van Leeuwen, T. N. (2020). Open Access uptake by universities worldwide. *PeerJ*, *8*, e9410. <https://doi.org/10.7717/peerj.9410>

Suber, P. (2012). *Open Access*. The MIT Press. <https://library.oopen.org/handle/20.500.12657/26065>

Open Science training facilities

History

Version	Revision date	Revision	Author
1.0	2023-07-12	Initial draft	T. van Leeuwen

Description

Open Science training is mostly organized at the institutional and even faculty level. This is often organized by the university or faculty libraries when it comes to scholarly publishing, that is, Open Access publishing. In addition, there may be national Open Science initiative that might provide Open Science training facilities. This might involve national research centres or institutions but could also be separately organised initiatives aimed specifically at Open Science.

The creation of metrics should take into consideration what elements of Open Science practices are being selected for training staff members. Given this potential diversity, and the absence of having generic or systematic data sources at hand for this aspect of Open Science, the data need to be collected by qualitative means, that is going through websites of institutions and/or faculties to identify Open Science training facilities.

This indicator can perhaps best be collected with information regarding Open Science support facilities, as training can be seen as part of such a support system.

Metrics

Number/% of institutions that offer OS training.

Open Science training might be often provided at the university level. This could vary from PhD courses to training targeted at employees more broadly. This metric provides an institutional view of the extent to which Open Science training facilities are offered.

MEASUREMENT.

This metric could be constructed by searching through university and faculty level websites to identify the presence of Open Science training facilities.

Potential problems and limitations are:

- This is a time-consuming effort, as one has to go through many websites, in particular when one is dealing with a large country, with many institutions, not necessarily limited to universities, but for example in the case of Germany, many publicly funded research institutions under the flag of Max Planck, Leibniz, Helmholtz, etc.
- The assessment of this indicator also requires mastering the language to the country under investigation as the general principles of Open Access can be expressed by different notions in different linguistic context.
- This kind of information might not be present on the website (yet), and as such create an underestimation of the actual situation.
- How does one compare the breadth of the Open Science training, here one only aims at the presence, but the range of facilities to train staff members on Open Science practices is here not yet included.

There are presently no generic and systematic data sources that contain such information.

Automation of this metric is currently not possible.

Number/% of national OS training initiatives.

In some cases, there might be national Open Science initiatives or other national institutions that provide Open Science training initiatives. This metric describes the extent to which there are national training initiatives.

MEASUREMENT.

This metric could be constructed by searching through policy initiatives that support Open Science practices on the national level and identify the training element out of the total of information collected.

Potential problems and limitations are:

- This is a time-consuming effort, as one has to go through many websites, in particular with respect to the definition of 'national'. In some countries, governance of science is not only a national issue, but might in the case of a federal government, also have to deal with federal initiative regarding the support of Open Science practices.
- A similar issue relates to the presence of funders, and their support of Open Science practices, these are also considered national research funding agencies, and how do these relate to the abovementioned issue of the 'national' level.
- How does one compare the breadth of the Open Science support, here one only aims at the presence, but the range of facilities to support Open Science is here not yet included.
- Language can be a problem here as well.

There are presently no generic and systematic data sources that contain such information.

Automation of this metric is currently not possible.

Breadth of Open Science training.

Training in Open Science can be broader or narrower. This metric provides an overview of the breadth of Open Science training, that is, to which extent it covers the diverse types of Open Science training facilities, such as publishing, research data, Open Code, peer review, pre-registration and registered reports.

MEASUREMENT.

Based on data that might be collected for the first two metrics, we might be able to construct a third metric. That is, when collecting information about training facilities, we could collect additional information. This metric could be constructed by searching through any relevant training facilities, either at the institutional or national level, as explained above.

Potential problems and limitations are:

- This is a time-consuming effort, as one has to go through many websites, on both overall institution level, as well as on faculty level.
- One has to decide upon a common denominator, the support for Open Science practices that are frequent across the system, but how does one deal with discipline specific support facilities that are not common?
- How does one include the various stages of development regarding Open Science across faculties into this metric? Some scholarly disciplines do have a different perspective on Open Science practices, e.g., regarding Open Access publishing of books in the humanities, or regarding qualitative research data in some parts of the social sciences, e.g., anthropology).

There are presently no generic and systematic data sources that contain such information.

Automation is currently not possible.

Duration of Open Science training (hour/day/week).

By providing an overview of the duration of Open Science training, we may capture the intensity of Open Science training. Having just a few hours of training, or more extensive training of course makes a difference with regards to the amount of knowledge that can be transferred.

MEASUREMENT.

This metric can be constructed in a similar manner as the previous metric, namely by looking at websites for what is offered as Open Science practice training facilities and collect the duration of such training facilities.

Potential problems and limitations are:

- This is a time-consuming effort, as one has to go through many websites, on both overall institution level, as well as on faculty level.
- Her one has to work with what is offered as Open Science training facilities, one does not have any idea of what is being 'consumed' from the offered training facilities?
- An issue that popped up previously, one has to be aware of the fact that perhaps not all of the Open Science support facilities are (yet) online visible on websites.

There are presently no generic and systematic data sources that contain such information.

Automation is currently not possible.

Prevalence of national Open Science policies

History

Version	Revision date	Revision	Author
1.2	2023-06-30	Minor edits	L. Stoy
1.1	2023-02-28	Extended content	L. Stoy
1.0	2023-02-14	Initial version of document	L. Stoy

Description

This indicator aims to capture the existence of country-level policies for Open Science. Since Open Science can be considered a catch-all term for different practices and objectives, as elaborated in the Recommendation on Open Science from the United Nations Educational, Scientific and Cultural Organization (UNESCO), this may include different elements or distinct policies such as Open Access to publications, Open Data, fair data, research assessment reform, Citizen Science etc. A

Usually, "national policy" will imply the policy being adopted by a country-level organisation with an important structuring role for the research and innovation sector. Edge cases could be federal countries in which some elements of decision-making and institutions relevant for research and innovations are devolved to regional or community level (e.g., Germany, Belgium, Switzerland). Practically, a policy could be adopted by a parliamentary decision, a ministry strategy, a government agency in charge of research and innovation, a national research council, an alliance of research organisations etc.

Measuring the proliferation of national Open Science policies is relevant for several reasons.

- First, knowing whether a specific policy exists and what it entails is crucial part of the impact analysis. A policy would take the place of the activity and intervention that, through specific outputs and outcomes, creates impact. In other words, it one of the first steps within Key Impact Pathways, in particular those of interest for national-level policy making.
- Second, there is an ongoing effort to monitor progress of Open Science policy implementation by proxy of national policies. This is evident, for example, in the UNESCO recommendation on Open Science, which asks member states to develop policies (Art. 17) and to invest in Open Science (Art. 18 and 19). In the European context, Open Science policies are monitored as part of the European Open Science Cloud (EOSC) initiative and the European

Research Area (ERA). In the case of EOSC, there is a specific ambition to also measure impact indicators in the future (O'Neill, 2022).

Metrics

Existence of Open Science policy

Measuring the existence of a national Open Science policy can be conceptualised at different levels and degrees of complexity.

'Basic': Existence of a country-level policy document or strategy on Open Science in general (yes/no)

'Intermediate': Existence of specific country-level policies on specific areas such Open Access to publications, FAIR/Open Data, Citizen Science, etc. (yes/no)

'Advanced': Specific measures (regulatory aspects, compliance levels, financial measures, target groups, implementing authority etc.) defined as part Open Science policy (Open Access to publications, FAIR/Open Data etc.)

Limitations but also advantages stemming from these approaches depend on the specific level:

- In the 'basic' case, the measurement can be conducted with relative ease. However, the lack of granularity of specific measures might limit the types of analyses that can be done with the data.
- In the 'intermediate' case, granularity is provided through the specific areas, which allows more detailed analyses than the 'basic' case. Data collection can be expected to be done relatively simply by capturing which topics are covered in the policy.
- The 'advanced' case will provide more detailed information about area, nature, and scope of Open Science policies which in turn allows more complex analyses. Data collection will be more challenging due to the need to define specific dimensions and their measurements and the need for an in-depth analysis of the source documents. Unfortunately, there are no controlled vocabularies for this type of data collection.

MEASUREMENT.

As outlined in the previous section, the exact measurement of the existence of a national policy for Open Science depends on the chosen approach.

In the basic case, the existence of a policy, which means the existence of a national (or perhaps, regional) policy and strategy supported by government actors and ideally stakeholders, can probably be captured with a simple binary variable with yes, there is a policy, and with no, there is no policy.

The only challenge would be to provide a common definition of what type of document constitutes a policy.

The intermediate case would work on a similar binary logic (yes/no) with the only difference being that this measurement is conducted for each Open Science policy area.

Only in the advanced case, additional refinements and more complex measurements and codes will be required. For example, differences in policy requirements (compliance, target groups), could be coded on an ordinal or nominal scale.

In practice, various studies have attempted to combine different levels of the conceptual approach outlined above.

SPARC Europe has regularly commissioned **reports on Open Science policies** in Europe. The latest report included information on the “scope, data definition, mandates, exceptions, mentions of FAIR, DMPs, data citation, data availability statements, re-use, Intellectual Property Rights (IPR) and licensing, and costs” (Sveinsdottir, Davidson and Proudman, 2021, p 10). For each element and several follow-up items, a binary or nominal coding was used:

- Definition of data: yes / not specified
- Mandates: unspecified / suggested or recommended / required
- Exceptions to data sharing: yes / unspecified
- Mentions of FAIR: yes / unspecified
- DMPs: recommended / required / other
- Expectation of data citation: yes / unspecified / other
- Data availability statements: yes / unspecified
- IPR: yes / unspecified
- Preferred licenses: yes / unspecified / other
- Costs for RDM: yes / unspecified / other

The information collected on behalf of SPARC Europe moreover contains textual or numeric information on additional structural information about the policies:

- Type of policy: statute, government ministry, funder policy
- Entry into force: year
- Sponsoring organisation
- Scope and coverage beyond data
- Link to Open Access: yes / no
- Soft or hard policy: soft / hard
- Coverage of skills and training: yes / no
- Monitoring and/or compliance: yes / no

Another case of standardized data collection on Open Science policies is the **EOSC Observatory** (<https://eoscobservatory.eosc-portal.eu/home>). The EOSC Observatory has been developed as part of the EOSC-Future project and aims to “support the EOSC community in tracking the implementation of EOSC and the policy makers in developing actionable policies”.

As part of the observatory, an EOSC Monitoring Framework has been developed which collects information, on a regular basis, together with representatives of EU member states and associated countries. The monitoring framework collects standardized information on 8 policy areas and the existence of country-level policies and accompanying investments in Open Science.

	Policies	Practices
Publications	Countries with a National Policy	Countries with National Monitoring
Data		
Software	Countries with a Financial Strategy	Countries with Use Cases
Services		
Infrastructure	Country RPOs with a Policy	Country Investments
Skills/Training	Country RFOs with a Policy	Country Outputs
Assessment		
Engagement		

Gareth O’Neill. (2022). *Monitoring Framework for National Contributions to EOSC (Version V1)*. Zenodo. <https://doi.org/10.5281/zenodo.7410762> (Licensed using [Creative Commons Attribution 4.0 International](#))

In the current framework, this matrix of 8 categories crossed with policies and practices areas results in 96 indicators, 48 each for policies and practices. Detailed items for each indicator are included in the “Survey on National Contributions to EOSC 2022” (O’Neill et al., 2023). Practically, the large majority of items are measured through binary yes-no questions. In the case of Open Access to publications, the following items are included in the survey:

- Existence of national policy on Open Access to publications: yes / no
 - Is this policy mandatory: yes / no
- Existence of a policy on immediate Open Access to publications: yes / no
 - Is this policy mandatory: yes / no
- Is there a specific policy on retention of IPR on publications: yes / no
 - Is this policy mandatory: yes / no

For each item, the questionnaire also includes an item about the public availability of the policy (measured as yes / no variable) and an open-text field to enter a URL to the corresponding website.

DATASOURCES EOSC OBSERVATORY

As introduced above, the EOSC Observatory is a “policy intelligence tool” which contains information about Open Science policies in EU member states and associated countries. The EOSC Observatory

is available as an online tool (<http://eoscobobservatory.eosc-portal.eu>). Underlying data is made reusable via a Zenodo community available at <https://zenodo.org/communities/eoscobservatory/>.

A specificity of the EOSC Observatory is that the data collected through the survey is directly answered and validated by member states and associated countries' representatives in the EOSC Steering Board. As part of the EOSC monitoring activities, the EOSC Observatory also provides crucial information for the monitoring of progress in the context of the European Research Area.

While originally started with the EOSC-Future project (2021-2023), the EOSC Observatory has secured follow-up funding through the Horizon Europe work programme for the next years. This makes it a stable source of information about the evolution of Open Science policies in the European Union and associated countries.

A detailed list of items asked for each policy category is included in the questionnaire. The latest version is available on Zenodo:

- Gareth O'Neill, Volker Beckmann, Sofia Abrahamsson, Thomas Neidenmark, & Stephan Siemen. (2023). Survey on National Contributions to EOSC 2022 (Version V1). Zenodo. <https://doi.org/10.5281/zenodo.7550798>

In addition, data is made available openly. The latest version is shared in .tsv format and therefore easily reusable. This availability also means that data can be re-coded with relatively little effort to adapt for specific research questions. Moreover, the data contains links to relevant policy sources, which reduces the effort to collect this information by researchers re-using the data.

- Gareth O'Neill, & Stefania Martziou. (2022). Data of Survey on National Contributions to EOSC 2021 (Version V1) [Data set]. Zenodo. <https://doi.org/10.5281/zenodo.7431678>

SPARC EUROPE

SPARC Europe has commissioned several studies on national Open Science policies in Europe. The study is regularly updated and is now available in its 7th version.

- Sveinsdottir, Thordis, Davidson, Joy, & Proudman, Vanessa. (2021). An Analysis of Open Science Policies in Europe, v7 (Version 7). Zenodo. <https://doi.org/10.5281/zenodo.4725817>

The data includes information on variables such as

- Definition of data: yes / not specified
- Mandates: unspecified / suggested or recommended / required
- Exceptions to data sharing: yes / unspecified
- Mentions of FAIR: yes / unspecified
- DMPs: recommended / required / other
- Expectation of data citation: yes / unspecified / other
- Data availability statements: yes / unspecified

- IPR: yes / unspecified
- Preferred licenses: yes / unspecified / other
- Costs for RDM: yes / unspecified / other
- Type of policy: statute, government ministry, funder policy
- Entry into force: year
- Sponsoring organisation
- Scope and coverage beyond data:
- Link to Open Access: yes / no
- Soft or hard policy: soft / hard
- Coverage of skills and training: yes / no
- Monitoring and/or compliance: yes / no

The information is presented as narrative content of the study report, although several tables are included too. It is not available in spreadsheet format publicly. In addition, the reports contain country fiches which provide narrative data about national Open Science policies. In the seventh and latest version, the specific focus was on FAIR and Open Data. Previous versions focussed on other areas, such as Open Access to publications.

OPENAIRE

OpenAIRE is providing information on national Open Science policies via its country pages (<https://www.openaire.eu/os-eu-countries>). The country pages are maintained by the respective National Open Access desks (NOAD). The information is split into several categories:

- Overview
- Open Science Policy
- Infrastructure & EOSC
- Training & Support
- Statistics
- News

This information is unstructured and not presented in a standardized fashion to facilitate re-use.

COUNCIL FOR NATIONAL OPEN SCIENCE COORDINATION (CoNOSC)

The Council for National Open Science Coordination (CoNOSC) is providing information about member countries' national Open Science system and policies at <https://conosc.org/os-policies/#page-content>.

This information is unstructured and not presented in a standardized fashion to facilitate re-use.

NATIONAL POINTS OF REFERENCE ON SCIENTIFIC INFORMATION

The European Expert Group on National Points of Reference on Scientific Information provides occasional updates about the state of Open Access and Open Science policies in EU member states and associated countries. Information is available via the EU expert group registry: <https://ec.europa.eu/transparency/expert-groups-register/screen/expert-groups/consult?lang=en&groupID=3477>

Reports of the group do contain qualitative and quantitative information, which is however not made available in standardized formats.

ROARMAP

The ROARMAP database contains information about policies concerning In total 87 funder-only policies are registered.

Open Access to publications from several types of organisations worldwide (<https://roarmap.eprints.org>) . For national Open Science policies, most relevant might be the category of research funder, which includes national research councils, government agencies etc. The database contains more than 770 policies at the time of writing. Data is available in standardized formats and with a unique identifier. All entries are reviewed by the curators of ROARMAP.

Information collected about policies includes attributes such as adoption, effective, and revision date; content types specified by the mandate; Open Access conditions; link to research evaluation; rights retention; etc.

References

O'Neill, Gareth. 'Monitoring Framework for National Contributions to EOSC'. Zenodo, 14 December 2022. <https://doi.org/10.5281/zenodo.7410762>.

O'Neill, Gareth, Volker Beckmann, Sofia Abrahamsson, Thomas Neidenmark, and Stephan Siemen. 'Survey on National Contributions to EOSC 2022', 19 January 2023. <https://doi.org/10.5281/zenodo.7550798>.

Sveinsdottir, Thordis, Joy Davidson, and Vanessa Proudman. 'An Analysis of Open Science Policies in Europe, V7'. Zenodo, 28 April 2021. <https://doi.org/10.5281/zenodo.4725817>.

UNESCO Recommendation on Open Science. <https://unesdoc.unesco.org/ark:/48223/pf0000379949.locale=en>

Prevalence of open/FAIR data practices

History

Version	Revision date	Revision	Author
1.1	2023-07-20	Edited & revised	T. Willemse
1.0	2023-04-13	First draft	T. Willemse

Description

The Open Science discourse has been fuelled by the prospect of a more open and collaborative scientific effort that can accelerate scientific development and innovation. A big step in this direction is the development of Open Data practices that make it easier for scientist to share and reuse data for research. Often this agenda is pushed forward with the Findability, Accessibility, Interoperability and Reusability (FAIR) principles (Wilkinson et al., 2016). Taking these principles in mind serves to open up data practices in science and thereby improve scientific data practices, data reuse and reproducibility. Accordingly, it is important to get an indication of the prevalence of such practices in the scientific system to get an overview of the status of data sharing and Open Science in general.

The FAIR principles and Open Data management try to establish a data environment in which high quality data is easily accessible in the long term and where this data can be simply discovered, evaluated and reused (Wilkinson et al., 2016). Making data more findable could be achieved by using identifiers, adding rich metadata and registration in a searchable resource. Accessibility could be improved by data being retrieved by their identifier in a standardized format, as well as by keeping metadata accessible even if the data is no longer available. Interoperability could be enhanced by using applicable language and vocabularies along with qualified references to other data. Reusability of data can be increased by unambiguous and comprehensive storing and describing practices.

Different stakeholders, such as researchers, data publishers and funding agencies stand to benefit from these practices. More insight in the application and presence of FAIR data principles could be very relevant in their profession. Questions typically relate to how to improve the implementation and development of FAIR principles. In order to improve (FAIR) data sharing practices, it is important to first have an overview of the current practices. Hence, relevant questions are where, how and what FAIR data practices are used in an area of interest.

Metrics

Number of publications with shared data

The number of publications with shared data can serve as a quick measure to assess to what extent Open Data practices are used in the area of interest. It does however not take into account how and to what extent the FAIR principles were followed and the nature of the data itself. Due to these shortcomings, it can give a skewed representation in areas where poor quality or partly available data is documented as shared data. An additional challenge is to identify not only shared data that is shared through official repositories, but also data that is shared as supplementary material.

In similar fashion the percentage of publications with shared data in the area of interest can give a quick representation of how widespread the use of Open Data practices is. When looking specifically at the prevalence of Open Data practices this would be the preferred metric over the total number of publications with shared data. However, the percentage measure suffers the same shortcomings as mentioned before for the total number.

In addition, it is important to note that more targeted indications can be used than if a publication shares data or not. Alternatives like number of publications with data shared in a repository, data availability statements or including the level of fairness of the shared data can be used to reach more specific results.

MEASUREMENT.

The number of publications with shared data can be represented as a simple count measure of the sample of interest. The percentage can be calculated by dividing the total number of publications with shared data with the total number of units included.

DATASOURCES

[DATACITE/CROSSREF](#)

DataCite and Crossref are both organizations that provide services for identifying and citing research data. They maintain large databases of metadata about research articles and associated datasets, including information on Open Access data. Besides providing DOIs for datasets, DataCite and Crossref also maintain metadata about the datasets, including information on data availability, access restrictions, and licensing information. This metadata can be accessed programmatically through APIs provided by DataCite and Crossref, making it a valuable data source for researchers interested in Open Access data.

EXISTING METHODOLOGIES

[EXTRACT DATASET SHARING BASED ON NATURAL LANGUAGE PROCESSING](#)

The Public Library of Science (PLOS) is a non-profit publisher of open-access journals. PLOS provides indicators on data repository use in PLOS articles as well as overall data repository use. PLOS uses a combination of manual curation and automated methods to generate information on Open Access

data. This includes reviewing data availability statements provided by authors, checking data repositories for publicly accessible data associated with articles, and using natural language processing and machine learning algorithms to identify mentions of data availability in articles. PLOS also encourages authors to provide detailed data availability statements and to deposit their data in public repositories to facilitate Open Data access.

PLOS has developed the indicators on data sharing and data use through DataSeer. DataSeer provides a Natural Language Processing (NLP) and AI backed algorithm that can automatically link data sources to doi's and check if these data sources are Open Source. Both the [machine learning](#) code and [web app](#) code are openly available.

PLOS also provides API's to search its database. This [page](#) provides some example Solr queries, the specific queries will depend on the research question.

Level of FAIRness of data

Metrics on the level of FAIRness of data (sources) can support in establishing the prevalence of open/FAIR data practices. This metric attempts to show in a more nuanced manner where FAIR data practices are used and in some cases even to what extent they are used. Assessing whether or not a data source practices FAIR principles is not trivial with a quick glance, but there are initiatives that developed methodologies that assist to determine this for (a large number of) data sources.

MEASUREMENT.

EXISTING METHODOLOGIES

RESEARCH DATA ALLIANCE

The Research Data Alliance developed a [FAIR Data Maturity Model](#) that can help to assess whether or not data adheres to the FAIR principles. This document is not meant to be a normative model, but provide guidelines for informed assessment.

The [document](#) includes a set of indicators for each of the four FAIR principles that can be used to assess whether or not the principles are met. Each indicator is described in detail and its relevance is annotated (essential, important or useful). The model recommends to evaluate the maturity of each indicator with the following set of maturity categories:

0 – not applicable

1 – not being considered yet

2 – under consideration or in planning phase

3 – in implementation phase

4 – fully implemented

By following this methodology, one could assess to what extent the FAIR data practices are adhered to and create comprehensive overviews, for instance by showing the scores in radar charts.

Data life cycle assessment

Determining the level of FAIR data practices can involve assessing how well data adheres to the FAIR principles at each stage of the data lifecycle, from creation to sharing and reuse (Jacob, 2019).

Identify the stages of the data lifecycle: The data lifecycle typically includes stages such as planning, collection, processing, analysis, curation, sharing, and reuse. Identify the stages that are relevant to the data to be assessed.

Evaluate adherence to FAIR principles at each stage: For each stage of the data lifecycle, evaluate the extent to which the data adheres to the FAIR principles. Use for instance the FAIR Data Maturity Model to score the adherence to the FAIR principles, assign a score for each principle and stage of the data lifecycle.

Determine the overall level of FAIR data practices: Once the scores for each principle and stage have been assigned, determine the overall level of FAIR data practices. This can be done by using a summary score that takes into account the scores for each principle and stage, or by assigning a level of FAIR data practices based on the average score across the principles and stages.

Availability of data statement

A data availability statement in a publication describes how the reader could get access to the data of the research. Having such a statement in place improves transparency on data availability and can thus be considered as an Open Data practice. However, having a data availability statement in place does not necessarily imply that the data is openly available or that it is more likely that the data can be shared (Gabelica et al., 2022). Nevertheless, a description of how to access an Open Data repository, how to make a request for data access or an explanation why some data cannot be shared due to ethical considerations are all examples of Open Data practices that make data reuse more accessible and transparent (Federer et al., 2018). The availability of a data statement can therefore be considered as an Open Data practice.

MEASUREMENT

EXISTING METHODOLOGY

All PLOS journals require publications to include a data availability statement. Moreover, it is strongly recommended that procedures on how to access research data are described in the data availability statement and that the data is stored in a public repository. Other practices that comply with this recommendation are including a data file, data requests through an approving committee and providing contact information for a third party that owns the data (Federer et al., 2018). A detailed

description of how to use PLOS data availability statements for quantitative research can be found in the (Colavizza et al., 2020) publication.

Known correlates

Some research suggests that openly sharing data is positively related to the citation rate of publications (Piwowar et al., 2007; Piwowar & Vision, 2013).

References

Colavizza, G., Hrynaszkiewicz, I., Staden, I., Whitaker, K., & McGillivray, B. (2020). The citation advantage of linking publications to research data. *PloS one*, 15(4), e0230416.

Federer, L. M., Belter, C. W., Joubert, D. J., Livinski, A., Lu, Y. L., Snyders, L. N., & Thompson, H. (2018). Data sharing in PLOS ONE: an analysis of data availability statements. *PloS one*, 13(5), e0194768.

Gabelica, M., Bojčić, R., & Puljak, L. (2022). Many researchers were not compliant with their published data sharing statement: a mixed-methods study. *Journal of Clinical Epidemiology*, 150, 33-41.

Jacob, D. (2019, April). FAIR principles, an new opportunity to improve the data lifecycle. In ado2019: journée thématique sur les autorités de données.

Piwowar, H. A., Day, R. S., & Fridsma, D. B. (2007). Sharing Detailed Research Data Is Associated with Increased Citation Rate. *PLoS ONE*, 2(3), e308. <https://doi.org/10.1371/journal.pone.0000308>

Piwowar, H. A., & Vision, T. J. (2013). Data reuse and the Open Data citation advantage. *PeerJ*, 1, e175. <https://doi.org/10.7717/peerj.175>

Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.-W., da Silva Santos, L. B., Bourne, P. E., Bouwman, J., Brookes, A. J., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, C. T., Finkers, R., ... Mons, B. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, 3(1), Article 1. <https://doi.org/10.1038/sdata.2016.18>

Prevalence of Open Method practices

History

Version	Revision date	Revision	Author
1.1	2023-08-29	Draft for initial publication	P. Stavropoulos
1.0	2023-02-12	First draft	I. Grypari, N. Manola, H. Papageorgiou, P. Stavropoulos

Description

*"Methods describe the processes, procedures, and materials used in a research investigation. Methods can take many forms depending on the field and approach, including study designs, protocols, code, materials and reagents, databases and more."*¹

Open methods refer to:

1. Open Access to the various elements of the scientific method (datasets, software/code, protocols, materials, etc.)
2. FAIRness of the same elements (i.e. following the FAIR principles), and,
3. an Open and FAIR documentation, that facilitates reproducibility and reuse of the study methods.

Metrics

Number/share of publications with Open Source code

Description: Number of publications with Open Source code: This is the number of publications for which the code that was developed in this context to produce the scientific outcomes included in the publication is shared/can be found with Open Access in Open Repositories.

¹ <https://plos.org/open-science/open-methods/>

Benefits: Providing Open Access to software/code of the scientific method is of key importance of understanding, evaluating, replicating and extending the study.

Limitations: In most cases, detailed documentation must accompany the code including data management, cleansing and pre-processing, configuration files and workflows.

Differences: ...?

MEASUREMENT.

There is no comprehensive list of publication-software pairs. The metric is partially covered by well-established repositories like GitHub or existing data sources like OpenAIRE. Automated approaches have been built (e.g., PapersWithCode) or are under development in the scientific community, scanning the publication text and spotting mentions of software/code and its FAIR metadata.

There is no well-established methodology to automatically find publications with code. Some suggested methods are the following:

1. Text mining and machine learning algorithm to find mentions of Open Software/Code and its FAIR metadata in the publication text.
2. Search relevant fields or tags in specific databases to find publications with software/code.

Prevalence of Open Peer Review

History

Version	Revision date	Revision	Author
1.0	2023-05-21	Initial draft	L. Waltman

Description

Peer review of research articles and other scientific works is often seen as essential for the trustworthiness of the scientific literature. Traditionally peer review is a closed process. Review reports are not published. They are made available only to the authors of a scientific work and to editors that need to decide whether the work is suitable for publication in a scientific journal or some other publication venue. Readers of a scientific work do not have access to review reports. In addition, reviewers are anonymous in a traditional peer review process. Authors and readers of a scientific work do not know by whom the work was reviewed. Reviewing a scientific work also requires an invitation from an editor. Without such an invitation, it is not possible to participate in a traditional peer review process.

There is increasing support for more open approaches to peer review. Open peer review, sometimes called transparent peer review, is an umbrella term that refers to various forms of openness in peer review ([ref](#)). It often refers to the publication of review reports. One approach is to publish review reports only if the outcome of a peer review process is positive and the scientific work under review is considered suitable for publication in a scientific journal or some other publication venue. Another approach is to always publish review reports, also if the outcome of a peer review process is negative. Open peer review may also refer to the publication of the identities of reviewers, but this form of openness is more controversial. Another possibility is to open participation in peer review by allowing anyone with certain minimum qualifications to participate.

Our focus is on Open Peer Review of research articles. This can be either Open Peer Review organized by scientific journals or other forms of Open Peer Review, in particular Open Peer Review of research articles published on preprint servers ([ref](#)). We do not consider Open Peer Review of other scientific works, such as books and conference contributions. In addition, we restrict ourselves to openness of review reports, since this is the most popular form of openness in peer review.

Metrics

Journals supporting Open Review reports

This metric provides the number or the percentage of journals that support Open Review reports. A further distinction can be made between journals that publish review reports for all articles and journals that publish review reports only for articles for which the authors and/or the reviewers agree with the publication of review reports. Another distinction that can be made is between journals that publish review reports only for accepted articles and journals that publish review reports for all articles that undergo peer review, including articles for which the outcome of the peer review process is negative.

Open review reports do not need to include the identities of the reviewers. Reviewers may remain anonymous.

MEASUREMENT.

There is no datasource that provides comprehensive data on the peer review models used by journals. However, there are a few datasources that provide partial data on journals that support Open Review reports:

- Transpose (TRANsparency in Scholarly Publishing for Open Scholarship Evolution): <https://transpose-publishing.github.io/>. Transpose is a database of journal policies, focusing on open peer review, co-reviewing, and preprinting policies. At the moment (July 2023), the database includes 3168 journals. A subset of these journals support open review reports. Many journals are not included in the database, and therefore the data on journals that support open review reports is incomplete.
- ASAPbio: <https://asapbio.org/letter>. ASAPbio maintains a list of journals that support open review reports. At the moment (July 2023), the list includes 377 journals. The completeness of the list is not clear.
- Data set compiled by Dietmar Wolfram and colleagues: <https://doi.org/10.5281/zenodo.3737197>. In April 2020, Wolfram and colleagues published a data set of 617 journals that use some form of open peer review. Many of these journals support open review reports. The data set has not been updated after April 2020.

Each of the above datasources is incomplete and some of the data is likely to be outdated.

Journal articles with Open Review reports

This metric provides the number or the percentage of journal articles that have Open Review reports. Open review reports do not need to include the identities of the reviewers. Reviewers may remain anonymous.

MEASUREMENT.

Some journals publish review reports and register DOIs for these reports at Crossref or DataCite. For these journals, metadata from Crossref and DataCite can be used to determine the number of articles with open review reports. This approach was used in [this blog post](#), in which it was shown that eLife, PeerJ, and Wiley journals publish a relatively large number of articles with open review reports. For an article with a Crossref DOI, the Crossref API (see <https://api.crossref.org/swagger-ui/index.html>) can be used to identify links from the article to open review reports. If the metadata of the article includes links to open review reports, these links can be found in the 'has-review' field in the API output. Conversely, for an open review report with a Crossref DOI, a link to the corresponding article can be found in the 'is-review-of' field in the API output.

Many journals publish review reports without registering DOIs for these reports. This is for instance the case for journals published by BMJ, EMBO, MDPI, PLOS, and Springer Nature. For these journals, there is no straightforward way to determine the number of articles with open review reports. However, if a journal is known (based on the datasources mentioned in the previous section) to publish open review reports for all its articles, the number of articles with open review reports can be determined by determining the total number of articles in the journal.

Preprints with Open Review reports

This metric provides the number or the percentage of preprints that have Open Review reports. There are many different forms in which feedback can be given on preprints, ranging from brief informal comments to detailed formal feedback. A decision needs to be made on which forms of feedback are considered to constitute peer review ([ref](#)).

Open review reports do not need to include the identities of the reviewers. Reviewers may remain anonymous.

MEASUREMENT.

There are a substantial number of services for peer review of preprints, especially in the life sciences ([ref](#)). While these services all provide Open Review reports, most of them do not register DOIs for these reports. For many of these services, data on review reports can be obtained from Sciety, an aggregator of Open Review reports for preprints. Sciety can be accessed through a website. An API is not yet publicly available.

For preprint review services that do register DOIs for review reports, data can be obtained from Crossref and DataCite, as shown in [this blog post](#). For a review report with a Crossref DOI, the Crossref API (see <https://api.crossref.org/swagger-ui/index.html>) can be used to identify a link from the review report to the corresponding preprint. The link can be found in the 'is-review-of' field in the API output.

When compiling statistics on preprints with open review reports, it needs to be decided which preprint review services are included in the compilation of statistics on preprints with Open Review reports. Some of the services covered by Sciety for instance operate in a fully algorithmic way and do not provide review reports written by humans. A decision could be made to exclude these services.

Some preprint servers allow review reports and other comments to be posted directly on the preprint server (rather than on a separate platform for preprint peer review). However, there is no straightforward way to identify preprints for which review reports are available directly on the preprint server.

Prevalence of Open Science funding policies

History

Version	Revision date	Revision	Author
1.1	2023-06-08	Second draft	L. Stoy, I. Karasz
1.0	2023-03-01	First draft	L. Stoy

Description

This indicator aims to capture the prevalence of funding being provided for Open Science. It is closely related to the indicator for prevalence of national policies for Open Science. Since Open Science can be considered a catch-all term for different practices and objectives, this may include different elements or distinct policies such as Open Access to publications, Open Data, fair data, research assessment reform, Citizen Science etc. Usually, the question whether funding is provided is part of a policy. I.e., the availability of funding implies the existence of a policy as well.

Practically, funding might be available at various levels. It can be international, national, regional, institutional (RPOs, HEIs, RFOs etc.) or even departmental..

Knowing the existence of Open Science funding policies can be helpful to evaluate the overall commitment of funding organizations, governments, and private entities to advancing Open Science practices and fostering a culture of transparency and collaboration in research. Researchers can assess the extent to which funding agencies prioritize Open Access publishing, data sharing, and other Open Science practices, which can influence their decisions on where to submit grant proposals and seek financial support for their projects.

Furthermore, being aware of these policies enables researchers to align their research projects with the values and objectives of funding organizations that actively promote Open Science. By adhering to Open Science principles in their proposals and research activities, researchers can increase their chances of securing funding and potentially gain access to additional resources, such as Open Science infrastructure and collaborative networks.

Beyond individual researchers, understanding the landscape of Open Science funding policies can also inform policymakers, institutions, and advocacy groups about the progress and impact of Open Science initiatives. It allows them to identify areas where more support and resources are needed to strengthen the Open Science ecosystem and accelerate the transition to a more open and transparent research culture. Paired with impact and other Open Science indicators, the existence of policies and funding can be expected to be an important determinant for Open Science practices to increase.

In summary, knowledge of Open Science funding policies empowers researchers and stakeholders to make informed decisions, promote Open Science practices, and contribute to a more inclusive and accessible global research community.

Metrics

Public supranational Open Science funding policies

Public supranational Open Science funding policies refer to the funding initiatives and programs that are organized and administered at a level beyond that of individual countries or regions. These funding policies are typically established by international organizations or supranational entities that aim to promote Open Science practices on a global scale. Here are some key characteristics of public supranational Open Science funding policies:

- **Global Collaboration:** Supranational Open Science funding policies foster collaboration among researchers, institutions, and countries worldwide. The focus is on encouraging projects that bring together expertise from diverse regions to address global challenges and advance scientific knowledge.
- **Open Access and Open Data:** One of the primary objectives of supranational Open Science funding policies is to promote Open Access to research publications and open sharing of research data. This means that funded research should be published in open-access journals or repositories, and research data should be made available to the public for further analysis and reuse.
- **Cross-Disciplinary Research:** Funding policies at the supranational level often support cross-disciplinary research projects that combine insights and methodologies from different scientific fields. This approach aims to foster innovation and address complex societal problems from multiple angles.
- **Support for Developing Countries:** Supranational Open Science funding policies may include specific provisions to support researchers and institutions in developing countries. These provisions aim to reduce disparities in access to research funding and resources and promote knowledge exchange between developed and developing regions.
- **Alignment with Global Agendas:** Funding policies at the supranational level are often designed to align with and contribute to global agendas and initiatives, such as the United Nations Sustainable Development Goals (SDGs). Researchers are encouraged to address pressing global challenges, such as climate change, health, and poverty alleviation.
- **Open Science Infrastructure:** Public supranational funding initiatives may allocate resources to develop and maintain Open Science infrastructure, including data repositories, collaborative platforms, and tools that facilitate research collaboration and data sharing on an international scale.

- Ethical Considerations: Supranational Open Science funding policies may emphasize the importance of adhering to ethical standards in research, including data privacy, research integrity, and the protection of research participants' rights.

Examples of organizations that promote supranational Open Science funding policies include the European Union through its research and innovation programs like Horizon Europe, as well as other international bodies such as the United Nations Educational, Scientific and Cultural Organization (UNESCO) and the World Health Organization (WHO).

DATASOURCES

UNESCO RECOMMENDATIONS

<https://unesdoc.unesco.org/ark:/48223/pf0000383709>

IMPLEMENTATION OF THE UNESCO RECOMMENDATIONS

<https://www.unesco.org/en/open-science/implementation#open-science-working-groups>

ROAR MAP

https://roarmap.eprints.org/cgi/search/archive/advanced?screen=Search&dataset=archive&_action_search=Search&policymaker_type=funder&policymaker_name_merge=ALL&policymaker_name=&policy_adoption=&policy_effecive=&mandate_content_types_merge=ANY&apc_fun_url_merge=ALL&apc_fun_url=&satisfyall=ALL&order=policymaker_name

Public European Open Science funding policies

Please provide a general description of the metric. Please provide an argument why this metric is a good operationalisation of the indicator. Please also describe any limitations of this metric. Highlight differences with respect to other possible metrics.

European Open Science funding policies are an essential component of the European Union's research and innovation strategy. Open Science refers to the practice of making research findings, data, and methodologies openly accessible and transparent to the broader scientific community and the public. The EU is committed to fostering an open and collaborative research environment to accelerate scientific progress, improve the reproducibility of research, and maximize the societal impact of publicly funded research. Here are some key aspects of public European Open Science funding policies:

Open Access to Publications: European Open Science funding policies encourage researchers to publish their research findings in open-access journals or repositories. Open Access allows anyone, anywhere, to access and read the published research without facing paywalls or subscription fees, promoting the dissemination of knowledge.

Open Research Data: Funding policies support the sharing of research data generated through publicly funded projects. Researchers are encouraged to deposit their data in Open Repositories, making it possible for other researchers to reuse and validate the data, thus enhancing the reproducibility and reliability of scientific results.

Data Management Plans: When applying for research funding, applicants are often required to submit a data management plan (DMP). A DMP outlines how researchers will handle, store, and share their data during and after the project. This plan ensures that research data is managed in accordance with Open Science principles.

Collaboration and Networking: Open Science funding policies may prioritize projects that foster collaboration and networking between researchers, institutions, and countries. International cooperation is encouraged to leverage diverse expertise and resources in addressing global challenges.

Open Science Tools and Infrastructure: Funding programs may support the development and maintenance of Open Science tools, platforms, and infrastructure. These resources facilitate data sharing, collaboration, and the use of open-source software and methodologies.

Citizen Science and Public Involvement: Some funding initiatives may promote Citizen Science projects, where the public actively participates in research activities. Public engagement and involvement in research are seen as ways to strengthen the link between science and society.

Compliance and Evaluation: Funders may require grantees to comply with Open Science principles as a condition for funding. Additionally, the impact of research funded through Open Science policies is evaluated not only based on academic metrics but also on its broader societal and economic impacts.

DATASOURCES

EOSC OBSERVATORY

<https://zenodo.org/communities/eoscobservatory/?page=1&size=20>

The datasource has a special focus on EOSC related investments which is the backbone of the European infrastructure of OS. It directly gathers data from Member States and Associated Countries on a yearly basis through individual surveys and then in-depth interviews within the frames of the EOSC Future project. After 2023 September, the information gathering activity is questionable.

EXISTING METHODOLOGIES

INTERVIEWS, SURVEYS

ROAR MAP

https://roarmap.eprints.org/cgi/search/archive/advanced?screen=Search&dataset=archive&_action_search=Search&policymaker_type=funder&policymaker_name_merge=ALL&policymaker_name=

&policy_adoption=&policy_effecive=&mandate_content_types_merge=ANY&apc_fun_url_merge=ALL&apc_fun_url=&satisfyall=ALL&order=policymaker_name

Public national Open Science funding policies

Public national Open Science funding policies are specific strategies and guidelines established by individual countries to support and promote Open Science practices within their research funding programs. Open Science aims to increase transparency, accessibility, and collaboration in the scientific community by making research findings, data, and methodologies openly available to the public and other researchers. Here are some key characteristics of public national Open Science funding policies:

- **Open Access to Publications:** Funding policies encourage or mandate researchers to publish their research findings in open-access journals or deposit them in Open Repositories. Open Access ensures that research outputs are freely accessible to anyone, facilitating the dissemination of knowledge without financial barriers.
- **Data Sharing and Management:** National Open Science funding policies emphasize the importance of data sharing and may require researchers to develop data management plans (DMPs). DMPs outline how research data will be collected, organized, stored, and made available to other researchers and the public.
- **Research Data Repositories:** National funding policies may provide resources to establish and maintain research data repositories where researchers can deposit and share their data in a standardized and accessible manner.
- **Reproducibility and Transparency:** Open Science funding policies promote research reproducibility by encouraging researchers to share their methodologies, code, and analytical workflows. This transparency allows others to verify and build upon existing research.
- **Collaborative Research Platforms:** Funding initiatives may support the development of collaborative research platforms and tools that enable researchers to work together and share data across institutions and disciplines.
- **Open Educational Resources (OER):** Some funding policies extend to supporting the creation and sharing of Open Educational Resources, such as course materials and textbooks. OER can benefit educators and learners by providing free and accessible learning materials.
- **Citizen Science and Public Engagement:** Public national Open Science funding policies may include provisions for Citizen Science projects, where the public actively participates in research activities, fostering greater public engagement and involvement in the scientific process.
- **Evaluation and Incentives:** Funding agencies may consider a researcher's commitment to Open Science principles when evaluating grant proposals, promoting the adoption of Open Science practices in the academic community. Moreover, researchers who engage in Open Science may receive recognition and incentives in funding decisions.

It's important to note that the specific details and extent of Open Science policies may vary among different countries, reflecting the unique research landscape and priorities of each nation. Researchers and applicants seeking funding through national research programs should refer to the official guidelines and requirements provided by their respective funding agencies.

Private global Open Science funding policies

Private global Open Science funding policies refer to the strategies and initiatives established by private organizations and foundations with a global reach to support and promote Open Science practices. These organizations recognize the importance of open and collaborative research in advancing scientific knowledge, addressing global challenges, and maximizing the societal impact of research. Private global Open Science funding policies often complement public funding efforts and provide additional resources and support to researchers worldwide. Here are some key characteristics of private global Open Science funding policies:

- **Open Access Publishing:** Private global Open Science funding policies may support Open Access publishing initiatives, where researchers are encouraged or required to publish their research findings in open-access journals. By doing so, research outputs become freely available to the public without subscription or payment barriers.
- **Open Data and Data Sharing:** Funding policies emphasize the sharing of research data and support efforts to create and maintain Open Data repositories. Researchers are encouraged to share their data openly, allowing other scientists to access and reuse the data for further analysis and validation.
- **Collaborative Research Platforms:** Private organizations may invest in collaborative research platforms and tools that facilitate data sharing, knowledge exchange, and interdisciplinary collaboration among researchers worldwide.
- **Research Fellowships and Grants:** Private global Open Science funding policies offer research fellowships and grants to individual researchers and research teams to conduct Open Science projects. These funding opportunities may have specific requirements for data sharing and Open Access to research outputs.
- **Open Science Prizes and Awards:** Private organizations may establish awards and prizes to recognize researchers and institutions that demonstrate exceptional commitment to Open Science principles. These awards serve to incentivize and celebrate Open Science practices.
- **Open Educational Resources (OER):** Some private organizations with a focus on global Open Science funding may also support the creation and dissemination of Open Educational Resources, making learning materials more accessible worldwide.
- **Global Challenges and Impact:** Private global Open Science funding policies often prioritize research that addresses pressing global challenges, such as climate change, healthcare, poverty, and education. The aim is to support research that can have a positive and meaningful impact on society and the environment.

- **Public-Private Partnerships:** Private organizations may collaborate with public entities, academic institutions, and other stakeholders to establish joint initiatives and funding programs that promote Open Science practices on a global scale.

It's important to note that private global Open Science funding policies may vary widely among different organizations and foundations. Researchers seeking funding from private entities should carefully review the specific guidelines, requirements, and focus areas of each funding opportunity to ensure their research aligns with the objectives of the funding program.

DATASOURCES

ROAR MAP

https://roarmap.eprints.org/cgi/search/archive/advanced?screen=Search&dataset=archive&_action_search=Search&policymaker_type=funder&policymaker_name_merge=ALL&policymaker_name=&policy_adoption=&policy_effective=&mandate_content_types_merge=ANY&apc_fun_url_merge=ALL&apc_fun_url=&satisfyall=ALL&order=policymaker_name

Private national Open Science funding policies

Private national Open Science funding policies refer to the strategies and guidelines established by private organizations or corporations within individual countries to support and promote Open Science practices. These funding policies are separate from public national research funding and are provided by private entities that recognize the value of Open Science principles in advancing research, innovation, and knowledge sharing. Private national Open Science funding policies typically focus on specific research areas or industries, complementing public funding efforts and contributing to the broader Open Science ecosystem. Here are some key characteristics of private national Open Science funding policies:

- **Open Access Initiatives:** Private national Open Science funding policies may support Open Access publishing, where researchers are encouraged or required to publish their research findings in open-access journals. This ensures that the research is freely available to the public without subscription fees or access barriers.
- **Research Data Sharing:** Private funding policies may incentivize or mandate researchers to share their research data openly, either through public repositories or other designated platforms. Open Data sharing enhances research transparency and reproducibility.
- **Open Science Tools and Platforms:** Private organizations may invest in the development and maintenance of Open Science tools and platforms that facilitate collaboration, data sharing, and knowledge exchange among researchers.
- **Collaborative Research Projects:** Private national Open Science funding may support collaborative research projects involving multiple research institutions or interdisciplinary teams. These projects often encourage open collaboration and data sharing.

- Industry-Specific Initiatives: Some private funding policies may be industry-specific, targeting research and innovation in particular sectors, such as technology, healthcare, energy, or agriculture. These policies may have Open Science requirements tailored to the specific needs and characteristics of the industry.
- Research Fellowships and Grants: Private organizations may offer research fellowships and grants to individual researchers or research teams working on Open Science projects. These funding opportunities may prioritize projects that align with Open Science principles.
- Open Educational Resources (OER): Private national Open Science funding policies may also support the creation and dissemination of Open Educational Resources, making educational materials freely accessible to learners.
- Social and Environmental Impact: Private organizations may prioritize funding research projects that have a positive impact on society or the environment, aligning with their corporate social responsibility (CSR) objectives.

It's essential to note that private national Open Science funding policies can vary significantly among different private organizations and industries within a country. Researchers seeking private funding should carefully review the specific guidelines, focus areas, and requirements of each funding opportunity to ensure their research aligns with the objectives of the funding program. Additionally, policies and initiatives may have evolved or changed since my last update, so researchers are encouraged to refer to the most recent guidelines provided by private funding organizations.

DATASOURCES

ROAR MAP

https://roarmap.eprints.org/cgi/search/archive/advanced?screen=Search&dataset=archive&_action_search=Search& policymaker_type=funder& policymaker_name_merge=ALL& policymaker_name=& policy_adoption=& policy_effecive=& mandate_content_types_merge=ANY& apc_fun_url_merge=ALL& apc_fun_url=& satisfyall=ALL& order=policymaker_name

Notes

This is an ongoing process that needs time until it can be considered as elaborated.

References

Beckmann, V., Abrahamsson, S., Neidenmark, T., Siemen S., O'Neill, G. (2021). Survey on National Contributions to EOSC. *EOSC Observatory Channel*, <https://zenodo.org/record/7423953#.ZIHWdS9BzfY>

Beckmann, V., Abrahamsson, S., Neidenmark, T., Siemen S., O'Neill, G. (2022). Survey on National Contributions to EOSC. *EOSC Observatory Channel*, <https://zenodo.org/record/7550798#.ZIHai9BzfY>

O'Neill, G. (2022). Monitoring Framework for National Contributions to EOSC. *EOSC Observatory Channel*, <https://zenodo.org/record/7410762#.ZIHPC9BzfY>

Komljenović, V., Draščić, C, M., O'Neill, G., Karasz, I. (2022). Calculating National Financial Contributions to EOSC, <https://zenodo.org/record/7951324#.ZIHqC9BzfY>

Prevalence of Open Science support

History

Version	Revision date	Revision	Author
1.0	2023-07-12	Initial draft	T. van Leeuwen

Description

Open Science support is mostly organized at the institutional and even faculty level. This is often organized by the university or faculty libraries when it comes to scholarly publishing, that is, Open Access publishing. Here the support covers mostly things such as suggesting journals and helping with license details. When it relates to research data, support is nowadays organized in many institutions by data stewards. They take care of data management in general and consider issues related to for example FAIR principles.

The creation of metrics should take into consideration what elements of Open Science practices are being selected for support. Given this potential diversity, and the absence of having generic or systematic data sources at hand for this aspect of Open Science, the data need to be collected by qualitative means, that is going through websites of institutions and/or faculties to identify Open Science support facilities.

Metrics

Number/% of institutions that offer OS support

Open Science support might be often provided at the university level, or within the university at the faculty or even departmental level. This metric provides an institutional view of the extent to which Open Science support is offered.

MEASUREMENT.

This metric can be constructed by searching through university and faculty level websites to identify the presence of Open Science support facilities, and when present identify what kind of Open Science activity is being supported.

Potential problems and limitations are:

- This is a time-consuming effort, as one has to go through many websites, in particular when one is dealing with a large country, with many institutions, not necessarily limited to universities, but for example in the case of Germany, many publicly funded research institutions under the flag of Max Planck, Leibniz, Helmholtz, etc.
- This kind of information might not be present on the website (yet), and as such create an underestimation of the actual situation.
- How does one compare the breadth of the Open Science support, here one only aims at the presence, but the range of facilities to support Open Science is here not yet included.

There are presently no generic and systematic data sources that contain such information.

Automation of this metric is currently not possible.

Number/% of national OS support initiatives.

In some cases, there might be national Open Science initiatives or other national institutions that provide Open Science support. For instance, this could cover information around Open Access, or support for opening up data. This metric describes the extent to which there are national initiatives for Open Science support.

MEASUREMENT.

This metric can be constructed by searching through policy initiatives that support Open Science practices on the national level.

Potential problems and limitations are:

- This is a time-consuming effort, as one has to go through many websites, in particular with respect to the definition of 'national'. In some countries, governance of science is not only a national issue, but might in the case of a federal government, also have to deal with federal initiative regarding the support of Open Science practices.
- A similar issue relates to the presence of funders, and their support of Open Science practices, these are also considered national research funding agencies, and how do these relate to the abovementioned issue of the 'national' level.
- How does one compare the breadth of the Open Science support, here one only aims at the presence, but the range of facilities to support Open Science is here not yet included.

There are presently no generic and systematic data sources that contain such information.

Automation of this metric is currently not possible.

Breadth of Open Science support (all Open Science areas, only limited to specific areas).

Open Science support can be broader or narrower. This metric provides an overview of the breadth of Open Science support, that is, to which extent it covers the diverse types of Open Science support, such as publishing, research data, Open Code, peer review, pre-registration and registered reports.

MEASUREMENT.

Based on data that might be collected for the first two metrics, we might be able to construct a third metric. That is, when collecting information about training facilities, we could collect additional information. This metric could be constructed by searching through any relevant training facilities, either at the institutional or national level, as explained above.

Potential problems and limitations are:

- This is a time-consuming effort, as one has to go through many websites, on both overall institution level, as well as on faculty level.
- One has to decide upon a common denominator, the support for Open Science practices that are frequent across the system, but how does one deal with discipline specific support facilities that are not common?

There are presently no generic and systematic data sources that contain such information.

Automation is currently not possible.

Prevalence of preprinting

History

Version	Revision date	Revision	Author
1.1	2023-08-29	Draft for publication	I. Grypari
1.0	2023-05-09	First draft	N. Manola, I. Grypari, P. Stavropoulos, H. Papageorgiou

Description

Description:

- The “prevalence of preprinting” indicator measures the frequency of preprints that are made publicly available prior to peer review and publication in a journal. Preprinting allows researchers to share their work quickly and receive feedback from their peers before publication, and it also provides Open Access to research that might otherwise be locked behind journal paywalls.

Usefulness:

- The “prevalence of preprinting” indicator can help assess the impact of preprinting on scholarly communication and the dissemination of research findings.
- It provides insight into the growing trend of sharing research findings before they are formally published in a peer-reviewed journal.
- It can help measure the effectiveness of preprinting in speeding up the dissemination of research results.

Limitations:

- Not all fields or research areas may have a culture of preprinting, which can affect the applicability of the indicator in different contexts.
- As a standalone this indicator also does not account for other forms of informal communication, such as conference presentations, personal communications, or social media discussions, which can also play a role in the dissemination of research findings.

Metrics

Number of preprints

Description:

- The “number of preprints” metric measures the absolute quantity of preprints.

Usefulness:

- The “number of preprints” metric provides a quantitative measure of the prevalence of preprints in a particular field or research area or any other unit of analysis.

Limitations:

- The pre-prints that are early in the research cycle (early work in progress) may not be ready for dissemination and may not be useful in a repository (not written in a clear way, too much of a rough draft).
- View also the limitations in section I.

MEASUREMENT.

Count

DATASOURCES

OPENAIRE GRAPH

1. Identify the preprints in the OpenAIRE graph using the publication classification type “Preprint”
2. Identify any repositories that only take preprints (preprint servers) and use those publications.
3. In addition, find all the Green OA publications in OpenAIRE and
 1. From those, those that are published in Closed Access (but Green OA) they are preprints.
 2. Add to those, those that are not published yet, they are a preprint
 3. For the rest, if they are published check the date of deposition in repository, if it is earlier than the publication date, it is *most likely* a preprint and not the version of record. **Checking how far before publication it was deposited is a useful measure in itself.**
4. Count

Limitations

-

–

- Coverage of data for 1. (in other words, many preprints that are not described as preprints). If there are repositories that only accept preprints

- For 2. Different preprint servers may have different policies regarding the types of preprints they accept, the length of time they allow preprints to be available, and the types of metadata that are included with preprints. These variations can affect the completeness and accuracy of the data used to calculate the metric.
- For step 3.c. date of deposition is not a metadata element regularly exposed by repositories. This provides an additional problem as this metric is most meaningful if one can identify how long before publication something was deposited (if it is right before it shouldn't be counted for this indicators).

% of Publications that have preprints

Description:

- The “share of publications that have a preprint” metric measures the proportion of articles that are available as preprints.

Usefulness:

- The “share of publications that have a preprint” metric can help assess the adoption of preprinting as a form of scholarly communication in a particular field or research area, and using a share as opposed to a relative value is more useful for comparisons, e.g. over time.

Limitations:

- -
 - See limitations in previous metric.

MEASUREMENT.

$\% = 100 * (\text{pubs with preprint}) / (\text{total number of pubs})$

[DATASOURCES](#)

[OPENAIRE GRAPH](#)

[METHODOLOGY](#)

1. Take all the publications for unit of analysis
2. Identify those with a preprint available, from the set of preprints as in previous metric
3. Calculate share

Limitations

-
-
- Like previous metric

% of preprints that are published

Description:

-
-
- The “% of preprints that are published” metric measures the proportion of preprints that are eventually published in peer-reviewed journals.

Usefulness:

-
-
- The “% of preprints that are published” metric provides insight into the effectiveness of preprinting as a tool for disseminating research findings and as a precursor to formal publication in peer-reviewed journals.
- It can help researchers, institutions, and funding agencies to evaluate the impact of preprinting on the overall scholarly communication landscape.
- This metric can also be used to identify trends in preprint adoption and publishing across different fields or research areas.

Limitations:

-
-
- The “% of preprints that are published” metric has limitations because it does not account for the time lag between preprinting and formal publication in a peer-reviewed journal, which can vary widely depending on the field or research area.
- This metric also does not account for the quality or impact of the preprint, which can affect its likelihood of being published in a peer-reviewed journal.

MEASUREMENT.

$$\% = 100 * (\text{preprints that are published}) / (\text{total number of preprints})$$

DATASOURCES

OPENAIRE GRAPH

1. Identify preprints as in previous metrics
2. Find those that are published
3. Calculate share

Limitations

- - - Like previous metric

Known correlates

We are unsure whether these are known but we would assume the following are correlates:

- - - Field of research (preprint culture)
 - Time (adoption over time)
 - Funding source and Open Access policies of institutions(if it is a mandate)
 - Stage of career (preprint culture)
 - Collaboration networks, e.g. if large social media presence of network maybe likely to adopt preprinting practices.

Prevalence of replication studies

History

Version	Revision date	Revision	Author
1.1	2023-05-11	Revisions	P. Stavropoulos
1.0	2023-02-12	First draft	I. Grypari, N. Manola, H. Papageorgiou, P. Stavropoulos

Description

Reproducibility and replicability are closely related concepts in scientific research. Reproducibility refers to the ability of independent researchers to obtain consistent results when using the same data and methods. It involves analyzing existing data in a similar manner to validate the findings. On the other hand, replicability focuses on the ability to repeat the entire experiment or study using new data collection and similar methods employed in the original study. This includes collecting new data to ensure the accuracy and robustness of the findings.

Reproducibility serves as a necessary but not sufficient condition for replicability. While reproducibility emphasizes obtaining consistent results from the same data and methods, replicability extends this by aiming to achieve the same results from new data and methods. In essence, replicability confirms the generalizability and applicability of the original research findings.

Replication studies play a crucial role in addressing the concepts of reproducibility and replicability. They are research studies that attempt to validate the findings of a prior piece of research by repeating the study using similar methods and circumstances. They can be exact or conceptual and aim to confirm the accuracy and broad applicability of the original research. Replication studies are important because they increase confidence in the findings of the original research and provide opportunities to test existing theories, hypotheses, or models.

Prevalence of replication studies is an indicator of the extent to which replication studies are being conducted in a particular field or scientific community. It aims to capture the adoption of Open Science practices related to reproducibility and transparency. Replication studies are important for validating and building upon existing research findings, and promoting Open Science practices can increase the reliability and credibility of scientific research.

This indicator can be used to assess the level of adoption of Open Science practices related to replication in a particular field, and to identify potential barriers or incentives for the adoption of such practices. Additionally, the prevalence of replication studies can be used to assess the impact

of Open Science policies and initiatives aimed at promoting reproducibility and transparency in research.

Metrics

Number of replication studies

Number of replication studies is a metric that counts the number of studies that replicate previous research findings.

Limitations of this metric include the potential for biased or incomplete reporting of replication studies, as well as the possibility that some replication studies may not be identified or counted due to differences in methodology or interpretation.

MEASUREMENT.

To measure the number of replication studies, besides manually testing, a suitable automatic approach is to use text mining and machine learning techniques to find candidate studies that define themselves as replications or provide evidence of replication. Another approach is to search for a specific tag or field in relevant databases that indicate replication studies, such as the Open Science Framework's Registered Reports database (<https://osf.io/registries/discover?provider=OSF>) or the Replication Wiki (http://replication.uni-goettingen.de/wiki/index.php/Main_Page). An alternative method would be to conduct a citation analysis, where relevant citations from the original publications are collected, referenced within the original text, and assessed using an automated tool to determine if they can be classified as replication studies.

Potential measurement problems and limitations include the possibility of incomplete or inaccurate reporting of replication studies, as well as differences in definitions and methodologies for identifying replication studies. Additionally, not all replication studies may be published in a way that makes them easily identifiable, which could lead to undercounting. There may also be variations in the level of adoption of Open Science practices related to replication across different fields or scientific communities, which could make it challenging to compare the prevalence of replication studies across different domains.

DATASOURCES

OPEN SCIENCE FRAMEWORK (OSF) - REGISTERED REPORTS

The Open Science Framework (OSF) provides a platform for researchers to pre-register their studies, including replication studies, as Registered Reports. These reports undergo peer review prior to data collection, ensuring transparency and credibility. The OSF Registered Reports database can be accessed at <https://osf.io/registries/discover?provider=OSF>. The database can be queried using specific search terms to find replication

For the calculation of the metric, the database can be queried using specific search terms to find the relevant replication studies.

Limitations of this datasource include the reliance on researchers voluntarily registering their studies as Registered Reports, which may introduce selection bias. Additionally, not all replication studies may be registered or reported in this database, limiting its coverage.

REPLICATION WIKI

The Replication Wiki (http://replication.uni-goettingen.de/wiki/index.php/Main_Page) is an online resource that catalogs a curated list of replication studies.

However, it is important to note that the Replication Wiki may not cover all replication studies, and its coverage may vary across different fields or scientific domains.

EXISTING METHODOLOGY

To our knowledge there is no existing methodology to automatically find replication studies. Some suggested methods are the following:

1. Text mining and machine learning algorithm to find candidate studies that define themselves as replications or provide evidence of replication.
2. Search relevant fields or tags in specific databases to find publications that are replication studies.
3. Perform citation analysis, where relevant citations from the original publications are collected, referenced within the original text, and assessed using an automated tool to determine if they can be classified as replication studies.

Number (%) of publications that are replication studies

Metric that measures the percentage of publications within a specific field or scientific community that are replication studies. This metric provides a more nuanced perspective on the prevalence of replication studies than simply counting the number of replication studies, as it considers the total number of publications within a specific scientific field or task.

Limitations of this metric include the potential for biased or incomplete reporting of replication studies, as well as the possibility that some replication studies may be misclassified as non-replication studies due to differences in methodology or interpretation. Additionally, this metric may be influenced by factors such as the availability of funding for replication studies and the incentives for researchers to conduct replication studies.

Transformative publishing agreements

History

Version	Revision date	Revision	Author
1.1	2023-06-30	Added OpenAPC; minor revisions to rest	L. Stoy
1.0	2023-03-01	First draft	L. Stoy

Description

Transformative publishing agreements are a catch-all term for different types of Open Access contracts that aim to facilitate Open Access publishing at national level. Usually, a transformative agreement is concluded between a publisher and a consortium of research organisations and/or their libraries. Single research institutions may conclude transformative agreements, too. Information on transformative agreements is collected by the ESAC initiative (<https://esac-initiative.org/about/transformative-agreements/>).

The 'transformative' meaning stems from the objective of these type of agreements to steer the payment modalities for scholarly publishing from pay-to-read (subscriptions to electronic resources) to pay-to-publish or similar models. This includes so called offsetting agreements, read-and-publish agreements, and publish-and-read agreements. Authors affiliated to the institutions can then publish their articles in Open Access. The exact model and modalities vary with publisher and agreement.

The power of transformative agreements lies in 'flipping' the entire publishing output of a consortium/institution from closed to open. Especially when large publishers are concerned, this can have strong effects on the absolute and relative amount of Open Access articles published in a given country.

Transformative agreements also touch upon the financial flows taking place in the research system. This concerns e.g., the allocation of costs between institutions to participate in the transformative agreement, the allocation of costs within libraries (from collections to publishing) and the costs the individual authors may face eventually. This can have structuring effects on both the academic sector and the scholarly publishing industry, and in particular the financial flows between actors.

Metrics

Transformative agreements

Considering that transformative agreements usually take place between a publisher and a national consortium, a basic metric is the existence of an agreement which could be coded as a binary yes-no indicator for a country-publisher pair.

A challenge is the fact that many publishers exist, and each might be of different relevance for a given country. In other words, the impact of a transformative agreement is only directly related to the Open Access-status of publications with this publisher – and this part might be of little overall impact to a country's or institution's overall publishing output. Measuring this impact over time, e.g., affiliated authors choosing to publish through the transformative agreement, might be an interesting case to study as well.

MEASUREMENT.

The basic measurement is the existence of a national level agreement between a consortium and a publisher. This can be done using a binary variable (yes/no). However, more detailed information might be needed, depending on the type of analysis. Of major interest would likely be to capture the overall amount of Open Access publications, the share of OA publications published through this transformative agreement, and other information that yields insights into the impact of the agreement.

This information may include:

- The overall coverage of a country's publishing output by transformative agreements with various publishers
- The amount of transformative agreements concluded in a given country
- The start and end date of an agreement
- The type(s) of articles covered by the agreement(s)
- The size and membership of the consortium (i.e., not all consortia cover all research organisations from a country; some cover other institutions; some consortia function differently depending on the publisher)
- The relevance (market share) of the publisher for the consortium
- The cost per article
- Other contract elements (archiving, workflows)
- The functioning of the consortium (e.g., opt-in or opt-out)

These are measurements which are generally feasible but may require more in-depth study, for instance using bibliometric research, desk research into contracts and other methods and sources.

DATASOURCES

ESAC REGISTRY OF TRANSFORMATIVE AGREEMENTS

The ESAC Registry of Transformative Agreements (<https://esac-initiative.org/about/transformative-agreements/agreement-registry/>) is a community-organised database of transformative agreements.

As of writing, the database contains 662 ongoing or past agreements. These agreements can be done through national agreements or individual institutions.

The database is available for download and online use. Structural information for each entry includes:

- Publisher
- Country / countries
- Organisation (consortium/university)
- Annual publishing output
- Start date
- End date
- ID = unique identifier given by ESAC to each agreement
- URL = unique website with more detailed information

More detailed information for each transformative agreement includes:

- Impact of costs / costs development
- Financial shift
- Risk sharing
- OA coverage
- OA license
- Article types
- Access types
- Access costs
- Access coverage
- Perpetual access rights
- Workflow assessment
- Overall assessment and comments

Notably, the total cost of an agreement and per-article costs is not available through the ESAC Registry. This information is often publicly available but must be found individually for each agreement through the respective consortium or research conducted elsewhere.

ESAC MARKET WATCH

A sister service of the ESAC Registry, the ESAC Market Watch (<https://esac-initiative.org/market-watch/>) provides information about the market share of publishers and the role of transformative agreements.

Information available online entails

- Global and national market share of publishers

- Amount of articles published through transformative agreements (cumulative/by year)
- The number of transformative agreements per country and per publisher
- A country overview of publishing outputs divided by transformative deals, other OA, and hybrid/closed
- Overviews over publisher journal portfolios
- Article processing charges
- Market positions

It should be noted that the data is not fully available for download.

The limitations disclaimer should also be consulted before using the ESAC Market Watch information.

OPENAPC

The OpenAPC initiative provides some statistical information about transformative agreements. Data is available for the number of publications published through a transformative agreement for 12 countries. The data can be filtered by publisher, institution, hybrid status, country, journal, year, and agreement.

Visualisation: <https://treemaps.openapc.net/apcdata/transformative-agreements/#agreement/>

Data is available via the GitHub repository, which includes a more detailed description of the dataset: https://github.com/OpenAPC/openapc-de/tree/master/data/transformative_agreements

Specific further analyses are provided for:

- The German DEAL agreements: <https://treemaps.openapc.net/apcdata/deal/>
- A combination of Transformative Agreement and APC data: <https://treemaps.openapc.net/apcdata/combined/>

In all cases, the data is not necessarily complete and should be used with care.

References

Hinchliffe, Lisa Janicke. 'Read-and-Publish? Publish-and-Read? A Primer on Transformative Agreements by @lisalibrarian.' The Scholarly Kitchen, 23 April 2019. <https://scholarlykitchen.sspnet.org/2019/04/23/transformative-agreements/>.

Part II: Academic Impact

Academic readership

History

Description

[Add text here]

Metrics

Number of times publications are saved in reference manager (Avg. / Total).

[Add text here]

MEASUREMENT.

[Add text here]

EXISTING DATASOURCES:

<DATASOURCE NAME>

[Add text here]

EXISTING METHODOLOGIES

<METHODOLOGY NAME>

[Add text here]

Number of times publications are accessed (views, clicks, downloads)

[Add text here]

MEASUREMENT.

[Add text here]

EXISTING DATASOURCES:

<DATASOURCE NAME>

[Add text here]

EXISTING METHODOLOGIES

<METHODOLOGY NAME>

[Add text here]

Known correlates

[Add text here]

Notes

[Add text here]

References

[Add Zotero bibliography here]

Citation impact

History

Description

[Add text here]

Metrics

Avg. Citations (MCS)

[Add text here]

MEASUREMENT.

[Add text here]

EXISTING DATASOURCES:

<DATASOURCE NAME>

[Add text here]

EXISTING METHODOLOGIES

<METHODOLOGY NAME>

[Add text here]

Total Citations (TCS)

[Add text here]

MEASUREMENT.

[Add text here]

EXISTING DATASOURCES:

<DATASOURCE NAME>

[Add text here]

EXISTING METHODOLOGIES

<METHODOLOGY NAME>

[Add text here]

Highly-cited publications (Number / %)

[Add text here]

MEASUREMENT.

[Add text here]

EXISTING DATASOURCES:

<DATASOURCE NAME>

[Add text here]

EXISTING METHODOLOGIES

<METHODOLOGY NAME>

[Add text here]

Avg. Normalised Citations (MNCS)

[Add text here]

MEASUREMENT.

[Add text here]

EXISTING DATASOURCES:

<DATASOURCE NAME>

[Add text here]

EXISTING METHODOLOGIES

<METHODOLOGY NAME>

[Add text here]

Total Normalised Citations (TNCS)

[Add text here]

MEASUREMENT.

[Add text here]

EXISTING DATASOURCES:

<DATASOURCE NAME>

[Add text here]

EXISTING METHODOLOGIES

<METHODOLOGY NAME>

[Add text here]

Highly-(normalised)-cited publications (Number / %)

[Add text here]

MEASUREMENT.

[Add text here]

EXISTING DATASOURCES:

<DATASOURCE NAME>

[Add text here]

EXISTING METHODOLOGIES

<METHODOLOGY NAME>

[Add text here]

Known correlates

[Add text here]

Notes

[Add text here]

References

[Add Zotero bibliography here]

Collaboration intensity

History

Description

[Add text here]

Metrics

**Avg. number of co-authors
(author/institution/country)**

[Add text here]

MEASUREMENT.

[Add text here]

EXISTING DATASOURCES:

<DATASOURCE NAME>

[Add text here]

EXISTING METHODOLOGIES

<METHODOLOGY NAME>

[Add text here]

**% of co-authored publications
(author/institution/country)**

[Add text here]

MEASUREMENT.

[Add text here]

EXISTING DATASOURCES:

<DATASOURCE NAME>

[Add text here]

EXISTING METHODOLOGIES

<METHODOLOGY NAME>

[Add text here]

Known correlates

[Add text here]

Notes

[Add text here]

References

[Add Zotero bibliography here]

History

Description

[Add text here]

Metrics

Avg. number of different institutions / countries.

[Add text here]

MEASUREMENT.

[Add text here]

EXISTING DATASOURCES:

<DATASOURCE NAME>

[Add text here]

EXISTING METHODOLOGIES

<METHODOLOGY NAME>

[Add text here]

Diversity of institutions (in terms of ranking).

[Add text here]

MEASUREMENT.

[Add text here]

EXISTING DATASOURCES:

<DATASOURCE NAME>

[Add text here]

EXISTING METHODOLOGIES

<METHODOLOGY NAME>

[Add text here]

Avg. number of different disciplines.

[Add text here]

MEASUREMENT.

[Add text here]

EXISTING DATASOURCES:

<DATASOURCE NAME>

[Add text here]

EXISTING METHODOLOGIES

<METHODOLOGY NAME>

[Add text here]

Avg. number of different genders.

[Add text here]

MEASUREMENT.

[Add text here]

EXISTING DATASOURCES:

<DATASOURCE NAME>

[Add text here]

EXISTING METHODOLOGIES

<METHODOLOGY NAME>

[Add text here]

Avg. number of different ethnicities.

[Add text here]

MEASUREMENT.

[Add text here]

EXISTING DATASOURCES:

<DATASOURCE NAME>

[Add text here]

EXISTING METHODOLOGIES

<METHODOLOGY NAME>

[Add text here]

Known correlates

[Add text here]

Notes

[Add text here]

References

[Add Zotero bibliography here]

Industry collaboration

History

Description

[Add text here]

Metrics

% of publications with industry co-authors

[Add text here]

MEASUREMENT.

[Add text here]

EXISTING DATASOURCES:

<DATASOURCE NAME>

[Add text here]

EXISTING METHODOLOGIES

<METHODOLOGY NAME>

[Add text here]

% of co-funded grants with industry (e.g. Horizon2020)

[Add text here]

MEASUREMENT.

[Add text here]

EXISTING DATASOURCES:

<DATASOURCE NAME>

[Add text here]

EXISTING METHODOLOGIES

<METHODOLOGY NAME>

[Add text here]

Known correlates

[Add text here]

Notes

[Add text here]

References

[Add Zotero bibliography here]

Interdisciplinarity

History

Description

[Add text here]

Metrics

Number of scientific (sub)fields associated to publications

[Add text here]

MEASUREMENT.

[Add text here]

EXISTING DATASOURCES:

<DATASOURCE NAME>

[Add text here]

EXISTING METHODOLOGIES

<METHODOLOGY NAME>

[Add text here]

Relative scientific distance of authors/cited articles

[Add text here]

MEASUREMENT.

[Add text here]

EXISTING DATASOURCES:

<DATASOURCE NAME>

[Add text here]

EXISTING METHODOLOGIES

<METHODOLOGY NAME>

[Add text here]

Known correlates

[Add text here]

Notes

[Add text here]

References

[Add Zotero bibliography here]

Novelty

History

Description

[Add text here]

Metrics

Relative scientific distance between citing/cited articles

[Add text here]

MEASUREMENT.

[Add text here]

EXISTING DATASOURCES:

<DATASOURCE NAME>

[Add text here]

EXISTING METHODOLOGIES

<METHODOLOGY NAME>

[Add text here]

Relative scientific distance between keywords

[Add text here]

MEASUREMENT.

[Add text here]

EXISTING DATASOURCES:

<DATASOURCE NAME>

[Add text here]

EXISTING METHODOLOGIES

<METHODOLOGY NAME>

[Add text here]

Known correlates

[Add text here]

Notes

[Add text here]

References

[Add Zotero bibliography here]

Societal collaboration

History

Description

[Add text here]

Metrics

% of publications with societal organisations (e.g. NGO, non-profit) as co-authors

[Add text here]

MEASUREMENT.

[Add text here]

EXISTING DATASOURCES:

<DATASOURCE NAME>

[Add text here]

EXISTING METHODOLOGIES

<METHODOLOGY NAME>

[Add text here]

Known correlates

[Add text here]

Notes

[Add text here]

References

[Add Zotero bibliography here]

Topic trends

History

Description

[Add text here]

Metrics

Share of output in emerging topics

[Add text here]

MEASUREMENT.

[Add text here]

EXISTING DATASOURCES:

<DATASOURCE NAME>

[Add text here]

EXISTING METHODOLOGIES

<METHODOLOGY NAME>

[Add text here]

Known correlates

[Add text here]

Notes

[Add text here]

References

[Add Zotero bibliography here]

Use of code in research

History

Description

[Add text here]

Metrics

Number (Avg.) of times code is cited/mentions in publications

[Add text here]

MEASUREMENT.

[Add text here]

EXISTING DATASOURCES:

<DATASOURCE NAME>

[Add text here]

EXISTING METHODOLOGIES

<METHODOLOGY NAME>

[Add text here]

Number (Avg.) of times code is used in replication material

[Add text here]

MEASUREMENT.

[Add text here]

EXISTING DATASOURCES:

<DATASOURCE NAME>

[Add text here]

EXISTING METHODOLOGIES

<METHODOLOGY NAME>

[Add text here]

GitHub statistics (# Forks/Stars).

[Add text here]

MEASUREMENT.

[Add text here]

EXISTING DATASOURCES:

<DATASOURCE NAME>

[Add text here]

EXISTING METHODOLOGIES

<METHODOLOGY NAME>

[Add text here]

Software dependencies.

[Add text here]

MEASUREMENT.

[Add text here]

EXISTING DATASOURCES:

<DATASOURCE NAME>

[Add text here]

EXISTING METHODOLOGIES

<METHODOLOGY NAME>

[Add text here]

Number of views/clicks/downloads from repository

[Add text here]

MEASUREMENT.

[Add text here]

EXISTING DATASOURCES:

<DATASOURCE NAME>

[Add text here]

EXISTING METHODOLOGIES

<METHODOLOGY NAME>

[Add text here]

Known correlates

[Add text here]

Notes

[Add text here]

References

[Add Zotero bibliography here]

Use of data in research

History

Description

[Add text here]

Metrics

Number (Avg.) of times data is cited in publications

[Add text here]

MEASUREMENT.

[Add text here]

EXISTING DATASOURCES:

<DATASOURCE NAME>

[Add text here]

EXISTING METHODOLOGIES

<METHODOLOGY NAME>

[Add text here]

Number (Avg.) of times data is mentioned in publications

[Add text here]

MEASUREMENT.

[Add text here]

EXISTING DATASOURCES:

<DATASOURCE NAME>

[Add text here]

EXISTING METHODOLOGIES

<METHODOLOGY NAME>

[Add text here]

Number (Avg.) of views/clicks/downloads from repository

[Add text here]

MEASUREMENT.

[Add text here]

EXISTING DATASOURCES:

<DATASOURCE NAME>

[Add text here]

EXISTING METHODOLOGIES

<METHODOLOGY NAME>

[Add text here]

Known correlates

[Add text here]

Notes

[Add text here]

References

[Add Zotero bibliography here]

Use of methods in research

History

Description

[Add text here]

Metrics

Number of citations of methodological publications

[Add text here]

MEASUREMENT.

[Add text here]

EXISTING DATASOURCES:

<DATASOURCE NAME>

[Add text here]

EXISTING METHODOLOGIES

<METHODOLOGY NAME>

[Add text here]

Number of occurrences of methodological keywords in publications

[Add text here]

MEASUREMENT.

[Add text here]

EXISTING DATASOURCES:

<DATASOURCE NAME>

[Add text here]

EXISTING METHODOLOGIES

<METHODOLOGY NAME>

[Add text here]

Known correlates

[Add text here]

Notes

[Add text here]

References

[Add Zotero bibliography here]

Use of patents in research

History

Description

[Add text here]

Metrics

Number (Avg.) of times patents are referenced in publications

[Add text here]

MEASUREMENT.

[Add text here]

EXISTING DATASOURCES:

<DATASOURCE NAME>

[Add text here]

EXISTING METHODOLOGIES

<METHODOLOGY NAME>

[Add text here]

Known correlates

[Add text here]

Notes

[Add text here]

References

[Add Zotero bibliography here]

Part III: Societal Impact

Uptake in citizen science

History

Description

[Add text here]

Metrics

#/% of non-experts mobilized/authoring in research outputs

[Add text here]

MEASUREMENT.

[Add text here]

EXISTING DATASOURCES:

<DATASOURCE NAME>

[Add text here]

EXISTING METHODOLOGIES

<METHODOLOGY NAME>

[Add text here]

#/% of "citizen science" projects per country/discipline/institutes/type of research

[Add text here]

MEASUREMENT.

[Add text here]

EXISTING DATASOURCES:

<DATASOURCE NAME>

[Add text here]

EXISTING METHODOLOGIES

<METHODOLOGY NAME>

[Add text here]

#/% of research outputs mentioned in “citizen science” projects

[Add text here]

MEASUREMENT.

[Add text here]

EXISTING DATASOURCES:

<DATASOURCE NAME>

[Add text here]

EXISTING METHODOLOGIES

<METHODOLOGY NAME>

[Add text here]

#/% of research outputs mentioned in Wikipedia

[Add text here]

MEASUREMENT.

[Add text here]

EXISTING DATASOURCES:

<DATASOURCE NAME>

[Add text here]

EXISTING METHODOLOGIES

<METHODOLOGY NAME>

[Add text here]

#/% of research outputs mentioned by skeptics websites

[Add text here]

MEASUREMENT.

[Add text here]

EXISTING DATASOURCES:

<DATASOURCE NAME>

[Add text here]

EXISTING METHODOLOGIES

<METHODOLOGY NAME>

[Add text here]

#/% of research outputs mentioned by fact-checking websites

[Add text here]

MEASUREMENT.

[Add text here]

EXISTING DATASOURCES:

<DATASOURCE NAME>

[Add text here]

EXISTING METHODOLOGIES

<METHODOLOGY NAME>

[Add text here]

Known correlates

[Add text here]

Notes

[Add text here]

References

[Add Zotero bibliography here]

Uptake in education

History

Description

[Add text here]

Metrics

#/% of research outputs in educational handbooks

[Add text here]

MEASUREMENT.

[Add text here]

EXISTING DATASOURCES:

<DATASOURCE NAME>

[Add text here]

EXISTING METHODOLOGIES

<METHODOLOGY NAME>

[Add text here]

#/% of research outputs in higher education syllabus

[Add text here]

MEASUREMENT.

[Add text here]

EXISTING DATASOURCES:

<DATASOURCE NAME>

[Add text here]

EXISTING METHODOLOGIES

<METHODOLOGY NAME>

[Add text here]

Known correlates

[Add text here]

Notes

[Add text here]

References

[Add Zotero bibliography here]

Science literacy

History

Description

[Add text here]

Metrics

#/% of references to research outputs in science popularization websites (e.g. Wikipedia)

[Add text here]

MEASUREMENT.

[Add text here]

EXISTING DATASOURCES:

<DATASOURCE NAME>

[Add text here]

EXISTING METHODOLOGIES

<METHODOLOGY NAME>

[Add text here]

#/% citizens aware of the existence/principle of open science

[Add text here]

MEASUREMENT.

[Add text here]

EXISTING DATASOURCES:

<DATASOURCE NAME>

[Add text here]

EXISTING METHODOLOGIES

<METHODOLOGY NAME>

[Add text here]

Known correlates

[Add text here]

Notes

[Add text here]

References

[Add Zotero bibliography here]

Uptake by policy makers

History

Description

[Add text here]

Metrics

of the mentions of the idea of Open Science in policy documents

[Add text here]

MEASUREMENT.

[Add text here]

EXISTING DATASOURCES:

<DATASOURCE NAME>

[Add text here]

EXISTING METHODOLOGIES

<METHODOLOGY NAME>

[Add text here]

#/% of references to research outputs in policy documents

[Add text here]

MEASUREMENT.

[Add text here]

EXISTING DATASOURCES:

<DATASOURCE NAME>

[Add text here]

EXISTING METHODOLOGIES

<METHODOLOGY NAME>

[Add text here]

#/% of references to research outputs in official institutions reports/websites

[Add text here]

MEASUREMENT.

[Add text here]

EXISTING DATASOURCES:

<DATASOURCE NAME>

[Add text here]

EXISTING METHODOLOGIES

<METHODOLOGY NAME>

[Add text here]

Known correlates

[Add text here]

Notes

[Add text here]

References

[Add Zotero bibliography here]

Uptake in the legal sector

History

Description

[Add text here]

Metrics

#/% of references to research outputs in legal publications

[Add text here]

MEASUREMENT.

[Add text here]

EXISTING DATASOURCES:

<DATASOURCE NAME>

[Add text here]

EXISTING METHODOLOGIES

<METHODOLOGY NAME>

[Add text here]

#/% of open data based legal projects / bills

[Add text here]

MEASUREMENT.

[Add text here]

EXISTING DATASOURCES:

<DATASOURCE NAME>

[Add text here]

EXISTING METHODOLOGIES

<METHODOLOGY NAME>

[Add text here]

Known correlates

[Add text here]

Notes

[Add text here]

References

[Add Zotero bibliography here]

Uptake in the public debate

History

Description

[Add text here]

Metrics

#/% of references to OS in social media

[Add text here]

MEASUREMENT.

[Add text here]

EXISTING DATASOURCES:

<DATASOURCE NAME>

[Add text here]

EXISTING METHODOLOGIES

<METHODOLOGY NAME>

[Add text here]

#/% of research outputs mentioned in Wikipedia/use of OS datasets/software

[Add text here]

MEASUREMENT.

[Add text here]

EXISTING DATASOURCES:

<DATASOURCE NAME>

[Add text here]

EXISTING METHODOLOGIES

<METHODOLOGY NAME>

[Add text here]

#/% of references to OS in the press / use of OS datasets/software

[Add text here]

MEASUREMENT.

[Add text here]

EXISTING DATASOURCES:

<DATASOURCE NAME>

[Add text here]

EXISTING METHODOLOGIES

<METHODOLOGY NAME>

[Add text here]

#/% of research outputs mentioning societal issues

[Add text here]

MEASUREMENT.

[Add text here]

EXISTING DATASOURCES:

<DATASOURCE NAME>

[Add text here]

EXISTING METHODOLOGIES

<METHODOLOGY NAME>

[Add text here]

Known correlates

[Add text here]

Notes

[Add text here]

References

[Add Zotero bibliography here]

Uptake in education

History

Description

[Add text here]

Metrics

#/% of references to research outputs in medical guidelines

[Add text here]

MEASUREMENT.

[Add text here]

EXISTING DATASOURCES:

<DATASOURCE NAME>

[Add text here]

EXISTING METHODOLOGIES

<METHODOLOGY NAME>

[Add text here]

of patients having shared their data

[Add text here]

MEASUREMENT.

[Add text here]

EXISTING DATASOURCES:

<DATASOURCE NAME>

[Add text here]

EXISTING METHODOLOGIES

<METHODOLOGY NAME>

[Add text here]

Known correlates

[Add text here]

Notes

[Add text here]

References

[Add Zotero bibliography here]

Uptake by patient groups

History

Description

[Add text here]

Metrics

#/% of references to OS publications / use of OS datasets/software in patient group websites / reports

[Add text here]

MEASUREMENT.

[Add text here]

EXISTING DATASOURCES:

<DATASOURCE NAME>

[Add text here]

EXISTING METHODOLOGIES

<METHODOLOGY NAME>

[Add text here]

Known correlates

[Add text here]

Notes

[Add text here]

References

[Add Zotero bibliography here]

Uptake by civil society

History

Description

[Add text here]

Metrics

#/% of references to research outputs in activist groups / NGOs reports/websites

[Add text here]

MEASUREMENT.

[Add text here]

EXISTING DATASOURCES:

<DATASOURCE NAME>

[Add text here]

EXISTING METHODOLOGIES

<METHODOLOGY NAME>

[Add text here]

Known correlates

[Add text here]

Notes

[Add text here]

References

[Add Zotero bibliography here]

Uptake by the general public

History

Description

[Add text here]

Metrics

#/% of references to OS publications/datasets online in personal blogs & websites

[Add text here]

MEASUREMENT.

[Add text here]

EXISTING DATASOURCES:

<DATASOURCE NAME>

[Add text here]

EXISTING METHODOLOGIES

<METHODOLOGY NAME>

[Add text here]

#/% of references to OS publications/datasets online in personal social media accounts

[Add text here]

MEASUREMENT.

[Add text here]

EXISTING DATASOURCES:

<DATASOURCE NAME>

[Add text here]

EXISTING METHODOLOGIES

<METHODOLOGY NAME>

[Add text here]

Known correlates

[Add text here]

Notes

[Add text here]

References

[Add Zotero bibliography here]

Uptake by policy makers

History

Description

[Add text here]

Metrics

#/% of artistic institutions websites references to research outputs

[Add text here]

MEASUREMENT.

[Add text here]

EXISTING DATASOURCES:

<DATASOURCE NAME>

[Add text here]

EXISTING METHODOLOGIES

<METHODOLOGY NAME>

[Add text here]

#/% of artists/artistic projects working with research outputs

[Add text here]

MEASUREMENT.

[Add text here]

EXISTING DATASOURCES:

<DATASOURCE NAME>

[Add text here]

EXISTING METHODOLOGIES

<METHODOLOGY NAME>

[Add text here]

Known correlates

[Add text here]

Notes

[Add text here]

References

[Add Zotero bibliography here]

Effect on democracy

History

Description

[Add text here]

Metrics

#/% of references to research outputs in activist groups/NGOs reports/websites focussing on democracy

[Add text here]

MEASUREMENT.

[Add text here]

EXISTING DATASOURCES:

<DATASOURCE NAME>

[Add text here]

EXISTING METHODOLOGIES

<METHODOLOGY NAME>

[Add text here]

research outputs addressing democracy

[Add text here]

MEASUREMENT.

[Add text here]

EXISTING DATASOURCES:

<DATASOURCE NAME>

[Add text here]

EXISTING METHODOLOGIES

<METHODOLOGY NAME>

[Add text here]

General indicator on democracy

[Add text here]

MEASUREMENT.

[Add text here]

EXISTING DATASOURCES:

<DATASOURCE NAME>

[Add text here]

EXISTING METHODOLOGIES

<METHODOLOGY NAME>

[Add text here]

Known correlates

[Add text here]

Notes

[Add text here]

References

[Add Zotero bibliography here]

Effect on ethnic inequality

History

Description

[Add text here]

Metrics

#/% of references to research output in activist groups/NGOs reports/websites focussing on ethnic inequality

[Add text here]

MEASUREMENT.

[Add text here]

EXISTING DATASOURCES:

<DATASOURCE NAME>

[Add text here]

EXISTING METHODOLOGIES

<METHODOLOGY NAME>

[Add text here]

research outputs addressing ethnic inequality

[Add text here]

MEASUREMENT.

[Add text here]

EXISTING DATASOURCES:

<DATASOURCE NAME>

[Add text here]

EXISTING METHODOLOGIES

<METHODOLOGY NAME>

[Add text here]

General indicator on ethic inequality

[Add text here]

MEASUREMENT.

[Add text here]

EXISTING DATASOURCES:

<DATASOURCE NAME>

[Add text here]

EXISTING METHODOLOGIES

<METHODOLOGY NAME>

[Add text here]

Known correlates

[Add text here]

Notes

[Add text here]

References

[Add Zotero bibliography here]

Effect on gender inequality

History

Description

[Add text here]

Metrics

#/% of references to research outputs in activist groups/NGOs reports/websites focussing on gender inequality

[Add text here]

MEASUREMENT.

[Add text here]

EXISTING DATASOURCES:

<DATASOURCE NAME>

[Add text here]

EXISTING METHODOLOGIES

<METHODOLOGY NAME>

[Add text here]

research outputs addressing gender inequality

[Add text here]

MEASUREMENT.

[Add text here]

EXISTING DATASOURCES:

<DATASOURCE NAME>

[Add text here]

EXISTING METHODOLOGIES

<METHODOLOGY NAME>

[Add text here]

General indicator on gender inequality

[Add text here]

MEASUREMENT.

[Add text here]

EXISTING DATASOURCES:

<DATASOURCE NAME>

[Add text here]

EXISTING METHODOLOGIES

<METHODOLOGY NAME>

[Add text here]

Known correlates

[Add text here]

Notes

[Add text here]

References

[Add Zotero bibliography here]

Effect on SDGs

History

Description

[Add text here]

Metrics

#/% of references to research outputs in activist groups/NGOs reports/websites on SDGs

[Add text here]

MEASUREMENT.

[Add text here]

EXISTING DATASOURCES:

<DATASOURCE NAME>

[Add text here]

EXISTING METHODOLOGIES

<METHODOLOGY NAME>

[Add text here]

research outputs addressing SDGs

[Add text here]

MEASUREMENT.

[Add text here]

EXISTING DATASOURCES:

<DATASOURCE NAME>

[Add text here]

EXISTING METHODOLOGIES

<METHODOLOGY NAME>

[Add text here]

General indicators on SDGs

[Add text here]

MEASUREMENT.

[Add text here]

EXISTING DATASOURCES:

<DATASOURCE NAME>

[Add text here]

EXISTING METHODOLOGIES

<METHODOLOGY NAME>

[Add text here]

Known correlates

[Add text here]

Notes

[Add text here]

References

[Add Zotero bibliography here]

Part IV: Economic Impact

Science-industry collaboration

History

Description

[Add text here]

Metrics

#/% of collaborative projects using OS inputs

[Add text here]

MEASUREMENT.

[Add text here]

EXISTING DATASOURCES:

<DATASOURCE NAME>

[Add text here]

EXISTING METHODOLOGIES

<METHODOLOGY NAME>

[Add text here]

EUR/% of collaborative projects using OS inputs

[Add text here]

MEASUREMENT.

[Add text here]

EXISTING DATASOURCES:

<DATASOURCE NAME>

[Add text here]

EXISTING METHODOLOGIES

<METHODOLOGY NAME>

[Add text here]

#/% of patents filed by industry in collaboration with academia that cites OS inputs

[Add text here]

MEASUREMENT.

[Add text here]

EXISTING DATASOURCES:

<DATASOURCE NAME>

[Add text here]

EXISTING METHODOLOGIES

<METHODOLOGY NAME>

[Add text here]

#/% of publications produced by academia in collaboration with industry that cites OS inputs

[Add text here]

MEASUREMENT.

[Add text here]

EXISTING DATASOURCES:

<DATASOURCE NAME>

[Add text here]

EXISTING METHODOLOGIES

<METHODOLOGY NAME>

[Add text here]

Known correlates

[Add text here]

Notes

[Add text here]

References

[Add Zotero bibliography here]

Innovation output

History

Description

[Add text here]

Metrics

#/% of new products or services developed fully using OS inputs

[Add text here]

MEASUREMENT.

[Add text here]

EXISTING DATASOURCES:

<DATASOURCE NAME>

[Add text here]

EXISTING METHODOLOGIES

<METHODOLOGY NAME>

[Add text here]

#/% of new products or services developed partly using OS inputs

[Add text here]

MEASUREMENT.

[Add text here]

EXISTING DATASOURCES:

<DATASOURCE NAME>

[Add text here]

EXISTING METHODOLOGIES

<METHODOLOGY NAME>

[Add text here]

#/% of new technologies developed fully using OS inputs

[Add text here]

MEASUREMENT.

[Add text here]

EXISTING DATASOURCES:

<DATASOURCE NAME>

[Add text here]

EXISTING METHODOLOGIES

<METHODOLOGY NAME>

[Add text here]

#/% of new technologies developed partly using OS inputs

[Add text here]

MEASUREMENT.

[Add text here]

EXISTING DATASOURCES:

<DATASOURCE NAME>

[Add text here]

EXISTING METHODOLOGIES

<METHODOLOGY NAME>

[Add text here]

#/% of patents filed citing OS inputs

[Add text here]

MEASUREMENT.

[Add text here]

EXISTING DATASOURCES:

<DATASOURCE NAME>

[Add text here]

EXISTING METHODOLOGIES

<METHODOLOGY NAME>

[Add text here]

Known correlates

[Add text here]

Notes

[Add text here]

References

[Add Zotero bibliography here]

Uptake of research result by industry

History

Description

[Add text here]

Metrics

#/% of patents filed by the industry citing OS inputs

[Add text here]

MEASUREMENT.

[Add text here]

EXISTING DATASOURCES:

<DATASOURCE NAME>

[Add text here]

EXISTING METHODOLOGIES

<METHODOLOGY NAME>

[Add text here]

#/% of new technologies developed by the industry using OS inputs

[Add text here]

MEASUREMENT.

[Add text here]

EXISTING DATASOURCES:

<DATASOURCE NAME>

[Add text here]

EXISTING METHODOLOGIES

<METHODOLOGY NAME>

[Add text here]

#/% of new products or services developed by the industry using OS inputs

[Add text here]

MEASUREMENT.

[Add text here]

EXISTING DATASOURCES:

<DATASOURCE NAME>

[Add text here]

EXISTING METHODOLOGIES

<METHODOLOGY NAME>

[Add text here]

Average increase in companies' patent portfolio value thanks to patent filed using OS inputs

[Add text here]

MEASUREMENT.

[Add text here]

EXISTING DATASOURCES:

<DATASOURCE NAME>

[Add text here]

EXISTING METHODOLOGIES

<METHODOLOGY NAME>

[Add text here]

Known correlates

[Add text here]

Notes

[Add text here]

References

[Add Zotero bibliography here]

Socially relevant products and processes

History

Description

[Add text here]

Metrics

#/% of new medical treatments developed using OS inputs

[Add text here]

MEASUREMENT.

[Add text here]

EXISTING DATASOURCES:

<DATASOURCE NAME>

[Add text here]

EXISTING METHODOLOGIES

<METHODOLOGY NAME>

[Add text here]

#/% of new drugs developed using OS inputs

[Add text here]

MEASUREMENT.

[Add text here]

EXISTING DATASOURCES:

<DATASOURCE NAME>

[Add text here]

EXISTING METHODOLOGIES

<METHODOLOGY NAME>

[Add text here]

#/% of sustainable agriculture practices using OS inputs

[Add text here]

MEASUREMENT.

[Add text here]

EXISTING DATASOURCES:

<DATASOURCE NAME>

[Add text here]

EXISTING METHODOLOGIES

<METHODOLOGY NAME>

[Add text here]

#/% of new renewable energy technologies developed using OS inputs

[Add text here]

MEASUREMENT.

[Add text here]

EXISTING DATASOURCES:

<DATASOURCE NAME>

[Add text here]

EXISTING METHODOLOGIES

<METHODOLOGY NAME>

[Add text here]

Known correlates

[Add text here]

Notes

[Add text here]

References

[Add Zotero bibliography here]

Economic growth of companies

History

Description

[Add text here]

Metrics

Difference (+/-) in turnover due to OS

[Add text here]

MEASUREMENT.

[Add text here]

EXISTING DATASOURCES:

<DATASOURCE NAME>

[Add text here]

EXISTING METHODOLOGIES

<METHODOLOGY NAME>

[Add text here]

Difference (+/-) in profit due to OS

[Add text here]

MEASUREMENT.

[Add text here]

EXISTING DATASOURCES:

<DATASOURCE NAME>

[Add text here]

EXISTING METHODOLOGIES

<METHODOLOGY NAME>

[Add text here]

Difference (+/-) in intangible assets due to OS

[Add text here]

MEASUREMENT.

[Add text here]

EXISTING DATASOURCES:

<DATASOURCE NAME>

[Add text here]

EXISTING METHODOLOGIES

<METHODOLOGY NAME>

[Add text here]

Difference (+/-) in tangible assets due to OS

[Add text here]

MEASUREMENT.

[Add text here]

EXISTING DATASOURCES:

<DATASOURCE NAME>

[Add text here]

EXISTING METHODOLOGIES

<METHODOLOGY NAME>

[Add text here]

Difference (+/-) in CAPEX due to OS

[Add text here]

MEASUREMENT.

[Add text here]

EXISTING DATASOURCES:

<DATASOURCE NAME>

[Add text here]

EXISTING METHODOLOGIES

<METHODOLOGY NAME>

[Add text here]

Difference (+/-) in ROA due to OS

[Add text here]

MEASUREMENT.

[Add text here]

EXISTING DATASOURCES:

<DATASOURCE NAME>

[Add text here]

EXISTING METHODOLOGIES

<METHODOLOGY NAME>

[Add text here]

Difference (+/-) in ROE due to OS

[Add text here]

MEASUREMENT.

[Add text here]

EXISTING DATASOURCES:

<DATASOURCE NAME>

[Add text here]

EXISTING METHODOLOGIES

<METHODOLOGY NAME>

[Add text here]

Difference (+/-) in OPEX due to OS

[Add text here]

MEASUREMENT.

[Add text here]

EXISTING DATASOURCES:

<DATASOURCE NAME>

[Add text here]

EXISTING METHODOLOGIES

<METHODOLOGY NAME>

[Add text here]

Difference (+/-) in productivity due to OS

[Add text here]

MEASUREMENT.

[Add text here]

EXISTING DATASOURCES:

<DATASOURCE NAME>

[Add text here]

EXISTING METHODOLOGIES

<METHODOLOGY NAME>

[Add text here]

Difference (+/-) in number of employees due to OS

[Add text here]

MEASUREMENT.

[Add text here]

EXISTING DATASOURCES:

<DATASOURCE NAME>

[Add text here]

EXISTING METHODOLOGIES

<METHODOLOGY NAME>

[Add text here]

Difference (+/-) in cost of personnel due to OS

[Add text here]

MEASUREMENT.

[Add text here]

EXISTING DATASOURCES:

<DATASOURCE NAME>

[Add text here]

EXISTING METHODOLOGIES

<METHODOLOGY NAME>

[Add text here]

Known correlates

[Add text here]

Notes

[Add text here]

References

[Add Zotero bibliography here]

Labour market impacts

History

Description

[Add text here]

Metrics

Number of new job positions in the industry thanks to OS (e.g., FAIR expert)

[Add text here]

MEASUREMENT.

[Add text here]

EXISTING DATASOURCES:

<DATASOURCE NAME>

[Add text here]

EXISTING METHODOLOGIES

<METHODOLOGY NAME>

[Add text here]

Number of new job positions in public research thanks to OS

[Add text here]

MEASUREMENT.

[Add text here]

EXISTING DATASOURCES:

<DATASOURCE NAME>

[Add text here]

EXISTING METHODOLOGIES

<METHODOLOGY NAME>

[Add text here]

Number of new job positions in academia thanks to OS

[Add text here]

MEASUREMENT.

[Add text here]

EXISTING DATASOURCES:

<DATASOURCE NAME>

[Add text here]

EXISTING METHODOLOGIES

<METHODOLOGY NAME>

[Add text here]

Number of jobs displacements because of OS

[Add text here]

MEASUREMENT.

[Add text here]

EXISTING DATASOURCES:

<DATASOURCE NAME>

[Add text here]

EXISTING METHODOLOGIES

<METHODOLOGY NAME>

[Add text here]

Number of new start-ups/spin-offs founded thanks to OS

[Add text here]

MEASUREMENT.

[Add text here]

EXISTING DATASOURCES:

<DATASOURCE NAME>

[Add text here]

EXISTING METHODOLOGIES

<METHODOLOGY NAME>

[Add text here]

Known correlates

[Add text here]

Notes

[Add text here]

References

[Add Zotero bibliography here]

Cost savings

History

Description

[Add text here]

Metrics

Access costs savings thanks to OS

[Add text here]

MEASUREMENT.

[Add text here]

EXISTING DATASOURCES:

<DATASOURCE NAME>

[Add text here]

EXISTING METHODOLOGIES

<METHODOLOGY NAME>

[Add text here]

Labour costs savings given availability of OS inputs

[Add text here]

MEASUREMENT.

[Add text here]

EXISTING DATASOURCES:

<DATASOURCE NAME>

[Add text here]

EXISTING METHODOLOGIES

<METHODOLOGY NAME>

[Add text here]

Savings for data storage given the availability of OS data online

[Add text here]

MEASUREMENT.

[Add text here]

EXISTING DATASOURCES:

<DATASOURCE NAME>

[Add text here]

EXISTING METHODOLOGIES

<METHODOLOGY NAME>

[Add text here]

Known correlates

[Add text here]

Notes

[Add text here]

References

[Add Zotero bibliography here]

Part VI: Reproducibility

IMPACT OF OPEN SCIENCE ON REPRODUCIBILITY

Large-scale computation and the rise of data-driven methodologies have transformed the way scientific research is conducted in many disciplines. Open Science with its overarching goals of sharing research outcomes (resources, methods, or tools) as well as the flow of the actual research processes has become a key enabler for scientific discovery and faster knowledge spillover, contributing or leading those major shifts in science.

In the backdrop of these changes, reproducibility and replicability have raised critical concerns about the development and evolution of science and the way we generate reliable knowledge. Open Science could streamline the requisite processes addressing reproducibility challenges and accelerate the uptake of good practices about research integrity.

In PathOS, reproducibility refers strictly to computational reproducibility and computational non-reproducibility. Concretely, we define reproducibility as a continuous, “ongoing” process, ranging from

- i. systematic efforts to regenerate/reproduce computationally a previous study,
- ii. studies that re-use or build upon or expand (part of) the research outputs of a previous study,
- iii. studies that verify or confirm the results of a previous study by collecting and analysing new data,
- iv. studies that provide evidence that support or refute scientific claims and inferences from a previous study, to
- v. studies conducting meta-analyses and research synthesis by consolidating, evaluating, interpreting, and contextualizing findings from previous studies on a particular topic.

In the following sections, we delve into the following aspects in the intersection of OS and reproducibility providing relevant indicators (summarized in the table below) while keeping a pragmatic approach to what is feasible in terms of measuring, monitoring, and evaluating reproducibility.

Table 2: Reproducibility aspects and indicators

Reproducibility Aspect

Relevant Indicators Provided

Availability and transparency of research outputs to other studies	<ul style="list-style-type: none"> • Reuse of Code in research • Reuse of Data in research
Verification	<ul style="list-style-type: none"> • Consistency in reported numbers
Coherence of the approach	<ul style="list-style-type: none"> • Pre-registration of method/protocol
Reviews and checks on reproducibility of OS research	<ul style="list-style-type: none"> • Levels of replication found • Polarity of publications
Integrity of OS datasets, code and methods	<ul style="list-style-type: none"> • Impact of Open Code in research • Impact of Open Data in research • Inclusion in systematic reviews or meta-analyses

Consistency in reported numbers

History

Version	Revision date	Revision	Author
1.2	2023-08-30	Revisions	Thomas Klebel & Eva Kormann
1.1	2023-07-20	Incorporating revisions	Thomas Klebel
1.0	2023-05-11	First draft	Thomas Klebel

Description

Not all publications allow for verification of computational reproducibility by re-running the analysis on the original data. Reasons include code and/or data not being available, or potentially computations being too costly. One proxy used by researchers aiming to determine whether the numbers reported in a publication are credible is to check them for internal consistency. The general idea is to assess whether a given combination of numbers is mathematically possible. In finite samples, only certain combinations of mean values and standard deviations, or t-statistics with degrees of freedom and p-values are possible. If the reported numbers are found to be inconsistent, this might have multiple reasons, among them “human error, sloppiness, or questionable research practices” (Nuijten & Polanin, 2020). Consequently, if reported numbers are found to be inconsistent, the publication’s findings are not reproducible, and results from meta-analyses might be biased due to the inclusion of erroneous results (Nuijten & Polanin, 2020). In the absence of more direct measures, such proxies can be a useful way to assess general levels of rigour and reproducibility.

It must be re-emphasised that inconsistencies in reporting can have multiple sources, some of which also relate to the re-analysis itself. While the re-analysis of p-values (as in the implementation of *statcheck*, reported below) is relatively straightforward, checking means and standard deviations for internal validity presupposes a correct understanding of the reported experiments (e.g., how many respondents were assigned to a certain group/category, etc.). Inconsistencies in reported numbers and statistics therefore do not by themselves provide proof for fraud or scientific misconduct more broadly.

Metrics

Inconsistent p-values

Null-hypothesis significance testing (NHST) is a wide-spread practice in academic research. Conclusions are often drawn based on the distinction between statistically significant and insignificant results. Yet, many reported statistics in leading psychological journals have been found to be inconsistent (Nuijten et al., 2016). Checking reported p-values for consistency is therefore a good metric to gauge the overall consistency and internal validity of reported findings.

The metric has limitations in at least two domains: (a) in terms of measurement (see below), and (b) in terms of its applicability. Not all studies use NHST, and no similar metrics exist for e.g., Bayesian analyses.

MEASUREMENT.

Many test statistics in the realm of NHST allow for the calculation of p-values. Given a test statistic (e.g., t , χ^2) and a value for *degrees of freedom*, the p-value can be calculated. Reporting standards differ, and not all necessary elements might be provided by authors. However, for manuscripts that adhere to the American Psychological Association (APA) format for reporting test statistics (which mandates reporting test statistic, degrees of freedom, and the p-value), it is possible to recalculate the reported p-value based on the test statistic and the reported degrees of freedom. A key precondition for this measurement is thus the availability of reported statistics that adhere to the APA format.

There is no single data source for this metric. The metric can be computed based on any set of publications for which full texts are available.

EXISTING METHODOLOGIES

STATCHECK

The R package “statcheck” (Nuijten et al., 2023) implements the proposed metric. The package can process text strings, PDF and HTML documents, and whole folders containing PDF or HTML documents. In addition, a [web-app based on statcheck](#) is available. As discussed above, a key precondition and thus also a limitation for the approach used by statcheck is the availability of reported statistics that adhere to the APA format. If all necessary statistics are provided, statcheck can analyse results from correlations (r) and t , F , χ^2 , Z tests and Q tests^[1].

statcheck takes into account rounding in reporting, reports the recomputed p-value, indicating whether the values are inconsistent or grossly inconsistent, in case the recomputed p-value leads to the opposite conclusion on the statistical significance of the results (e.g., the publication reporting a result as statistically significant, while the recomputed p-value is above the conventional threshold of 0.05).

Inconsistent means and standard deviations.

The rationale for this indicator is similar to the one above, in exploiting mathematical properties of common summary statistics. Given a certain sample size, only specific values for means and standard deviations are possible. In small samples ($n < 100$), this can be used to test the plausibility of means, using the GRIM test (Brown & Heathers, 2016), and standard deviations, using the GRIMMER test (Anaya, 2016).

MEASUREMENT.

Consider the case of Likert-style answers (scale 1-5) and sample sizes from $n = 1$ to $n = 3$. The mean will always be a full integer for $n = 1$, multiples of 0.5 for $n = 2$, and multiples of 0.33 for $n = 3$. The sample sizes in this example are implausibly small, but the general principle holds for samples up to $n = 100$ when means are reported as rounded towards two decimals (Brown & Heathers, 2016). The logic of granularity in reported statistics has been extended to analyse measures of variation (Anaya, 2016), and to more general solutions that aim to recover possible datasets for a given combination of summary statistics (e.g., sample size, mean SD).

There is no single data source for this metric. The metric can be computed based on any set of publications for which full texts are available.

EXISTING METHODOLOGIES

GRIM TEST

The GRIM test exploits the property of granularity of means in small samples ($n < 100$). Given a certain sample size, and the use of Likert-style items that are based on integers, only certain values for the mean are possible. If a reported mean falls outside the range of possible mean values, it can be understood as being inconsistent.

An online implementation is available [here](#), and multiple implementations in statistical software environments are also available, e.g., in Jung and Allard (2023).

GRIMMER TEST

The GRIMMER test is an extension of the GRIM test, introduced by Anaya (2016), to test “the validity of reported measures of variability”, i.e., testing variances, standard deviations, and standard errors. An online calculator is available [here](#), and other implementations can be found in common statistical software environments, e.g., in Jung and Allard (2023).

SPRITE

The SPRITE technique builds on the intuition of the other two methodologies introduced above but avoids the limitation to small samples of the GRIM and GRIMMER tests (Heathers, Anaya, et al., 2018). Furthermore, it also lets the analyst check whether a reported *combination* of mean and SD is mathematically possible. Through the inspection of graphical output from the algorithm, it is also possible to spot pairs of means and SDs which are mathematically possible but highly unlikely in

practice (e.g., responses on a five-point Likert scale, with most responses being “1”, a few responses being “5”, and no responses for values 2-4). The SPRITE algorithm is only approximate. If SPRITE is unable to recover a potential dataset given input parameters, exhaustive searches of potential datasets should be performed in addition (see below on CORVIDS).

Implementations of SPRITE for R, Python and Matlab are available at <https://osf.io/pwjad/> (Heathers, Brown, et al., 2018).

CORVIDS

Although computationally efficient, the SPRITE algorithm is only approximate, and can therefore fail to uncover underlying datasets for given summary statistics. Building on the logic behind Diophantine equations, the CORVIDS algorithm can reliably uncover whether a given combination of summary statistics is mathematically possible and can provide all possible datasets that could lead to these statistics (Wilner et al., 2018).

An implementation of the algorithm is available in Python (Wood, 2018/2021).

References

- Anaya, J. (2016). *The GRIMMER test: A method for testing the validity of reported measures of variability* (e2400v1). PeerJ Inc. <https://doi.org/10.7287/peerj.preprints.2400v1>
- Brown, N. J. L., & Heathers, J. A. (2016). *The GRIM test: A simple technique detects numerous anomalies in the reporting of results in psychology* [Preprint]. PeerJ Preprints. <https://doi.org/10.7287/peerj.preprints.2064v1>
- Heathers, J. A., Anaya, J., Zee, T. van der, & Brown, N. J. L. (2018). *Recovering data from summary statistics: Sample Parameter Reconstruction via Iterative TEchniques (SPRITE)* (e26968v1). PeerJ Inc. <https://doi.org/10.7287/peerj.preprints.26968v1>
- Heathers, J. A., Brown, N. J. L., Zee, T. van der, & Anaya, J. (2018). *SPRITE*. OSF. <https://osf.io/pwjad/>
- Jung, L., & Allard, A. (2023). *scrutiny: Error Detection in Science* (0.2.4). <https://cran.r-project.org/web/packages/scrutiny/index.html>
- Nuijten, M. B., Epskamp, S., Slegers, W., Rife, S., Sakaluk, J., Laken, P. van der, Hartgerink, C., & Haroz, S. (2023). *statcheck: Extract Statistics from Articles and Recompute P-Values* (1.4.0). <https://cran.r-project.org/web/packages/statcheck/>
- Nuijten, M. B., Hartgerink, C. H. J., van Assen, M. A. L. M., Epskamp, S., & Wicherts, J. M. (2016). The prevalence of statistical reporting errors in psychology (1985–2013). *Behavior Research Methods*, 48(4), 1205–1226. <https://doi.org/10.3758/s13428-015-0664-2>

Nuijten, M. B., & Polanin, J. R. (2020). "statcheck": Automatically detect statistical reporting inconsistencies to increase reproducibility of meta-analyses. *Research Synthesis Methods*, 11(5), 574–579. <https://doi.org/10.1002/jrsm.1408>

Wilner, S., Wood, K., & Simons, D. J. (2018). *Complete recovery of values in Diophantine systems (CORVIDS)*. PsyArXiv. <https://doi.org/10.31234/osf.io/7shr8>

Wood, K. (2021). *CORVIDS* [Python]. <https://github.com/katherinemwood/corvids> (Original work published 2018)

1. <https://michelenuijten.shinyapps.io/statcheck-web/> ¹

Impact of Open Code in research

History

Version	Revision date	Revision	Author
1.1	2023-07-20	Revisions	Petros Stavropoulos
1.0	2023-05-11	First Draft	Petros Stavropoulos

Description

The impact of Open Code in research aims to capture the effect of making research code openly accessible and reusable on enhancing the reproducibility of research results, as open and accessible code is a cornerstone for verification and validation in science.

This indicator can be used to assess the level of openness and accessibility of research code within a specific scientific community or field and to identify potential barriers or incentives for the adoption of Open Code practices. It can also be used to track the reuse and subsequent impact related to reproducibility of Open Code, as well as to evaluate the effectiveness of policies and initiatives promoting Open Code practices.

Metrics

NCI for publications that have introduced Open Code

This metric calculates the Normalised Citation Impact (NCI) for publications that have introduced Open Code. By introducing Open Code, researchers enable others to scrutinize and build upon their computational methods, thus enhancing the potential for reproducibility and advancement of the field. The NCI metric primarily measures the citation impact of a publication, adjusted for differences in citation practices across scientific fields. However, citation impact can also be an indicator of research quality and reproducibility. Therefore, the NCI for publications that have introduced Open Code can serve as an indicator of both the visibility, influence, and reproducibility of research findings.

One limitation of this metric is that the use of NCI has been criticized for its potential biases and limitations, such as the inability to fully account for differences in research quality or the influence of non-citation-based impact measures. Therefore, we recommend using this metric in conjunction with other metrics in this document, such as software mentions and citations of the code repository, to obtain a more comprehensive assessment of the impact of Open Code practices on research output.

MEASUREMENT.

To measure this metric, the process begins with the identification of publications that have introduced Open Code. This is typically achieved by scrutinizing metadata within the code repositories and the publications, such as the repository's unique identifiers or the DOI (Digital Object Identifier). Alternatively, explicit mentions of the code repository, such as GitHub or GitLab URLs, within the publication text can be extracted to verify their openness. This can be performed manually or using automated tools.

Upon identification of the relevant publications, it is crucial to categorize them into their respective disciplines. The assignment of disciplines is typically based on the journal where the paper is published, the author's academic department, or the thematic content of the paper. Several databases provide such categorizations, such as [OpenAIRE](#), and [Web of Science](#) can be utilized.

Finally, the NCI (Normalised Citation Impact) score for each publication is calculated. The NCI measures the citation impact of a publication relative to the average for similar publications in the same discipline, publication year, and document type. It is computed by dividing the total number of citations the publication receives by the average number of citations received by all similar publications.

One limitation of this approach is that not all Open Code may be registered in code repositories, making it challenging to identify all relevant publications. Additionally, the accuracy of the NCI score may be affected by the availability and quality of citation data in different scientific fields. Therefore, it is important to carefully consider the potential biases and limitations of the data sources and methodologies used to measure this metric.

DATASOURCES

SCOPUS

Scopus is a comprehensive expertly curated abstract and citation database that covers scientific journals, conference proceedings, and books across various disciplines. Scopus provides enriched metadata records of scientific articles, comprehensive author and institution profiles, citation counts, as well as calculation of the articles' NCI score using their API.

One limitation of Scopus is that the calculation of NCI from Scopus only considers documents that are indexed in the Scopus database. This could lead to both underestimation and overestimation of the NCI for some publications, depending on how these publications are cited in sources outside the Scopus database.

EXISTING METHODOLOGIES

SCI NOBO TOOLKIT

The SciNoBo toolkit (Gialitsis et al., 2022) (Kotitsas et al., 2023) can be used to classify scientific publications into specific fields of science, which can then be used to calculate their NCI score. The

tool utilizes the citation-graph of a publication and its references to identify its discipline and assign it to a specific Field-of-Science (FoS) taxonomy. The classification system of publications is based on the structural properties of a publication and its citations and references organized in a multilayer network.

Furthermore, a new component of the SciNoBo toolkit, currently undergoing evaluation, involves an automated tool that employs Deep Learning and Natural Language Processing techniques to identify code/software mentioned in the text of publications and extract metadata associated with them, such as name, version, license, URLs etc. This tool can also classify whether the code/software has been introduced by the authors of the publication.

To measure the proposed metric, the tool can be used to identify relevant publications that have introduced code/software in conjunction with code repositories in GitHub, GitLab, or Bitbucket where the code/software is openly located and calculate their NCI score.

NCI for publication that have (re)used Open Code

This metric calculates the Normalised Citation Impact (NCI) for publications that have (re)used Open Code. It is a measure of the citation impact of research publications that have utilized Open Code, adjusted for differences in citation practices across scientific fields. The NCI for publications that have (re)used Open Code can indicate the potential impact of code sharing and reuse practices on the visibility and influence of research findings.

A limitation of this metric is that the use of NCI has been criticized for its potential biases and limitations, such as the inability to fully account for differences in research quality or the influence of non-citation-based impact measures. Therefore, we recommend to use this metric in conjunction with other metrics in this document, such as software mentions and citations of the code repository, to obtain a more comprehensive assessment of the impact of Open Code practices on research output.

MEASUREMENT.

To measure this metric, the process begins with the identification of publications that have (re)used Open Code. This is achieved by extracting explicit mentions of software/code mentions or code repositories, such as GitHub or GitLab URLs, within the publication text and then verifying their (re)use and openness. This can be performed manually or using automated tools.

Upon identification of the relevant publications, it is crucial to categorize them into their respective disciplines. The assignment of disciplines is typically based on the journal where the paper is published, the author's academic department, or the thematic content of the paper. Several databases provide such categorizations, such as [OpenAIRE](#), [Scopus](#) and [Web of Science](#) can be utilized.

Finally, the NCI (Normalised Citation Impact) score for each publication is calculated. The NCI measures the citation impact of a publication relative to the average for similar publications in the same discipline, publication year, and document type. It is computed by dividing the total number of citations the publication receives by the average number of citations received by all other similar publications.

One potential limitation of this approach is that not all Open Code may be registered in code repositories, making it challenging to identify all relevant publications. Additionally, the accuracy of the NCI score may be affected by the availability and quality of citation data in different scientific fields. Therefore, it is important to carefully consider the potential biases and limitations of the data sources and methodologies used to measure this metric.

DATASOURCES

SCOPUS

Scopus is a comprehensive expertly curated abstract and citation database that covers scientific journals, conference proceedings, and books across various disciplines. Scopus provides enriched metadata records of scientific articles, comprehensive author and institution profiles, citation counts, as well as calculation of the articles' NCI score using their API.

One limitation of Scopus is that the calculation of NCI from Scopus only considers documents that are indexed in the Scopus database. This could lead to both underestimation and overestimation of the NCI for some publications, depending on how these publications are cited in sources outside the Scopus database.

EXISTING METHODOLOGIES

SCI NO BO TOOLKIT

The SciNoBo toolkit (Gialitsis et al., 2022) (Kotitsas et al., 2023) can be used to classify scientific publications into specific fields of science, which can then be used to calculate their NCI score. The tool utilizes the citation-graph of a publication and its references to identify its discipline and assign it to a specific Field-of-Science (FoS) taxonomy. The classification system of publications is based on the structural properties of a publication and its citations and references organized in a multilayer network.

Furthermore, a new component of the SciNoBo toolkit, currently undergoing evaluation, involves an automated tool that employs Deep Learning and Natural Language Processing techniques to identify code/software mentioned in the text of publications and extract metadata associated with them, such as name, version, license, URLs etc. This tool can also classify whether the code/software has been (re)used by the authors of the publication.

To measure the proposed metric, the tool can be used to identify relevant publications that have (re)used code/software in conjunction with code repositories in GitHub, GitLab, or Bitbucket where the code/software is openly located and calculate their NCI score.

Code downloads/usage counts/stars from repositories

This metric measures the number of times an Open Code repository has been downloaded, used, or favoured, which can indicate the level of interest and impact of the code on the scientific community.

In terms of reproducibility, high usage counts or stars may indicate that a code/software is well-documented and easy to use. Furthermore, a widely used code/software is more likely to be updated and maintained over time, which can improve its reproducibility.

However, this metric may have limitations in capturing the impact of code that is not hosted in a public repository or downloaded through other means, such as direct communication between researchers. Additionally, usage counts and stars may not necessarily reflect the quality or impact of the code, and may be influenced by factors such as marketing and social media outreach. Therefore, we recommend using this metric in conjunction with other metrics in this document to obtain a more comprehensive assessment of the impact of Open Code practices on research output.

MEASUREMENT.

To measure this metric, data can be obtained from code repositories such as GitHub, GitLab, or Bitbucket. The number of downloads, usage counts, and stars can be extracted from the repository metadata. For example, on GitHub, this data is available through the API or by accessing the repository page. However, it is important to note that not all repository hosting providers may make this information publicly available, and some may only provide partial or incomplete usage data.

Additionally, the accuracy of the usage data may be affected by factors such as the frequency of updates, the type of license, and the accessibility of the code to different research communities.

The data can be computationally obtained using web scraping tools, API queries, or by manually accessing the download/usage count/star data.

DATASOURCES

GITHUB

GitHub is a web-based platform used for version control and collaborative software development. It allows users to create and host code repositories, including those for Open Source software and datasets. The number of downloads, usage counts, and stars on GitHub can be used as a metric for the impact and popularity of Open Code.

To measure this metric, we can search for the relevant repositories on GitHub and extract the relevant download, usage, and star data. This data can be accessed via the GitHub API, which provides

programmatic access to repository data. The API can be queried using HTTP requests, and the resulting data can be parsed and analysed using programming languages such as Python.

Following is an API call example for retrieving the stars of the `indicator_handbook` repository for the PathoOS-project from Github.

```
import requests

owner = "PathOS-project"
repo = "indicator_handbook"

url = f"https://api.github.com/repos/{owner}/{repo}/stargazers"
headers = {"Accept": "application/vnd.github.v3.star+json"}

response = requests.get(url, headers=headers)
stars = len(response.json())

print(f"The {owner}/{repo} repository has {stars} stars.")
```

GITLAB

GitLab is a web-based Git repository manager that provides source code management, continuous integration and deployment, and more. It can be used as a data source for metrics related to the usage of open-source software projects, including the number of downloads, stars, and forks.

To calculate the metric of code downloads/usage counts/stars from GitLab, we need to identify the relevant repositories and extract the relevant information. The number of downloads can be obtained by looking at the download statistics for a particular release of the repository. The number of stars can be obtained by looking at the number of users who have starred the repository. The number of forks can be obtained by looking at the number of users who have forked the repository.

To access this information, we can use the GitLab API.

BITBUCKET

Bitbucket is a web-based Git repository hosting service that allows users to host their code repositories, collaborate with other users and teams, and automate their software development workflows. It can be used as a data source for metrics related to the usage of open-source software projects, including the number of downloads, stars, and forks.

To calculate the metric of code downloads/usage counts/stars from Bitbucket, we need to identify the relevant repositories and extract the relevant information. The number of downloads can be obtained by looking at the download statistics for a particular release of the repository. The number of stars can be obtained by looking at the number of users who have starred the repository. The number of forks can be obtained by looking at the number of users who have forked the repository.

To access this information, we can use the Bitbucket API, which provides programmatic access to repository data. The API can be queried using HTTP requests, and the resulting data can be parsed and analyzed using programming languages such as Python.

EXISTING METHODOLOGIES

ENSURING THAT REPOSITORIES CONTAIN CODE

To ensure that a code repository (i.e. Github, Gitlab, Bitbucket) primarily contains code and not data or datasets, one can consider the following checks:

- Repository labelling: Look for repositories that are explicitly labelled as containing code or software. Many repository owners provide clear labels or descriptions indicating the nature of the content.
- File extensions: Check for files with common code file extensions, such as .py, .java, or .cpp. These file extensions are commonly used for code files, while data files often have extensions like .csv, .txt, or .xlsx.
- Repository descriptions and README files: Examine the repository descriptions and README files to gain insights into the content. Authors often provide information about the type of code included, its functionality, and its relevance to the project or software.
- Documentation: Some repositories include extensive documentation that provides details on the software, its usage, and how to contribute to the project. This indicates a greater likelihood that the repository primarily contains code.
- Existence of script and source folders: In some cases, the existence of certain directories like '/src' for source files or '/scripts' for scripts can indicate that the repository is primarily for code.

By considering these checks, we can ensure that the repository primarily contains code rather than data or datasets.

References

Gialitsis, N., Kotitsas, S., & Papageorgiou, H. (2022). SciNoBo: A Hierarchical Multi-Label Classifier of Scientific Publications. *Companion Proceedings of the Web Conference 2022*, 800–809. <https://doi.org/10.1145/3487553.3524677>

Kotitsas, S., Pappas, D., Manola, N., & Papageorgiou, H. (2023). SCINOBO: A novel system classifying scholarly communication in a dynamically constructed hierarchical Field-of-Science taxonomy. *Frontiers in Research Metrics and Analytics*, 8. <https://www.frontiersin.org/articles/10.3389/frma.2023.1149834>

Impact of Open Data in research

History

Version	Revision date	Revision	Author
1.1	2023-07-19	Revisions	Petros Stavropoulos
1.0	2023-05-11	First draft	Petros Stavropoulos

Description

The impact of Open Data in research aims to capture the effect of making research data openly accessible and reusable on enhancing the reproducibility of research results, as open and accessible code is a cornerstone for verification and validation in science.

The indicator can be used to assess the level of openness and accessibility of research data within a specific scientific community or field, and to identify potential barriers or incentives for the adoption of Open Data practices.

Metrics

NCI for publications that have introduced Open Datasets

This metric calculates the Normalised Citation Impact (NCI) for publications that have introduced Open Datasets. By introducing Open Datasets, researchers enable others to access and verify their findings, thus enhancing the potential for reproducibility. The NCI metric primarily measures the citation impact of a publication, adjusted for differences in citation practices across scientific fields. However, citation impact can also be an indicator of research quality and reproducibility. Therefore, the NCI for publications that have introduced Open Datasets can serve as an indicator of both the visibility, influence, and reproducibility of research findings.

One limitation of this metric is that the use of NCI has been criticized for its potential biases and limitations, such as the inability to fully account for differences in research quality or the influence of non-citation-based impact measures. Therefore, we recommend using this metric in conjunction with other metrics in this document to obtain a more comprehensive assessment of the impact of Open Data practices on research output.

MEASUREMENT.

To measure this metric, the process begins with the identification of publications that have introduced Open Datasets. This is typically achieved by scrutinizing metadata within the datasets and the publications, such as the DOI (Digital Object Identifier). Alternatively, explicit mentions of the dataset within the publication text can be extracted and verify their openness. This can be performed manually or using automated tools.

Upon identification of the relevant publications, it's crucial to categorize them into their respective disciplines. The assignment of disciplines is typically based on the journal where the paper is published, the author's academic department, or the thematic content of the paper. Several databases provide such categorizations, such as [OpenAIRE](#), [Scopus and Web of Science](#) can be utilized.

Finally, the NCI (Normalised Citation Impact) score for each publication is calculated. The NCI measures the citation impact of a publication relative to the average for similar publications in the same discipline, publication year, and document type. It is computed by dividing the total number of citations the publication receives by the average number of citations received by all other similar publications.

One potential limitation of this approach is that not all Open Datasets may be registered in data repositories, making it challenging to identify all relevant publications. Additionally, the accuracy of the NCI score may be affected by the availability and quality of citation data in different scientific fields. Therefore, it is important to carefully consider the potential biases and limitations of the data sources and methodologies used to measure this metric.

DATASOURCES

OPENAIRE

OpenAIRE is a European platform that provides Open Access to research outputs, including publications, datasets, and software. OpenAIRE collects metadata from various data sources, including institutional repositories, data repositories, and publishers.

For the NCI for publications that have introduced Open Datasets metric, we can use OpenAIRE to identify publications that have introduced Open Datasets. We can search for publications by looking for OpenAIRE records that have a dataset identifier in the references section or by using OpenAIRE's API to search for publications that are linked to a specific dataset.

One limitation of using OpenAIRE for this metric is that not all Open Datasets may be registered in OpenAIRE, which could lead to underestimation of the number of publications that have introduced Open Datasets.

SCOPUS

Scopus is a comprehensive expertly curated abstract and citation database that covers scientific journals, conference proceedings, and books across various disciplines. Scopus provides enriched metadata records of scientific articles, comprehensive author and institution profiles, citation counts, as well as calculation of the articles' NCI score using their API.

One limitation of Scopus is that the calculation of NCI from Scopus only considers documents that are indexed in the Scopus database. This could lead to both underestimation and overestimation of the NCI for some publications, depending on how these publications are cited in sources outside the Scopus database.

EXISTING METHODOLOGIES

SciNoBo TOOLKIT

The SciNoBo toolkit (Gialitsis et al., 2022) (Kotitsas et al., 2023) can be used to classify scientific publications into specific fields of science, which can then be used to calculate their NCI score. The tool utilizes the citation-graph of a publication and its references to identify its discipline and assign it to a specific Field-of-Science (FoS) taxonomy. The classification system of publications is based on the structural properties of a publication and its citations and references organized in a multilayer network.

Furthermore, a new component of the SciNoBo toolkit, currently undergoing evaluation, involves an automated tool that employs Deep Learning and Natural Language Processing techniques to identify datasets mentioned in the text of publications and extract metadata associated with them, such as name, version, license, URLs etc. This tool can also classify whether the datasets has been introduced by the authors of the publication.

To measure the proposed metric, the tool can be used to identify relevant publications that have introduced datasets and calculate their NCI score.

NCI for publications that have (re)used Open Datasets

This metric calculates the Normalised Citation Impact (NCI) for publications that have (re)used Open Datasets. It is a measure of the citation impact of research publications that have utilized Open Datasets, adjusted for differences in citation practices across scientific fields. The NCI for publications that have (re)used Open Datasets can indicate the potential impact of data sharing and reuse practices on the visibility and influence of research findings. A higher NCI score indicates a greater level of scientific collaboration and data sharing within a specific scientific community or field, suggesting that the availability of Open Datasets can contribute to the impact and recognition of research, thus indirectly indicating its potential for reproducibility.

A limitation of this metric is that the use of NCI has been criticized for its potential biases and limitations, such as the inability to fully account for differences in research quality or the influence of non-citation-based impact measures. Therefore, we recommend using this metric in conjunction with other metrics in this document to obtain a more comprehensive assessment of the impact of Open Data practices on research output.

MEASUREMENT.

To measure this metric, the process begins with the identification of publications that have (re)used Open Datasets. This is typically achieved by scrutinizing metadata within the datasets and the publications, such as the DOI (Digital Object Identifier). Alternatively, explicit mentions of the dataset within the publication text can be extracted and verify their openness. This can be performed manually or using automated tools.

Upon identification of the relevant publications, it's crucial to categorize them into their respective disciplines. The assignment of disciplines is typically based on the journal where the paper is published, the author's academic department, or the thematic content of the paper. Several databases provide such categorizations, such as [OpenAIRE](#), [Scopus](#) and [Web of Science](#) and can be utilized.

Finally, the NCI (Normalised Citation Impact) score for each publication is calculated. The NCI measures the citation impact of a publication relative to the average for similar publications in the same discipline, publication year, and document type. It is computed by dividing the total number of citations the publication receives by the average number of citations received by all other similar publications.

One potential limitation of this approach is that not all Open Datasets may be registered in data repositories, making it challenging to identify all relevant publications. Additionally, the accuracy of the NCI score may be affected by the availability and quality of citation data in different scientific fields. Therefore, it is important to carefully consider the potential biases and limitations of the data sources and methodologies used to measure this metric.

DATASOURCES

SCOPUS

Scopus is a comprehensive expertly curated abstract and citation database that covers scientific journals, conference proceedings, and books across various disciplines. Scopus provides enriched metadata records of scientific articles, comprehensive author and institution profiles, citation counts, as well as calculation of the articles' NCI score using their API.

One limitation of Scopus is that the calculation of NCI from Scopus only considers documents that are indexed in the Scopus database. This could lead to both underestimation and overestimation of

the NCI for some publications, depending on how these publications are cited in sources outside the Scopus database.

EXISTING METHODOLOGIES

SciNoBo TOOLKIT

The SciNoBo toolkit (Gialitsis et al., 2022) (Kotitsas et al., 2023) can be used to classify scientific publications into specific fields of science, which can then be used to calculate their NCI score. The tool utilizes the citation-graph of a publication and its references to identify its discipline and assign it to a specific Field-of-Science (FoS) taxonomy. The classification system of publications is based on the structural properties of a publication and its citations and references organized in a multilayer network.

Furthermore, a new component of the SciNoBo toolkit, currently undergoing evaluation, involves an automated tool that employs Deep Learning and Natural Language Processing techniques to identify datasets mentioned in the text of publications and extract metadata associated with them, such as name, version, license, URLs etc. This tool can also classify whether the datasets has been (re)used by the authors of the publication.

To measure the proposed metric, the tool can be used to identify relevant publications that have (re)used datasets and calculate their NCI score.

Dataset downloads/usage counts/stars from repositories

This metric measures the number of downloads, usage counts, or stars (depending on the repository) of a given Open Dataset. It provides an indication of the level of interest and use of the dataset by the scientific community, and can serve as a proxy for the potential impact of the dataset on scientific research. It should be noted that this metric may not capture the full impact of Open Datasets on scientific research, as the number of downloads or usage counts may not necessarily reflect the quality or impact of the research that utilizes the dataset.

In terms of reproducibility, high usage counts or stars may indicate that a dataset is well-documented and easy to use. Furthermore, a widely used dataset is more likely to be updated and maintained over time, which can improve its reproducibility.

One limitation of this metric is that it only captures usage of Open Datasets from specific repositories and may not reflect usage of the same dataset that is hosted elsewhere. Additionally, differences in repository usage and user behaviour may affect the comparability of download/usage count/star data across repositories. Finally, this metric does not capture non-public uses of Open Datasets, such as internal use within an organization or personal use by researchers, which may also contribute to the impact of Open Datasets on scientific research.

MEASUREMENT.

To measure this metric, we can use data from various data repositories, such as DataCite and Zenodo, or data from OpenAIRE, which provide download or usage statistics for hosted datasets. We can also use platforms such as GitHub or GitLab, which provide star counts as a measure of user engagement with Open-Source code repositories that may include Open Datasets. However, it is important to note that different repositories may provide different types of usage statistics, and these statistics may not be directly comparable across repositories. Additionally, not all repositories may track usage statistics, making it difficult to obtain comprehensive data for all Open Datasets.

The data can be computationally obtained using web scraping tools, API queries, or by manually accessing the download/usage count/star data for each dataset.

DATASOURCES

DATA CITE

DataCite is a global registry of research data repositories and datasets, providing persistent identifiers for research data to ensure that they are discoverable, citable, and reusable. The dataset landing pages on DataCite contain information about the dataset, such as metadata, version history, and download statistics. This information can be used to measure the usage and impact of Open Datasets.

To calculate the usage count of a dataset, we can use the “Views” field provided on the dataset landing page on DataCite, which indicates the number of times the landing page has been accessed. To calculate the number of downloads, we can use the “Downloads” field, which indicates the number of times the dataset has been downloaded. The number of stars or likes can be used as a measure of the popularity of the dataset among users.

ZENODO

Zenodo is a general-purpose open-access repository developed by CERN to store scientific data. It accepts various types of research outputs, including datasets, software, and publications. Zenodo assigns a unique digital object identifier (DOI) to each deposited item, which can be used to track its usage and citations.

To calculate the metric of dataset views and downloads from Zenodo, we can extract the relevant metadata from the Zenodo API, which provides programmatic access to the repository’s contents. The API allows us to retrieve information about a specific item, such as its title, author, publication date, and number of views / downloads. We can then aggregate this data to obtain usage statistics for a particular dataset or set of datasets.

OPENAIRE

OpenAIRE is a European Open Science platform that provides access to millions of openly available research publications, datasets, software, and other research outputs. OpenAIRE aggregates content

from various sources, including institutional and thematic repositories, data archives, and publishers. This platform provides usage statistics for each research output in the form of downloads, views, and citations, which can be used to measure the impact and reuse of research outputs, including Open Datasets.

To calculate this metric using OpenAIRE, we can retrieve the download and view counts for the relevant Open Datasets, which can be accessed through the OpenAIRE REST API. The API returns JSON-formatted metadata for each research output, which includes information such as the title, authors, publication date, download counts, and view counts. The download and view counts can be used to calculate the total number of times the dataset has been accessed or viewed, respectively.

GITHub

GitHub is a web-based platform used for version control and collaborative software development. It allows users to create and host code repositories, including those for Open-Source software and datasets. The number of downloads, usage counts, and stars on GitHub can be used as a metric for the impact and popularity of Open Datasets.

To measure this metric, we can search for the relevant repositories on GitHub and extract the relevant download, usage, and star data. This data can be accessed via the GitHub API, which provides programmatic access to repository data. The API can be queried using HTTP requests, and the resulting data can be parsed and analysed using programming languages such as Python.

GITLAB

GitLab is a web-based Git repository manager that provides source code management, continuous integration and deployment, and more. It can be used as a data source for metrics related to the usage of open-source software projects, including the number of downloads, stars, and forks.

To calculate the metric of dataset downloads/usage counts/stars from GitLab, we need to identify the relevant repositories and extract the relevant information. The number of downloads can be obtained by looking at the download statistics for a particular release of the repository. The number of stars can be obtained by looking at the number of users who have starred the repository. The number of forks can be obtained by looking at the number of users who have forked the repository.

To access this information, we can use the GitLab API.

EXISTING METHODOLOGIES

ENSURING THAT REPOSITORIES CONTAIN DATA

To ensure that a repository (i.e. Github, Gitlab) primarily contains research data and not code, we can consider the following methodology:

- Repository labelling: Look for repositories that are explicitly labelled as containing data or datasets. Many repository owners provide clear labels or descriptions indicating the nature of the content.
- File extensions: Check for files with common data file extensions, such as .csv, .txt, or .xlsx. These file extensions are commonly used for data files, while code files often have extensions like .py, .java, or .cpp.
- Repository descriptions and README files: Examine the repository descriptions and README files to gain insights into the content. Authors often provide information about the type of data included and its relevance to research.
- Data availability statements: Some repositories include data availability statements that provide details on where the data supporting the reported results can be found. These statements may include links to publicly archived datasets or references to specific repositories.
- Supplementary materials: In some cases, authors may publish supplementary materials alongside their research articles. These materials can include datasets and provide additional information about the data and its relevance to the research.

By considering these checks, we can ensure that the repository primarily contains research data rather than code.

Downloads / views of published DMPs

This metric measures the number of downloads or views of published Data Management Plans (DMPs) from data repositories, such as DataCite, Zenodo, or institutional repositories. A DMP is a document that outlines how research data will be managed throughout a research project, including details on data collection, storage, sharing, and preservation. The number of downloads or views of published DMPs can indicate the level of interest and engagement of researchers and other stakeholders in Open Data practices and the importance of data management planning in the research process, thereby reflecting the adoption of good data management practices that indirectly contribute to overall research reproducibility.

A limitation of this metric is that it only captures the number of downloads or views of DMPs, which may not necessarily indicate the actual implementation of the DMP or the quality of the data management practices. Therefore, it is important to use this metric in conjunction with other metrics in this document to obtain a more comprehensive assessment of the impact of Open Data practices on research output.

MEASUREMENT.

To measure this metric, we can start by identifying published DMPs in data repositories, such as DataCite, Zenodo, or institutional repositories. To identify the relevant DMPs, we can utilize search features and application programming interfaces (APIs) provided by these data repositories, conduct keyword searches related to the specific research or project, and review the metadata associated with each DMP for relevance. Once we have identified the relevant DMPs, we can track the number of downloads or views of these DMPs over a specified period of time.

Potential measurement problems and limitations of this metric include the possibility of multiple downloads or views by the same user, which can inflate the metric. Additionally, the number of downloads or views may not reflect the actual use or implementation of the DMP, as some researchers may download or view DMPs out of curiosity or to gain insight into best practices. Therefore, it is important to interpret the results of this metric in the context of other metrics and qualitative data on the use and effectiveness of DMPs.

Existing data sources and methodologies for this metric include the data repositories and web analytics tools mentioned above. DataCite and Zenodo provide download counts for their published content, including DMPs, while Google Analytics can be used to track views of DMPs on institutional or funder websites. However, there may be gaps in the availability of download or view counts for DMPs published on other platforms or websites. In such cases, it may be necessary to manually track the number of downloads or views through user surveys or by contacting individual users who have downloaded or viewed the DMP.

DATASOURCES

DATA CITE

DataCite is a global registry for research data that provides persistent identifiers (DOIs) for research datasets. To measure the number of downloads or views of published DMPs in DataCite, we can use the DataCite REST API to search for DMPs by the keyword "Data Management Plan" and filter the results by the download count or view count metadata. The API also allows filtering by date range and repository location, which can provide additional context for the measurement.

One potential limitation of using DataCite for this metric is that not all DMPs may be registered with DataCite, and the search results may not capture all relevant DMPs. Additionally, the download or view count metadata may not always accurately reflect the actual use or engagement with the DMP, as these metrics can be affected by factors such as availability, accessibility, and discoverability of the DMP.

ZENODO

Zenodo is a data repository that allows researchers to upload and share research outputs, including DMPs. To calculate the number of downloads or views of published DMPs on Zenodo, we can use the Zenodo REST API to retrieve the relevant metadata for each DMP, such as the number of views

and downloads. This can be done by searching for DMPs on Zenodo using their unique identifiers or keywords, and then extracting the relevant metadata for each search result.

One limitation of using Zenodo to measure this metric is that not all DMPs may be published on this repository. Additionally, the number of views and downloads may not necessarily reflect the actual use or implementation of the DMP, as users may simply be browsing or downloading the document for reference purposes. Finally, the number of downloads and views may be influenced by factors such as the popularity of the topic or the visibility of the DMP on the repository.

Number of datasets reused inside DMPs

This metric measures the number of datasets that are reused in Data Management Plans (DMPs). A DMP is a document that outlines how research data will be managed throughout a research project, including details on data collection, storage, sharing, and preservation. The number of datasets reused in DMPs can indicate the level of engagement of researchers in Open Data practices and the potential impact of data sharing and reuse practices on research output. This metric can also serve as a proxy for reproducibility, as datasets explicitly cited and reused in multiple DMPs are likely to be more robust and have undergone scrutiny, thus facilitating other researchers in verifying or replicating results. Furthermore, the standardization of data storage, management, and processing practices encouraged by DMPs can indirectly promote reproducibility.

A limitation of this metric is that it may not capture the full range of Open Data practices that are being utilized by researchers, such as the sharing of data outside of DMPs or the creation of new datasets for reuse. Additionally, the metric may not capture the quality or impact of the datasets being reused in DMPs. Therefore, it is important to use this metric in conjunction with other metrics in this document to obtain a more comprehensive assessment of the impact of Open Data practices on research output.

MEASUREMENT.

To measure this metric, we can start by identifying published DMPs in data repositories, such as DataCite, Zenodo, or institutional repositories. We can then analyse the content of published DMPs to identify the datasets that are being reused through automated text mining techniques (e.g., using the SciNoBo toolkit).

However, there are some limitations to this approach. One limitation is that not all DMPs are publicly available, which may limit the scope of the analysis. Additionally, automated techniques may not capture all instances of dataset reuse if they are not explicitly mentioned in the text of the DMP.

DATASOURCES

DATA CITE

DataCite is a metadata repository that provides persistent identifiers for research datasets. It collects metadata from various sources, including data centres, publishers, and institutional repositories. The metadata includes information on the dataset, such as the title, author, publisher, date of publication, and the identifier of the dataset.

To measure the number of datasets reused inside DMPs using DataCite, we can search for published DMPs in DataCite, extract the metadata for each DMP, and analyse the content of the DMP to identify the datasets that are being reused. This can be done using automated text mining techniques to identify dataset names or identifiers mentioned in the DMP.

However, it is important to note that not all DMPs may be available in DataCite, and some datasets may not have persistent identifiers, which may limit the scope of the analysis. Additionally, automated text mining techniques may not capture all instances of dataset reuse if they are not explicitly mentioned in the text of the DMP.

To obtain the metadata for published DMPs in DataCite, we can use the DataCite REST API to search for DMPs that have been registered with DataCite. The metadata can be obtained in various formats, including JSON and XML.

ZENODO

Zenodo is a general-purpose data repository that allows users to upload any kind of research output, including datasets and data management plans (DMPs). Zenodo assigns a unique Digital Object Identifier (DOI) to each uploaded item, which can be used to track usage and reuse.

To measure the number of datasets reused inside DMPs based on Zenodo, we can search for published DMPs on Zenodo using keywords and filters, such as the “data management plan” keyword and the “DMP” tag. Once we have identified a set of DMPs, we can use automated text mining techniques to identify the datasets that are being reused. This can involve searching for mentions of dataset names or identifiers in the text of the DMPs.

However, it is important to note that not all DMPs on Zenodo may contain information on reused datasets, and some datasets may not be explicitly mentioned in the text of the DMP. Additionally, the automated text mining techniques used to identify reused datasets may not capture all instances of reuse, particularly if the datasets are referred to in a non-standard way or if they are combined with other datasets.

EXISTING METHODOLOGIES

SCI NOBO TOOLKIT

The SciNoBo toolkit (Gialitsis et al., 2022) (Kotitsas et al., 2023) has a new component, currently undergoing evaluation, which involves an automated tool that employs Deep Learning and Natural Language Processing techniques to identify datasets mentioned in scientific text (i.e., the text of a

DMP) and extract metadata associated with them, such as name, version, license, URLs etc. This tool can also classify whether the dataset has been (re)used by the authors of the DMP.

To use the tool to measure the proposed metric, we can provide a collection of DMPs as input to the tool to extract all the datasets mentioned in the text, along with their metadata. We can then analyse them to identify which datasets have been reused by the authors of the DMP, as classified by the machine learning algorithms of the tool.

One limitation of this methodology is that it may not capture all instances of dataset reuse if they are not explicitly mentioned in the text of the DMP. Additionally, the machine learning algorithms used by the tool may not always accurately classify whether a dataset has been reused, and may require manual validation.

References

Gialitsis, N., Kotitsas, S., & Papageorgiou, H. (2022). SciNoBo: A Hierarchical Multi-Label Classifier of Scientific Publications. *Companion Proceedings of the Web Conference 2022*, 800–809. <https://doi.org/10.1145/3487553.3524677>

Kotitsas, S., Pappas, D., Manola, N., & Papageorgiou, H. (2023). SCINOBO: A novel system classifying scholarly communication in a dynamically constructed hierarchical Field-of-Science taxonomy. *Frontiers in Research Metrics and Analytics*, 8. <https://www.frontiersin.org/articles/10.3389/frma.2023.1149834>

Inclusion in systematic reviews or meta-analyses

History

Version	Revision date	Revision	Author
1.2	2023-08-30	Revisions	Eva Kormann & Thomas Klebel
1.1	2023-07-20	Revisions	Thomas Klebel
1.0	2023-05-02	First draft	Eva Kormann

Description

Systematic reviews and meta-analyses are very useful methodologies to synthesize scientific literature on a certain topic. Inclusion of a paper in such a systematic review or meta-analysis can be used as an indicator for reproducibility, since in the process of reviewing literature and assessing inclusion criteria, judgements are made about the quality of a paper (e.g., of methods and results). For instance, the PRISMA guidelines include specification of risk of bias assessment (Page et al., 2021). Instead of directly investigating the quality of papers, inclusion in systematic reviews and meta-analyses can therefore be taken as a proxy. Papers passing the quality check of a systematic review or meta-analysis could be expected to be more reproducible than papers failing this check. This indicator, however, is dependent on the existence of systematic reviews or meta-analyses for a certain topic and gathering enough information for comparisons might be challenging. Publications available as Open Access might be more often included, since none of them are excluded from such studies due to unavailability.

Metrics

Number of citations in systematic reviews or meta-analyses

The inclusion of a research paper can be indicated through the number of times it has already been cited in a systematic review or meta-analysis. This metric, however, has some limitations. Citation by a systematic review or meta-analysis by itself is not a reliable indicator for whether a given study was included in a review. It could also be cited within the background or discussion section, or as an excluded source.

MEASUREMENT.

This metric could be measured by analysing the sources a paper is cited by. The type of publication needs to be extracted for all citing sources. The number of citing sources that are systematic reviews or meta-analyses can then be counted.

DATASOURCES

LITERATURE DATABASES

Data sources for this metric are literature databases. These are suitable insofar they provide information on all sources a paper is cited by. Examples for such databases are Web of Science (WoS), Scopus, OpenAlex or Dimensions. For all citing sources, the type of publication needs to be determined from available metadata (e.g., title, abstract and keywords).

Number of inclusions in systematic reviews or meta-analyses

For a specific paper the number of systematic reviews and meta-analyses that have included it in the data synthesis can be counted. Inclusion in a systematic review or meta-analysis does not only rely on the topic of a paper and the scope of the review or meta-analysis, but also on some quality assessment (e.g., of methods or results). This metric has its limitation, since quality assessment of studies might differ significantly between different conducted systematic reviews and meta-analyses in criteria, strictness, etc. While inclusion in a systematic review or meta-analysis indicates that some form of quality check was passed by a paper, the thoroughness of this assessment is not indicated by the number of inclusions.

MEASUREMENT.

All citing sources of a given paper are classified by their type, filtering out systematic reviews and meta-analyses. This could be done by retrieving keywords in the titles of publications, such as "systematic review", since publications adhering to the PRISMA guidelines must indicate this in the title. For the retrieved sources, one would have to manually determine whether they include the given paper in their data synthesis. The number of systematic reviews and meta-analyses where this is the case can then be counted.

A key question in applying this metric would be how to interpret the aggregated counts: what number of inclusions would indicate a "robust" or "reproducible" finding? Furthermore, absence of inclusion cannot be taken as a sign for low reproducibility, because there are many reasons why this might be the case: no systematic reviews conducted yet in this field, study not included in prior reviews due to general exclusion criteria (e.g., different language, sample population or study target, etc.).

DATASOURCES

There is no single data source for this metric. Data needs to be extracted newly by the described methodologies for papers of interest.

EXISTING METHODOLOGIES

SEMANTIC ANALYSIS OF FULL TEXT

One potential methodology is the semantic analysis of full texts of citing systematic reviews or meta-analyses themselves. Using a semantic analysis of the full text, it might be possible to determine where a specific paper is cited and whether a statement is made about inclusion. However, the methodology of extracting this specific information through semantic analysis is not yet developed.

SUPPLEMENTARY MATERIAL/DATA PROVIDED FOR SYSTEMATIC REVIEWS/META-ANALYSES

Some systematic reviews or meta-analyses make data available that was gathered during the screening and/or data charting process. From this data, information can be extracted about the inclusion status of a specific paper. This data might be available in repositories, e.g., Figshare, OSF, or Zenodo, but at the moment there is no systematic database covering this data.

Number of exclusions from systematic reviews or meta-analyses due to methodological issues or bias

Opposite to inclusion, the number of systematic reviews and meta-analyses that have excluded a specific paper in the data synthesis can also be counted. However, exclusion for the reason of being out of scope is no indicator of quality. Therefore, only exclusions specifically due to methodological issues (such as insufficient reporting, noticeable errors or questionable choice of methods) or suspected bias within a study can be of interest for this metric. To determine this, however, information on specific reasons for exclusion must be given which might be given less frequently than general information about inclusion.

MEASUREMENT.

All citing sources of a given paper should be classified by their type, and then be filtered on citing systematic reviews and meta-analyses. For these sources it should be checked whether they exclude the given paper due to methodological issues or bias that were identified (not because of scope). The number of systematic reviews and meta-analyses where this is the case can then be counted.

DATASOURCES

There is no single data source for this metric. Data needs to be extracted newly by the described methodologies for papers of interest.

EXISTING METHODOLOGIES

SEMANTIC ANALYSIS OF FULL TEXT

One potential methodology is the semantic analysis of full texts of citing systematic reviews or meta-analyses themselves. With a semantic analysis of the full text, it might be possible to determine where a specific paper is cited and whether a statement is made about exclusion. Clear information must be given in the full text about the reason of exclusion for it to be counted, e.g., through a variable specifying the reason for exclusion within the dataset. However, the methodology of extracting this specific information through semantic analysis is not yet developed.

SUPPLEMENTARY MATERIAL/DATA PROVIDED FOR SYSTEMATIC REVIEWS/META-ANALYSES

Some systematic reviews or meta-analyses make data available that was gathered during the screening and/or data charting process. From this data, information can be extracted about the exclusion status of a specific paper. Clear information must be given about the reason of exclusion for it to be counted, e.g., through a variable specifying the reason for exclusion within the dataset.

References

Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., Shamseer, L., Tetzlaff, J. M., Akl, E. A., Brennan, S. E., Chou, R., Glanville, J., Grimshaw, J. M., Hróbjartsson, A., Lalu, M. M., Li, T., Loder, E. W., Mayo-Wilson, E., McDonald, S., ... Moher, D. (2021). The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *BMJ*, 372, n71. <https://doi.org/10.1136/bmj.n71>

Level of replication

History

Version	Revision date	Revision	Author
1.2	2023-08-30	Revisions	Eva Kormann & Thomas Klebel
1.1	2023-07-20	Revisions	Thomas Klebel
1.0	2023-04-26	First draft	Eva Kormann

Description

Replication is often defined as the process of repeating a study with the same methodology: generating new data that can then be analysed similarly to the original study. A study is considered successfully replicated when the replication yields the same results as the original. The term replicability is closely related to the term reproducibility and sometimes used interchangeably. However, terms can be differentiated by referring to *reproduction* when repeating the analysis with the original study data and referring to *replication* when repeating the entire study creating new data to analyse (Goodman et al., 2016).

A certain number of replication attempts is expected to fail due to the chance of false positives / false negatives in the original or replication studies (Marino, 2018). However, higher proportions of failed replication attempts might, for instance, be signs of insufficient reporting, biases (cognitive or related to the publication process, i.e., publication bias) or methodological issues (such as low statistical power), and therefore challenge the validity and credibility of results. Low levels of replication indicate flaws in research practices and potential waste of research effort (Munafò et al., 2017).

The level of successful replication represents a direct indicator for reproducibility. It can also serve as an indicator for research quality in a broader sense, since issues related to reporting or methods increase the risk for failed replication. The extent to which research findings are replicable can be examined over time and in relation to the employed research practices.

Metrics

Number (%) of studies found to successfully replicate

The level of replication can be measured by counting the number or calculating the proportion of studies that were found to successfully replicate. Data on success of replication attempts, however,

is limited. Re-performing a study requires substantial resources. For a large share of the published literature no data on the number or share of successful replications is available. Additionally, it might be impossible for some studies to be replicated, e.g., because a one-time event was studied. For these types of studies, levels of replication cannot be assessed.

MEASUREMENT.

Levels of replication can directly be examined by analysing the proportion of successful replication attempts. Therefore, the number of replication attempts and their success/failure need to be measured. Then, the percentage can be calculated as the proportion of successful replications within all replication attempts. Difficulties, however, lie in the definition of what constitutes a successful replication. A common argument in the literature on replication and reproducibility is that exact replications are not possible, since the exact setting, context, sample, etc. usually cannot be recreated fully (Nosek et al., 2022; Schmidt, 2009). A study might be seen as a replication “when the differences to the original study are believed to be irrelevant for obtaining the evidence about the same finding” (Nosek et al., 2022), but this cannot easily be determined and appears quite subjective.

To calculate the number or percentage of successful replications, a dichotomous indicator is needed for success of replication (Nosek et al., 2022). Multiple ways of indicating success of replication are in use (see Nosek et al., 2022):

- The null hypothesis is rejected in the same direction ($p < \alpha$).
- An estimate is within a confidence or prediction interval.
- The detected effect size is consistent with the original study.
- The findings are similar when assessed subjectively.

There are also continuous measures that can be dichotomized:

- Bayes factors for comparison of original and replication findings.
- Bayesian tests to compare null distribution and posterior distribution of the original study.

DATASOURCES

There is no single data source for this metric. Data needs to be extracted from existing publications or gathered newly by employing these methodologies.

EXISTING METHODOLOGIES:

REPLICATION PROJECTS/STUDIES

Many studies or projects on the topic of replicability pursue the goal of conducting a multitude of replication attempts following the same process or using the same measures of success to determine the proportion of replicable findings. Some of these studies and projects are concentrated on specific fields of research, e.g., Open Science Collaboration (2015) and SCORE Project (including subprojects like the [repliCATS Project](#)) for social-behavioural science. The “Many Labs” studies also had their

initial focus on psychology but have since spread into other disciplines (Klein et al., 2014; Stroebe, 2019). Approaches to investigate replicability have reached a multitude of disciplines, e.g., the field of [humanities](#).

SCOPING REVIEW PAPERS

Since multiple reformed studies are needed to determine the percentage of successful replications, standalone replication attempts provide only limited information. Scoping reviews are one way to synthesize singular replication attempts to gain an overview and to estimate the percentage of successful replications more precisely. However, singular replication studies synthesized within a review might be inconsistent in their procedures and employ different measures of success, complicating synthesis or comparison.

Number (%) of studies reported to successfully replicate

Replication attempts might not be published (especially in the case of repeated replications of the same study) and might for instance only be conducted internally within a research group. To gather information about these replications attempts, reports can be obtained from researchers about the total number of replications they attempt and the number of studies out of those that they were able to replicate.

MEASUREMENT.

In addition to directly measuring the success of replication studies, the level of replication can also be assessed by surveying researchers. They can report retrospectively about their replication attempts and indicate or estimate the level of replication they encountered. These reports, however, might be less systematic, detailed or objective than studies or projects directly reperforming studies. However, they can be acquired with fewer resources.

DATASOURCES

There is no single data source for this metric. Data needs to be extracted from existing publications or gathered newly by employing these methodologies.

EXISTING METHODOLOGIES:

SURVEYS

Experiences of researchers with replications and the level of replication they have encountered in their work can be investigated through surveys. There, questions can be included about previous replication attempts and their success and about general estimates of replicability. While using the term "reproducibility" instead of "replicability", the Nature survey by Baker (2016) employed this method using similar questions.

Number (%) of studies predicted to successfully replicate

Since replication attempts require substantial resources, they cannot be conducted for all studies. Without studies or researcher reports available to assess levels of replication, the number of studies to successfully replicate can also be estimated through expert predictions.

MEASUREMENT.

Levels of replication can be measured prospectively. This is done through expert predictions of the replicability of studies (mostly captured as the predicted probability of successful replication), without directly attempting a replication. A percentage or number of studies predicted to successfully replicate can be calculated after dichotomizing this probability (e.g., interpreting probability $> .5$ as prediction of success). While these predictions might be less accurate compared to other measures of replicability, studies would not actually have to be reperformed when only employing this measure.

DATASOURCES

There is no single data source for this metric. Data needs to be extracted from existing publications or gathered newly by employing these methodologies. The studies cited below made their data available, which can be used to re-analyse or extend existing analyses. Note however that Forsell and colleagues promised to populate a repository with the data but have not done so as of June 2023.

EXISTING METHODOLOGIES

The following methodologies, namely surveys and prediction markets, are so far most often used in conjunction with each other. They also have been validated by subsequently conducting full replication studies.

SURVEYS

Experts (mainly researchers) can be asked in surveys about the probability they estimate for specific studies to be successfully reperformed based on information they are given about these studies (e.g., hypothesis, effect size, p-value, link to the original paper, etc.). Surveys have been validated with subsequent replication attempts and compared to prediction markets (see next section). While some studies see many prediction errors and low accuracy when predicting whether the null hypothesis will be rejected ($p < \alpha$) in the same direction (Dreber et al., 2015; Forsell et al., 2019), other findings show better general accuracy of these predictions (Gordon et al., 2021) and better performance predicting relative effect sizes compared to prediction markets (Forsell et al., 2019).

PREDICTION MARKETS

Prediction markets are used for trading bets on a certain outcome. The final market prices can then be taken as an indicator for the probability of an event. In the context of replicability, experts (mainly researchers) are given a budget to bet on studies they think will successfully replicate. The final market price is then a proxy for the probability of successful replication (reaching a previously defined replication criterion) that is estimated by the entire market. Prediction markets have been shown to reach accuracies higher than 70% for their predictions and to outperform surveys (Dreber et al., 2015; Forsell et al., 2019; Gordon et al., 2021). However, prediction markets do not yet appear to be established as a standalone measure for levels of replication.

References

- Baker, M. (2016). 1,500 scientists lift the lid on reproducibility. *Nature*, 533(7604), Article 7604. <https://doi.org/10.1038/533452a>
- Dreber, A., Pfeiffer, T., Almenberg, J., Isaksson, S., Wilson, B., Chen, Y., Nosek, B. A., & Johannesson, M. (2015). Using prediction markets to estimate the reproducibility of scientific research. *Proceedings of the National Academy of Sciences*, 112(50), 15343–15347. <https://doi.org/10.1073/pnas.1516179112>
- Forsell, E., Viganola, D., Pfeiffer, T., Almenberg, J., Wilson, B., Chen, Y., Nosek, B. A., Johannesson, M., & Dreber, A. (2019). Predicting replication outcomes in the Many Labs 2 study. *Journal of Economic Psychology*, 75, 102117. <https://doi.org/10.1016/j.joep.2018.10.009>
- Goodman, S. N., Fanelli, D., & Ioannidis, J. P. A. (2016). What does research reproducibility mean? *Science Translational Medicine*, 8(341). <https://doi.org/10.1126/scitranslmed.aaf5027>
- Gordon, M., Viganola, D., Dreber, A., Johannesson, M., & Pfeiffer, T. (2021). Predicting replicability—Analysis of survey and prediction market data from large-scale forecasting projects. *PLOS ONE*, 16(4), e0248780. <https://doi.org/10.1371/journal.pone.0248780>
- Klein, R. A., Ratliff, K. A., Vianello, M., Adams Jr., R. B., Bahník, Š., Bernstein, M. J., Bocian, K., Brandt, M. J., Brooks, B., Brumbaugh, C. C., Cemalcilar, Z., Chandler, J., Cheong, W., Davis, W. E., Devos, T., Eisner, M., Frankowska, N., Furrow, D., Galliani, E. M., ... Nosek, B. A. (2014). Investigating variation in replicability: A “many labs” replication project. *Social Psychology*, 45(3), 142. <https://doi.org/10.1027/1864-9335/a000178>
- Marino, M. J. (2018). How often should we expect to be wrong? Statistical power, P values, and the expected prevalence of false discoveries. *Biochemical Pharmacology*, 151, 226–233. <https://doi.org/10.1016/j.bcp.2017.12.011>
- Munafò, M. R., Nosek, B. A., Bishop, D. V. M., Button, K. S., Chambers, C. D., Percie du Sert, N., Simonsohn, U., Wagenmakers, E.-J., Ware, J. J., & Ioannidis, J. P. A. (2017). A manifesto for

reproducible science. *Nature Human Behaviour*, 1(1), 0021. <https://doi.org/10.1038/s41562-016-0021>

Nosek, B. A., Hardwicke, T. E., Moshontz, H., Allard, A., Corker, K. S., Dreber, A., Fidler, F., Hilgard, J., Kline Struhl, M., Nuijten, M. B., Rohrer, J. M., Romero, F., Scheel, A. M., Scherer, L. D., Schönbrodt, F. D., & Vazire, S. (2022). Replicability, Robustness, and Reproducibility in Psychological Science. *Annual Review of Psychology*, 73(1), 719–748. <https://doi.org/10.1146/annurev-psych-020821-114157>

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, 349(6251), aac4716. <https://doi.org/10.1126/science.aac4716>

Schmidt, S. (2009). Shall we Really do it Again? The Powerful Concept of Replication is Neglected in the Social Sciences. *Review of General Psychology*, 13(2), 90–100. <https://doi.org/10.1037/a0015108>

Stroebe, W. (2019). What Can We Learn from Many Labs Replications? *Basic and Applied Social Psychology*, 41(2), 91–103. <https://doi.org/10.1080/01973533.2019.1577736>

Polarity of publications

History

Version	Revision date	Revision	Author
1.1	2023-07-18	Revisions	Petros Stavropoulos
1.0	2023-05-11	First Draft	Petros Stavropoulos

Description

The polarity of publications refers to the overall sentiment expressed in research publications through their citations and can be used as an indicator of the scientific community's perception of a certain topic or concept. This indicator aims to measure the degree to which research publications use citations to support, refute, or take no position on a claim, methodology, results, or research output of another publication. The polarity of publications can be used to assess the impact of research on a particular topic, identify potential controversies, and inform future research directions.

The polarity of publications is also useful in assessing the impact of reproducibility efforts in research. For instance, if publications that report on successfully reproduced studies have a more positive polarity in their citations than those that do not, this could indicate that reproducibility efforts have a positive impact on the perception of the scientific community towards a certain topic. Furthermore, if many studies report findings that contradict a particular finding, it might be an indication that this study would not be able to be replicated. Additionally, if there is a trend of negative polarity towards studies that have failed to be reproduced, this could suggest that reproducibility efforts have led to greater scrutiny and higher standards for research quality.

However, it is important to note that polarity itself does not directly measure reproducibility. Rather, it provides insights into the perception and impact of research, including the potential influence of reproducibility efforts. Therefore, while polarity can be indicative of various factors, it should not be solely relied upon as a measure of reproducibility.

Metrics

Number of supporting citations for publications

This metric counts the number of citations in which a publication is cited in a way that supports its claims, methodology, results, or research output. It can be used to determine the level of support a

publication has received from other researchers and can be indicative of the scientific community's perception of the publication.

Limitations of this metric include the potential for biased or incomplete citation practices, as well as the possibility that a publication may receive support from researchers who share similar viewpoints or research interests, rather than from a wider scientific community.

Furthermore, it is important to note that the number of supporting citations can vary widely, depending on the field of study and the specific publication. In some fields, supporting citations are very common, while in others they are relatively rare. This variation can make it difficult to use the number of supporting citations as a reliable indicator of the scientific community's perception of a publication. (Lamers et al., 2021)

In addition to the variation in the number of supporting citations, there are other factors that can affect the interpretation of this metric. For example, the number of supporting citations may be influenced by the age of the publication, the number of citations overall, and the visibility of the publication. It is also important to consider the context in which the citations are made. For example, a citation that is used to support a claim may be different from a citation that is used to mention a publication or to refute a claim.

MEASUREMENT.

To measure the number of supporting citations for a publication, we can search for citations that explicitly mention the publication in a supportive way. This can be done by manually searching for citations, extracting them from the text, and classifying their mentions as "supporting", "refuting" or "neutral". However, this manual process can be time-consuming. Alternatively, automated tools can be leveraged to identify the supporting citation mentions (referred to as 'citances') from the publications text (Budi & Yaniasih, 2022).

One potential limitation of this metric is that it may be difficult to differentiate between citations that provide explicit support for a publication's claims and those that merely mention the publication in passing. In addition, not all citations may explicitly mention the publication's claims, methodology, results or research outputs, and some researchers may support a publication without necessarily citing it.

DATASOURCES

OPENAIRE RESEARCH GRAPH

OpenAIRE Research Graph is a metadata infrastructure that provides a gateway to research publications and their associated data. It is possible to create citation graphs for publications using the OpenAIRE Research Graph by accessing and analysing the metadata provided by the infrastructure.

Using the OpenAIRE Research Graph, it is possible to identify other publications that have cited a publication of interest. The SciNoBo toolkit, which is detailed in the methodologies section, can then be applied to these citations to determine the level of support towards the publication.

There are some limitations to the OpenAIRE Research Graph, such as incomplete or missing metadata, which can affect the accuracy of the citation graphs created. Additionally, the OpenAIRE Research Graph is limited to grant-supported research publications, which may not include all publications in each scientific field.

SCITE.AI

[Scite.ai](#) is a platform that uses natural language processing and machine learning algorithms to identify and classify the citances within a publication as supporting, refuting, or neutral.

Limitations of the platform include:

- Limited coverage of articles analysed by scite.ai
- This is an automated process, so there are limitations based on the underlying model's precision
- This is a paid service

EXISTING METHODOLOGIES

SCINOBO TOOLKIT

The SciNoBo toolkit (Gialitsis et al., 2022) (Kotitsas et al., 2023) has a new component, currently undergoing evaluation, which involves analysing a publication's text and identifying all citances to other publications. It then classifies these citances based on their intent (generic, reuse, comparison), polarity (supporting, refuting, neutral), and semantics (claim, methodology, results, artifact/output).

Limitations of the tool include potential errors in capturing all relevant citances, and correctly classifying their polarity.

References

Budi, I., & Yaniasih, Y. (2022). Understanding the meanings of citations using sentiment, role, and citation function classifications. *Scientometrics*, *128*, 1–25. <https://doi.org/10.1007/s11192-022-04567-4>

Lamers, W. S., Boyack, K., Larivière, V., Sugimoto, C. R., van Eck, N. J., Waltman, L., & Murray, D. (2021). Investigating disagreement in the scientific literature. *ELife*, *10*, e72737. <https://doi.org/10.7554/eLife.72737>

Gialitsis, N., Kotitsas, S., & Papageorgiou, H. (2022). SciNoBo: A Hierarchical Multi-Label Classifier of Scientific Publications. Companion Proceedings of the Web Conference 2022, 800–809. <https://doi.org/10.1145/3487553.3524677>

Kotitsas, S., Pappas, D., Manola, N., & Papageorgiou, H. (2023). SCINOBO: A novel system classifying scholarly communication in a dynamically constructed hierarchical Field-of-Science taxonomy. *Frontiers in Research Metrics and Analytics*, 8. <https://www.frontiersin.org/articles/10.3389/frma.2023.1149834>

Reuse of code in research

History

Version	Revision date	Revision	Author
1.1	2023-07-21	Revisions	Petros Stavropoulos
1.0	2023-05-11	First draft	Petros Stavropoulos

Description

The reuse of code or software in research refers to the practice of utilising existing code or software to develop new research tools, methods, or applications. It is becoming increasingly important in various scientific fields, including computer science, engineering, and data analysis, because it directly contributes to scientific reproducibility by enabling other researchers to validate the findings without the need to recreate the software or tools from scratch. Additionally, it is an indicator of research quality, as repeated use of code or software often signals robustness and reliability. Furthermore, a high percentage of research projects reusing code within a particular field could be an indication of strong collaboration and trust within the scientific community. This indicator aims to capture the extent to which researchers engage in the reuse of code or software in their research by quantifying the number and proportion of studies that utilise existing code or software. The indicator can be used to assess the level of collaboration and sharing of resources within a specific scientific community or field and to identify potential barriers or incentives for the reuse of code or software in research. Additionally, it can serve as a measure of the quality and reliability of research, as the reuse of code or software can increase the transparency, replicability, and scalability of research findings.

Metrics

Number of code/software reused in publications

This metric quantifies the number of times existing code or software has been reused in published research articles. A higher number of instances of code or software reuse in publications suggests a strong culture of code and resources dissemination and building upon existing research within a scientific community or field.

A limitation of this metric is that it may not capture all instances of code or software reuse, as some researchers may reuse code or software without explicitly citing the original source. This challenge is

further exacerbated by the fact that standards of code/software citation are still relatively poor, making the identification of all instances of code/software reuse across research fields problematic. Additionally, this metric may not account for the quality or appropriateness of the reused code or software for the new research questions. Furthermore, it may be challenging to compare the number of instances of code or software reuse in publications across different fields, as some fields may rely more heavily on developing new code or software rather than reusing existing resources.

MEASUREMENT.

An initial step to measure the number of reused code/software in publications can be to count the code/software citations linked with each code/software. This basic strategy, despite being prone to some noise, serves as a fundamental measure for this metric. For a more comprehensive and accurate estimate, we can use tools like text mining and machine learning, including Natural Language Processing (NLP) applied to full texts. These tools help us find code or software reuse statements, or directly pull out datasets from a publication and label them as reused.

However, these methods may face challenges such as inconsistencies in reporting of code or software reuse, variations in the degree of specificity in reporting of the reuse, and difficulties in distinguishing between code or software that is reused versus code or software that is developed anew but shares similarities with existing code or software. Furthermore, the availability and quality of the automated tools may vary across different research fields and may require domain-specific adaptations.

DATASOURCES

OPENAIRE

OpenAIRE is a European Open Science platform that provides access to millions of openly available research publications, datasets, software, and other research outputs. OpenAIRE aggregates content from various sources, including institutional and thematic repositories, data archives, and publishers. This platform provides usage statistics for each research output in the form of downloads, views, and citations, which can be used to measure the impact and reuse of research outputs, including code/software.

To measure the proposed metric, [OpenAIRE Explore](#) can be used to find and access Open Software, study their usage statistics, and identify the research publications that reference them.

However, it's important to note that OpenAIRE Explore does not provide comprehensive data for directly calculating the metric, but rather provides the publication references of each Open Software that need to be analysed.

CZI SOFTWARE MENTIONS

The CZI Software Mentions Dataset (Istrate et al., 2022) is a resource released by the Chan Zuckerberg Initiative (CZI) that provides software mentions extracted from a large corpus of scientific literature. Specifically, the dataset provides access to 67 million software mentions derived from 3.8 million

open-access papers in the biomedical literature from PubMed Central and 16 million full-text papers made available to CZI by various publishers.

A key limitation of this dataset is its focus on biomedical science, meaning it may not provide a comprehensive view of software usage in other scientific disciplines.

To calculate the proposed metric, one could use the CZI Software Mentions Dataset to identify the frequency and distribution of mentions of specific software tools across different scientific papers. The dataset also contains links to software repositories (like PyPI, CRAN, Bioconductor, SciCrunch, and GitHub) which can be used to gather more metadata about the software tools.

EXISTING METHODOLOGIES

SCI NO BO TOOLKIT

The SciNoBo toolkit (Gialitsis et al., 2022) (Kotitsas et al., 2023) has a new component, currently undergoing evaluation, which involves an automated tool, leveraging Deep Learning and Natural Language Processing techniques to identify code/software mentioned in the text of publications and extract metadata associated with them, such as name, version, license, etc. This tool can also classify whether the code/software has been reused by the authors of the publication.

To measure the proposed metric, the tool can be used to identify the reused code/software in the publication texts.

One limitation of this methodology is that it may not capture all instances of code/software reuse if they are not explicitly mentioned in the text of the publication. Additionally, the machine learning algorithms used by the tool may not always accurately classify whether a code/software has been reused and may require manual validation.

DATA SEER.AI

[DataSeer.ai](#) is a platform that utilizes machine learning and Natural Language Processing (NLP) to facilitate the detection and extraction of datasets, methods, and software mentioned in academic papers. The platform can be used to identify instances of software/code reuse within the text of research articles and extract associated metadata.

To measure the proposed metric, DataSeer.ai can scan the body of text in research articles and identify instances of code/software reuse.

However, it is important to note that the ability of DataSeer.ai to determine actual code/software reuse may depend on the explicitness of the authors' writing about their code/software usage, thus not capturing all instances of code/software reuse if they are not explicitly mentioned in the text. Moreover, the machine learning algorithms used by the tool may not always accurately classify whether a code or software has been reused, and may require manual validation.

Number (%) of publications with reused code/software

This metric quantifies the number or percentage of publications that explicitly mention the reuse of existing code or software. It provides an indication of the extent to which researchers are utilizing existing resources to develop new research tools, methods, or applications, within a specific scientific field or task.

A limitation of this metric is that it may not capture all instances of code or software reuse, as some researchers may reuse code or software without explicitly citing the original source. Additionally, it may not account for the quality or appropriateness of the reused code or software for the new research questions. Furthermore, it may be challenging to compare the number or percentage of publications with reused code or software across different fields, as some fields may rely more heavily on developing new code or software rather than reusing existing resources.

MEASUREMENT.

To measure the number or percentage of publications with reused code or software, automatic text mining and machine learning techniques can be used to search for code or software reuse statements, or to identify reused code or software within published research articles, such as the new component of the SciNoBo toolkit.

To measure the percentage of publications with reused code/software, we start by using automatic text mining and/or machine learning techniques to identify whether a publication uses/analyses code or software. This involves searching for keywords and phrases associated with the methodologies and use of code or software within the text of the publications. Next, among the identified publications, we search for code or software reuse statements, or directly extract the code/software from the publications and try to classify them as reused, reporting the percentage of those publications.

However, these methods may face challenges such as inconsistencies in reporting of code or software reuse, variations in the degree of specificity in reporting of the reuse, and difficulties in distinguishing between code or software that is reused versus code or software that is developed anew but shares similarities with existing code or software. Furthermore, the availability and quality of the automated tools may vary across different research fields and may require domain-specific adaptations.

DATASOURCES

OPENAIRE

OpenAIRE is a European Open Science platform that provides access to millions of openly available research publications, datasets, software, and other research outputs. OpenAIRE aggregates content from various sources, including institutional and thematic repositories, data archives, and publishers.

This platform provides usage statistics for each research output in the form of downloads, views, and citations, which can be used to measure the impact and reuse of research outputs, including code/software.

To measure the proposed metric, [OpenAIRE Explore](#) can be used to find and access Open Software, study their usage statistics, and identify the research publications that reference them.

However, it's important to note that OpenAIRE Explore does not provide comprehensive data for directly calculating the metric, but rather provides the publication references of each Open Software that need to be analysed.

CZI SOFTWARE MENTIONS

The CZI Software Mentions Dataset (Istrate et al., 2022) is a resource released by the Chan Zuckerberg Initiative (CZI) that provides software mentions extracted from a large corpus of scientific literature. Specifically, the dataset provides access to 67 million software mentions derived from 3.8 million open-access papers in the biomedical literature from PubMed Central and 16 million full-text papers made available to CZI by various publishers.

A key limitation of this dataset is its focus on biomedical science, meaning it may not provide a comprehensive view of software usage in other scientific disciplines.

To calculate the proposed metric, one could use the CZI Software Mentions Dataset to identify the frequency and distribution of mentions of specific software tools across different scientific papers. The dataset also contains links to software repositories (like PyPI, CRAN, Bioconductor, SciCrunch, and GitHub) which can be used to gather more metadata about the software tools.

EXISTING METHODOLOGIES

SciNoBo TOOLKIT

The SciNoBo toolkit (Gialitsis et al., 2022) (Kotitsas et al., 2023) has a new component, currently undergoing evaluation, which involves an automated tool, leveraging Deep Learning and Natural Language Processing techniques to identify code/software mentioned in the text of publications and extract metadata associated with them, such as name, version, license, etc. This tool can also classify whether the code/software has been reused by the authors of the publication.

To measure the proposed metric, the tool can be used to identify the reused code/software in the publication texts.

One limitation of this methodology is that it may not capture all instances of code/software reuse if they are not explicitly mentioned in the text of the publication. Additionally, the machine learning algorithms used by the tool may not always accurately classify whether code/software has been reused, and may require manual validation.

DATASEER.AI

[DataSeer.ai](#) is a platform that utilizes machine learning and Natural Language Processing (NLP) to facilitate the detection and extraction of datasets, methods, and software mentioned in academic papers. The platform can be used to identify instances of dataset reuse within the text of research articles and extract associated metadata.

To measure the proposed metric, DataSeer.ai can scan the body of text in research articles and identify instances of code/software reuse.

However, it is important to note that DataSeer.ai's ability to determine actual code/software reuse may depend on the explicitness of the authors' writing about their code/software usage, thus not capturing all instances of code/software reuse if they are not explicitly mentioned in the text. Moreover, the machine learning algorithms used by the tool may not always accurately classify whether a code or software has been reused, and may require manual validation.

References

Istrate, A.-M., Li, D., Taraborelli, D., Torkar, M., Veytsman, B., & Williams, I. (2022). *A large dataset of software mentions in the biomedical literature* (arXiv:2209.00693). arXiv. <https://doi.org/10.48550/arXiv.2209.00693>

Gialitsis, N., Kotitsas, S., & Papageorgiou, H. (2022). SciNoBo: A Hierarchical Multi-Label Classifier of Scientific Publications. *Companion Proceedings of the Web Conference 2022*, 800–809. <https://doi.org/10.1145/3487553.3524677>

Kotitsas, S., Pappas, D., Manola, N., & Papageorgiou, H. (2023). SCINOBO: A novel system classifying scholarly communication in a dynamically constructed hierarchical Field-of-Science taxonomy. *Frontiers in Research Metrics and Analytics*, 8. <https://www.frontiersin.org/articles/10.3389/frma.2023.1149834>

Reuse of data in research

History

Version	Revision date	Revision	Author
1.1	2023-20-07	Revisions	Petros Stavropoulos
1.0	2023-05-11	First draft	Petros Stavropoulos

Description

The reuse of data in research refers to the practice of utilizing existing data sets for new research questions. It is a common practice in various scientific fields, and it can lead to increased scientific efficiency, reduced costs, and enhanced scientific collaborations. Additionally, the reuse of well-documented data can serve as an independent verification of original findings, thereby enhancing the reproducibility of research. This indicator aims to capture the extent to which researchers engage in the reuse of data in their research, by quantifying the number and proportion of studies that utilize previously collected data. The indicator can be used to assess the level of scientific collaboration and sharing of data within a specific scientific community or field, and to identify potential barriers or incentives for the reuse of data in research. Additionally, it can serve as a measure of the quality and reliability of research, as the reuse of data can increase the transparency, validity, and replicability of research findings.

Metrics

Number of datasets reused in publications

This metric quantifies the number of datasets that have been reused in published research articles. A higher number of datasets reused in publications suggests a strong culture of data dissemination and building upon existing research within a scientific community or field.

A limitation of this metric is that it may not capture all instances of data reuse, as some researchers may reuse data sets without explicitly citing the original source. This challenge is further exacerbated by the fact that standards of data citation are still relatively poor, making the identification of all instances of data reuse across research fields problematic. Additionally, this metric may not account for the quality or appropriateness of the reused data sets for the new research questions. Furthermore, it may be challenging to compare the number of datasets reused in publications across different fields, as some fields may rely more heavily on new data collection rather than data reuse.

MEASUREMENT.

An initial step to measure the number of reused datasets in publications can be to count the data citations linked with each dataset. This basic strategy, despite being prone to some noise, serves as a fundamental measure for this metric. For a more comprehensive and accurate estimate, we can use tools like text mining and machine learning, including Natural Language Processing (NLP) applied to full texts. These tools help us find data reuse statements, data availability statements, or directly pull out datasets from a publication and label them as reused.

However, these methods may face challenges such as inconsistencies in reporting of reused data, and variations in the degree of specificity in the reporting of the reuse. Additionally, the availability and quality of the data extraction tools may vary across different research fields and may require domain-specific adaptations.

DATASOURCES

OPENAIRE

OpenAIRE is a European Open Science platform that provides access to millions of openly available research publications, datasets, software, and other research outputs. OpenAIRE aggregates content from various sources, including institutional and thematic repositories, data archives, and publishers. This platform provides usage statistics for each research output in the form of downloads, views, and citations, which can be used to measure the impact and reuse of research outputs, including Open Datasets.

To measure the proposed metric, [OpenAIRE Explore](#) can be used to find and access Open Datasets, study their usage statistics, and identify the research publications that reference them.

However, it's important to note that OpenAIRE Explore does not provide comprehensive data for directly calculating the metric, but rather provides the publication references of each Open Software that need to be analysed.

DATA CITE

DataCite is a global registry of research data repositories and datasets, providing persistent identifiers for research data to ensure that they are discoverable, citable, and reusable. The dataset landing pages on DataCite contain information about the dataset, such as metadata, version history, and download statistics.

To measure the proposed metric, we can employ the DataCite REST API to identify relevant datasets, along with to find their DOI, metadata and usage statistics.

EXISTING METHODOLOGIES

SCI NOBO TOOLKIT

The SciNoBo toolkit (Gialitsis et al., 2022) (Kotitsas et al., 2023) has a new component, currently undergoing evaluation, which involves an automated tool, leveraging Deep Learning and Natural Language Processing techniques to identify datasets mentioned in the text of publications and extract metadata associated with them, such as name, version, license, etc. This tool can also classify whether the dataset has been reused by the authors of the publication.

To measure the proposed metric, the tool can be used to identify the reused datasets in the publication texts.

One limitation of this methodology is that it may not capture all instances of dataset reuse if they are not explicitly mentioned in the text of the publication. Additionally, the machine learning algorithms used by the tool may not always accurately classify whether a dataset has been reused and may require manual validation.

DATASEER.AI

[DataSeer.ai](#) is a platform that utilizes machine learning and Natural Language Processing (NLP) to facilitate the detection and extraction of datasets, methods, and software mentioned in academic papers. The platform can be used to identify instances of dataset reuse within the text of research articles and extract associated metadata.

To measure the proposed metric, DataSeer.ai can scan the body of text in research articles and identify instances of dataset reuse.

However, it is important to note that DataSeer.ai's ability to determine actual data reuse may depend on the explicitness of the authors' writing about their data usage, thus not capturing all instances of dataset reuse if they are not explicitly mentioned in the text. Moreover, the machine learning algorithms used by the tool may not always accurately classify whether a dataset has been reused, and may require manual validation.

Number (%) of publications with reused datasets

This metric quantifies the number or percentage of publications that explicitly mention the reuse of previously collected datasets. It is a useful metric for assessing the extent to which researchers are engaging in the reuse of data in their research, within a specific scientific field or task.

A limitation of this metric is that it may not capture all instances of data reuse, as some researchers may reuse data sets without explicitly citing the original source. Additionally, it may not account for the quality or appropriateness of the reused data sets for the new research questions. Furthermore, it may be challenging to compare the number or percentage of publications with reused datasets across different fields, as some fields may rely more heavily on new data collection rather than data reuse.

MEASUREMENT.

To measure the percentage of publications with reused datasets, we start by using automatic text mining and/or machine learning techniques to identify whether a publication uses/analyses data. This involves searching for keywords and phrases associated with data analysis within the text of the publications. Next, among the identified data-analysing publications, we search for data reuse statements, data availability statements, or directly extract the datasets from the publications and try to classify them as reused, reporting the percentage of those publications.

However, these methods may face challenges such as inconsistencies in reporting of reused data, and variations in the degree of specificity in the reporting of the reuse. Additionally, the availability and quality of the data extraction tools may vary across different research fields and may require domain-specific adaptations.

DATASOURCES

OPENAIRE

OpenAIRE is a European Open Science platform that provides access to millions of openly available research publications, datasets, software, and other research outputs. OpenAIRE aggregates content from various sources, including institutional and thematic repositories, data archives, and publishers. This platform provides usage statistics for each research output in the form of downloads, views, and citations, which can be used to measure the impact and reuse of research outputs, including Open Datasets.

To measure the proposed metric, [OpenAIRE Explore](#) can be used to find and access Open Datasets, study their usage statistics, and identify the research publications that reference them.

However, it's important to note that OpenAIRE Explore does not provide comprehensive data for directly calculating the metric, but rather provides the publication references of each Open Software that need to be analysed.

DATA CITE

DataCite is a global registry of research data repositories and datasets, providing persistent identifiers for research data to ensure that they are discoverable, citable, and reusable. The dataset landing pages on DataCite contain information about the dataset, such as metadata, version history, and download statistics.

To measure the proposed metric, we can employ the DataCite REST API to identify relevant datasets, along with to find their DOI, metadata and usage statistics.

EXISTING METHODOLOGIES

SCI NOBO TOOLKIT

The SciNoBo toolkit (Gialitsis et al., 2022) (Kotitsas et al., 2023) has a new component, currently undergoing evaluation, which involves an automated tool, leveraging Deep Learning and Natural Language Processing techniques to identify datasets mentioned in the text of publications and extract metadata associated with them, such as name, version, license, etc. This tool can also classify whether the dataset has been reused by the authors of the publication.

To measure the proposed metric, the tool can be used to identify reused datasets in publication texts.

One limitation of this methodology is that it may not capture all instances of dataset reuse if they are not explicitly mentioned in the text of the publication. Additionally, the machine learning algorithms used by the tool may not always accurately classify whether a dataset has been reused and may require manual validation.

DATASEER.AI

[DataSeer.ai](https://dataseer.ai) is a platform that utilizes machine learning and Natural Language Processing (NLP) to facilitate the detection and extraction of datasets, methods, and software mentioned in academic papers. The platform can be used to identify instances of dataset reuse within the text of research articles and extract associated metadata.

To measure the proposed metric, DataSeer.ai can scan the body of text in research articles and identify instances of dataset reuse.

However, it is important to note that DataSeer.ai's ability to determine actual data reuse may depend on the explicitness of the authors' writing about their data usage, thus not capturing all instances of dataset reuse if they are not explicitly mentioned in the text. Moreover, the machine learning algorithms used by the tool may not always accurately classify whether a dataset has been reused and may require manual validation.

References

Gialitsis, N., Kotitsas, S., & Papageorgiou, H. (2022). SciNoBo: A Hierarchical Multi-Label Classifier of Scientific Publications. Companion Proceedings of the Web Conference 2022, 800–809.

<https://doi.org/10.1145/3487553.3524677>

Kotitsas, S., Pappas, D., Manola, N., & Papageorgiou, H. (2023). SCINOBO: A novel system classifying scholarly communication in a dynamically constructed hierarchical Field-of-Science taxonomy. *Frontiers in Research Metrics and Analytics*, 8.

<https://www.frontiersin.org/articles/10.3389/frma.2023.1149834>