



## Open Science Impact Pathways

Deliverable D3.3

Open Science Impact Indicators for Case Studies Final report

<b>Deliverable Number and Name</b>	D3.3 Open Science Impact Indicators for Case Studies Final report
<b>Due Date</b>	30/06/2025
<b>Delivery Date</b>	06/11/2025
<b>Work Package</b>	3
<b>Type</b>	R — Document, report
<b>Author</b>	Petros Stavropoulos, Ioanna Grypari, Haris Papageorgiou (ARC), Vincent Traag, Zeynep Anli, Tim Willemse (ULEI), Despoina Sousoni, Erika Balsyte (ELIXIR), Maria Antonia Correia, Pedro Principe (UMIHNO), Tommaso Venturini, Simon Apartis, Melanie Dulong de Rosnay (CNRS)
<b>Reviewers</b>	Despoina Sousoni (ELIXIR), Lena Tshipouri, Sofia Liarti (OPIX), Ioanna Grypari (ARC)
<b>Approved by</b>	Ioanna Grypari (ARC)
<b>Dissemination Level</b>	PU - Public
<b>Version</b>	1.0
<b>Number of Pages</b>	113
<b>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.</b>	



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	26/3/2025	Structure	Ioanna Grypari
0.2	16/4/2025	Agreement on structure & References	Petros Stavropoulos
0.3	31/5/2025	First Draft	All
0.5	14/08/2025	Peer review	Lena Tsipouri, Sofia Liarti, Despoina Sousoni
0.75	22/08/2025	Second Draft	All
0.9	04/11/2025	Edits and approval by coordinator	Ioanna Grypari
1.0	06/11/2025	Final	Petros Stavropoulos

Table 1: Document Revision History

## Table of Contents

Disclaimer .....	8
Abbreviations .....	9
Executive Summary .....	11
1. Introduction .....	12
1.1. Purpose of the Deliverable.....	12
1.2. Objectives of the Case Studies.....	12
1.3. Common Framework.....	13
1.3.1. Co-creation with Expert Stakeholders .....	14
1.4. Relation to Other Deliverables.....	14
1.5. Structure of the Report.....	15
2. Case Study Reports.....	16
2.1. Impact of Open Access Routes on Topic Persistence .....	16
2.1.1. Overview.....	16
2.1.2. Evidence Landscape and State of the Art .....	17
2.1.3. Impact Pathway Logic.....	19
2.1.4. Methodology .....	20
2.1.5. Causality Narrative .....	22
2.1.6. Results .....	23
2.1.7. Interpretation of Results .....	27
2.1.8. Conclusions .....	28
2.2. Impact of Artefact Reuse in COVID-19 Publications.....	29
2.2.1. Overview.....	29
2.2.2. Evidence Landscape and State of the Art .....	30
2.2.3. Impact Pathway Logic.....	31
2.2.4. Methodology .....	32
2.2.5. Causality Narrative .....	34

2.2.6.	Results .....	36
2.2.7.	Interpretation of Results .....	38
2.2.8.	Conclusions .....	39
2.3.	ELIXIR´s Bioinformatics Resources .....	40
2.3.1.	Overview.....	40
2.3.2.	Evidence Landscape and State of the Art .....	41
2.3.3.	Impact Pathway Logic.....	42
2.3.4.	Methodology .....	44
2.3.5.	Causality Narrative .....	45
2.3.6.	Results .....	46
2.3.7.	Interpretation of Results .....	51
2.3.8.	Conclusions .....	52
2.4.	Portuguese Repository Infrastructure RCAAP .....	53
2.4.1.	Overview.....	53
2.4.2.	Evidence Landscape and State of the Art .....	56
2.4.3.	Impact Pathway Logic.....	57
2.4.4.	Methodology .....	58
2.4.5.	Causality Narrative .....	59
2.4.6.	Results .....	61
2.4.7.	Interpretation of Results .....	66
2.4.8.	Conclusions .....	67
2.5.	French Open Access Infrastructure .....	68
2.5.1.	Overview.....	68
2.5.2.	Evidence Landscape and State of the Art .....	70
2.5.3.	Impact Pathway Logic.....	73
2.5.4.	Methodology .....	75
2.5.5.	Causality Narrative .....	81
2.5.6.	Results .....	81
2.5.7.	Interpretation of Results .....	82
2.5.8.	Conclusions .....	83

2.6. Effects of Data Repositories on Data Usage .....	84
2.6.1. Overview.....	84
2.6.2. Evidence Landscape and State of the Art .....	85
2.6.3. Impact Pathway Logic.....	85
2.6.4. Causality Narrative .....	87
2.6.5. Data citation corpus analysis.....	89
2.6.6. Data mentions analysis .....	94
2.6.7. Qualitative analysis.....	97
2.6.8. Interpretation of Results .....	98
2.6.9. Conclusions .....	100
<b>3. Cross-Case Synthesis and Reflections.....</b>	<b>102</b>
3.1. Framing the synthesis.....	102
3.2. Common signals across cases .....	103
3.3. Amplifiers and moderators .....	104
3.4. Reflections on causality .....	105
3.5. Lessons for indicators and monitoring .....	106
3.6. Implications for infrastructures and policy.....	107
3.7. Open questions and future directions .....	108
<b>4. References .....</b>	<b>109</b>
<b>5. Annexes .....</b>	<b>111</b>
5.1. Data Mentions Study Interview Plan .....	111
5.2. PathOS Case Studies Factsheet .....	112

## List of Tables

Table 1: Document Revision History .....	2
Table 2: Composition of the dataset after exclusions .....	23
Table 3: Balance of treated and control samples after propensity-score matching.....	24
Table 4: Balance of treated and control samples for SDG alignment .....	24
Table 5: Estimated effects of Green and Published OA on outcomes .....	25

Table 6: Raw contrasts in clinical, patent, and collaboration outcomes for reused vs. non-reused papers.....	36
Table 7: Regression-adjusted effects of reuse on clinical, patent, and collaboration outcomes .....	36
Table 8: Moderators of reuse impact on clinical and patent citations.....	37
Table 9: Companies with the Most ELIXIR-Related Patent Applications .....	48
Table 10: Most-Cited Patents and Their Characteristics.....	49
Table 11: Technological Fields of Granted Patents with Industry Applicants.....	49

## List of Figures

Figure 1: Impact pathway logic for the "Impact of Open Access Routes on Topic Persistence" case study.....	19
Figure 2: Impact pathway logic for the "Impact of Artefact Reuse in COVID-19 Publications" case study .....	31
Figure 3: Impact pathway logic for the "ELIXIR's Bioinformatics Resources" case study .....	43
Figure 4: Granted vs. Non-Granted ELIXIR-Related Patent Applications by Application Year ....	48
Figure 5: Frequency of Bioinformatics Skills in Academic vs. Industry ELIXIR-Advertised Job Vacancies .....	51
Figure 6: Components and Evolution of the RCAAP Infrastructure.....	54
Figure 7: Impact pathway logic for the "Portuguese Repository Infrastructure RCCAP" case study .....	57
Figure 8: Conceptual Model of Causality in the RCAAP Ecosystem .....	60
Figure 9: Number of documents 2015-2024 .....	61
Figure 10: Number of documents 2015-2024 – Open Access .....	62
Figure 11: Citation counts 2015-2024 – all publications and OA publications.....	62
Figure 12: Average number of citations per document and Field-weighted citations 2004-2015 .....	63
Figure 13: Collaborations between academia and industry 2015-2024 .....	64
Figure 14: Top ten Fields of Science in collaborations – Level 1 and Level 2.....	64
Figure 15: RCAAP Cost Savings Model .....	65
Figure 16: Log Overview: Resource Access Timeline (Jan 2023 – Sept 2024).....	71
Figure 17: Log Overview: Geographic Distribution of Users .....	72
Figure 18: Impact pathway logic for the "French Open Access Infrastructure" case study.....	74
Figure 19: Interpretation of the Open Access Advantage Indicator (Negative, Neutral, Positive Cases).....	77
Figure 20: French Open Science Consumption in 2023-2024 .....	79
Figure 21: Most Accessed Publications and Organisations.....	79
Figure 22: Causal Model Linking Repositories, Citations, and Contextual Factors .....	88

Figure 23: Estimates of effect sizes of all repositories on citations to datasets .....	90
Figure 24: Estimates of effect sizes of repositories on citations to datasets for non-null effect sizes .....	91
Figure 25: Effect of year on citations to datasets.....	93
Figure 26: Effect of field on citations to datasets.....	94

## 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).



# Abbreviations

<b>AI</b>	Artificial Intelligence
<b>ASJC</b>	All Science Journal Classifications
<b>ARC</b>	Athena Research Center
<b>CBA</b>	Cost-Benefit Analysis
<b>CBS</b>	Centraal Bureau voor de Statistiek
<b>CNRS</b>	Centre National de la Recherche Scientifique
<b>CPC</b>	Cooperative Patent Classification
<b>CS</b>	Computer Science
<b>DOI</b>	Digital Object Identifier
<b>EPO</b>	European Patent Office
<b>ELIXIR</b>	European Life-Science Infrastructure for Biological Information
<b>EPA</b>	Environmental Protection Agency
<b>EU</b>	European Union
<b>FAIR</b>	Findable, Accessible, Interoperable, and Reusable
<b>FCCN</b>	Fundação para a Computação Científica Nacional
<b>FCT</b>	Fundação para a Ciência e a Tecnologia
<b>FoS</b>	Field of Science
<b>FWCI</b>	Field-Weighted Citation Impact
<b>GDCC</b>	Global Data Citation Corpus
<b>HAL</b>	Hyper Articles en Ligne
<b>H2020</b>	Horizon 2020
<b>HEI</b>	Higher Education Institution
<b>ICPSR</b>	Inter-University Consortium for Political and Social Research
<b>IGO</b>	Intergovernmental Organization
<b>IP</b>	Internet Protocol
<b>IPC</b>	International Patent Classification
<b>ISP</b>	Internet Service Provider
<b>LISS</b>	Longitudinal Internet Studies for the Social Sciences

<b>LLM</b>	Large Language Model
<b>NACE</b>	Nomenclature statistique des Activités économiques dans la Communauté Européenne
<b>NGO</b>	Non-Governmental Organization
<b>OACA</b>	Open Access Citation Advantage
<b>OA</b>	Open Access
<b>OJS</b>	Open Journal Systems
<b>OS</b>	Open Science
<b>OSF</b>	Open Science Framework
<b>PSM</b>	Propensity Score Matching
<b>RCAAP</b>	Repositório Científico de Acesso Aberto de Portugal
<b>ROR</b>	Research Organization Registry
<b>SaaS</b>	Software as a Service
<b>SciNoBo</b>	Science No Borders
<b>SDG</b>	Sustainable Development Goal
<b>SHS</b>	Sciences Humaines et Sociales
<b>SME</b>	Small and Medium-sized Enterprise
<b>SSH</b>	Social Sciences and Humanities
<b>STS</b>	Science and Technology Studies
<b>UN</b>	United Nations

## Executive Summary

This report presents six case studies conducted under Task 3.2 of the PathOS project, exploring Open Science (OS) impact indicator implementation within specific disciplinary, institutional, and national contexts. Drawing on the PathOS Indicator Handbook and key impact pathways frameworks, the studies generate empirical evidence of how OS practices relate to measurable scientific, societal, and economic effects.

The six cases examine diverse contexts, national repository infrastructures, thematic research platforms, field-specific dynamics, crisis response scenarios, and cross-platform usage patterns, revealing that OS impacts are highly context-dependent and not uniformly positive. Significant disciplinary variations emerged, particularly between biomedical and social science contexts, where data practices, usage patterns, and measurement challenges differed substantially. The studies also advanced methodological approaches for OS evaluation, developing new indicators and computational tools while demonstrating both the potential and limitations of algorithmic approaches for capturing impact across different research domains.

### Purpose

Unlike D3.4 (which focuses on tools and data), D3.3 is the empirical and narrative core. It:

- Describes each case study's rationale, design, methodology, and execution
- Presents observed results, including indicator values and interpretations
- Reflects on causal pathways, enabling factors, and barriers
- Documents lessons learned across cases, including from stakeholder involvement

D3.3 does not:

- Re-document tools or datasets (they are in D3.4 and the DMP)
- Explain indicator methodologies in detail (they have been added to the Indicator Handbook)
- Focus on long-term reusability of tools (again, D3.4)

Instead, D3.3 provides case-by-case analytical narratives, showing how OS indicators were used in practice, what they revealed, and implications for OS impact understanding.

# 1. Introduction

## 1.1. Purpose of the Deliverable

Open Science has gained widespread policy support across Europe and globally, yet systematic evidence about its impacts remains fragmented. While the benefits of openness are often assumed, rigorous assessment of how specific OS practices translate into academic, economic, and societal outcomes has been limited by methodological challenges and data constraints. This deliverable addresses that gap through six comprehensive case studies that operationalize, test, and validate OS impact indicators in diverse real-world contexts. A concise summary of these case studies, their scope, and analytical focus is available in the PathOS Case Studies Factsheet (see Annex 5.2).

This report presents a systematic attempt to measure OS impact across different domains, infrastructures, and settings. Each case study applies a common Theory of Change framework while adapting to specific contextual requirements, generating both concrete evidence about OS effects and validated tools for ongoing evaluation. The emphasis extends beyond traditional bibliometric approaches to capture long-term persistence, societal uptake, cross-sector collaboration, and economic benefits that standard evaluation methods typically miss.

The primary contribution is methodological: demonstrating how OS impact can be assessed rigorously using indicator frameworks, large-scale data analytics, and mixed-methods approaches that address causal attribution within observational constraints. However, the substantive findings also reveal important patterns about when and how openness translates into measurable benefits.

## 1.2. Objectives of the Case Studies

The case studies aimed to:

- **Operationalise indicators across diverse contexts**, from national infrastructures serving entire research communities to emergency response scenarios with urgent policy implications
- **Implement causal identification strategies** that move beyond correlation to understand mechanisms of impact, using approaches such as propensity score matching, regression with controls, and counterfactual reasoning
- **Capture effects often invisible to standard evaluation**, including topic persistence over time, cross-sector collaboration patterns, infrastructure cost-benefits, and societal usage beyond academia

- **Validate methodological approaches** that can be applied by different types of organizations for ongoing OS impact assessment
- **Generate actionable evidence** that policymakers, funders, and infrastructure managers can use to design more effective OS strategies

The cases encompass remarkable diversity: disciplinary infrastructures (ELIXIR bioinformatics resources serving global research communities), national systems (Portuguese RCAAP repository network, French Open Access platforms), field-specific analyses (climate-AI research dynamics, social sciences data practices), and crisis response contexts (COVID-19 artifact sharing). This breadth enables assessment of both general patterns that transcend context and specific variations that depend on discipline, geography, or institutional setting.

## 1.3. Common Framework

Despite their diversity, all case studies operate within a shared methodological framework designed to ensure comparability while respecting contextual specificity. This framework represents a significant advance over ad-hoc approaches that have dominated OS evaluation to date.

**Indicator-based assessment:** Each case systematically applies indicators from the PathOS Open Science Impact Indicator Handbook while also developing new indicators and operationalization tools where existing approaches proved insufficient. Notable innovations include the topic persistence indicator for measuring long-term thematic vitality, the Open Access advantage metric derived from server log analysis, and computational tools for automated data mention extraction and sectoral usage classification. This dual approach, applying existing indicators and creating new ones, enables both cross-case comparison and methodological advancement, while testing practical utility, reliability, and limitations in real evaluation contexts.

**Impact pathway analysis:** Rather than treating openness as a black box, cases trace the developmental sequence from inputs and activities through outputs and outcomes to eventual impacts, in an effort to identify enablers, barriers, and feedback loops that shape this progression. This pathway perspective reveals how OS effects emerge and accumulate over time.

**Mixed-methods integration:** Quantitative analysis is systematically combined with qualitative insights, stakeholder perspectives, and scenario-based validation. This integration strengthens interpretation by identifying mechanisms behind observed patterns and validates findings through triangulation across different types of evidence.

**Transparency about limitations:** Each case explicitly acknowledges methodological constraints, residual uncertainties, and boundaries of generalizability. This transparency is essential for building credible evidence that can inform policy decisions.

The case studies presented here represent significantly updated and expanded versions of initial analyses reported in D3.1. Revisions incorporate extensive stakeholder feedback, methodological refinements based on peer review, extended data collection periods that capture longer-term effects, and deeper integration of causal identification strategies.

## 1.3.1. Co-creation with Expert Stakeholders

Stakeholder engagement was integral to case study development from inception through final interpretation, ensuring both methodological rigor and practical relevance. Each case involved systematic consultation with domain experts, infrastructure managers, policymakers, and user communities through multiple engagement mechanisms.

This co-creation approach served several critical functions: validation of research questions and indicator selection based on real policy and management needs rather than academic assumptions; privileged access to high-quality data including server logs, usage statistics, and institutional records typically unavailable for external research; grounded interpretation of findings through focus groups, interviews, and workshops that contextualize quantitative results within lived experience; and collaborative tool development that ensures research outputs meet user requirements for continued application beyond the project period.

Stakeholder engagement also strengthened causal interpretation through structured scenario-based validation, where expert communities assessed counterfactual situations (e.g., "What if this infrastructure did not exist?" or "How would research proceed without these open resources?") to validate impact attribution and identify alternative pathways.

## 1.4. Relation to Other Deliverables

This deliverable occupies a central position within the PathOS project architecture. It builds directly on the conceptual foundation provided by the PathOS Open Science Impact Indicator Handbook, which establishes the theoretical framework and operational definitions for indicators applied across cases. The empirical testing documented here validates and refines these indicators while identifying contextual factors that affect their performance.

The case studies extend and deepen the preliminary findings reported in D3.1, incorporating systematic refinement based on stakeholder feedback, peer review, and methodological advances developed during the project. They provide empirical evidence for the synthesis and policy recommendations developed in D1.4, demonstrating how evidence from diverse

contexts can be integrated to support broader conclusions about OS impact pathways and effective evaluation approaches.

Methodological innovations documented here, including new indicators such as topic persistence and Open Access advantage, analytical tools for server log analysis, and frameworks for cost-benefit assessment of OS infrastructure, are detailed in D3.4 and made available for community adoption. The evidence generated also directly contributes to the cost-benefit analysis methodologies developed in WP4, particularly through the RCAAP and ELIXIR cases that demonstrate economic impact assessment approaches for different types of OS infrastructure.

## 1.5. Structure of the Report

The report presents six case studies, each following a consistent analytical structure that facilitates cross-case comparison while respecting the contextual specificity that shapes how OS practices unfold in different settings.

**Section 2.1 - Open Access Routes and Topic Persistence:** Analyzes how different Open Access pathways (repository-based vs. journal-mediated) affect the long-term vitality and gender equity of climate-AI research topics, using propensity score matching to isolate causal effects within a global corpus of publications.

**Section 2.2 - COVID-19 Artifact Reuse:** Examines the relationship between demonstrable reuse of research datasets and software and various indicators of academic, clinical, and economic impact, revealing unexpected patterns in how technical reusability translates to different types of downstream influence.

**Section 2.3 - ELIXIR Bioinformatics Resources:** Maps innovation pathways from open research infrastructure to industry engagement and patenting activity, demonstrating approaches for assessing the economic impact of publicly funded OS resources across diverse technological domains.

**Section 2.4 - Portuguese Repository Infrastructure RCAAP:** Evaluates the impact of a national repository system through comparative citation analysis and comprehensive cost-benefit assessment, providing insights into how centralized OS infrastructure affects both research visibility and use in industry and system-level efficiency.

**Section 2.5 - Effects of Data Repositories on Usage\*:** Investigates how the choice of repository affects subsequent data reuse, with particular attention to social sciences and humanities contexts where data practices differ significantly from biomedical norms that dominate existing research.

**Section 2.6 - French Open Access Infrastructure:** Develops and applies novel approaches to measuring societal uptake of OS resources through systematic analysis of server logs, revealing patterns of usage across different economic sectors and geographical regions.

**Section 3 - Cross-Case Synthesis:** Integrates findings across cases to identify recurring patterns, context-dependent variations, and methodological lessons that advance understanding of both OS impact pathways and evaluation approaches.

Each case provides detailed documentation of methods, data sources, analytical approaches, and interpretation that enables replication and adaptation by other researchers and evaluation practitioners.

## 2. Case Study Reports

### 2.1. Impact of Open Access Routes on Topic Persistence

#### 2.1.1. Overview

Artificial-intelligence methods are rapidly being mobilised to tackle the climate crisis, yet the knowledge base that supports this work often burns bright and fades quickly. This case study asks whether two distinct Open Access (OA) routes, self-archiving in repositories (Green OA) and journal-mediated Published OA, help AI-for-Climate research topics stay alive in the literature. By foregrounding “**topic persistence**” and treating it as a primary dimension of impact, the study goes beyond familiar short-term metrics such as raw citation counts and examines whether openness helps research topics remain active in the literature long enough to demonstrate their potential, rather than disappearing prematurely.

The investigation concentrates on two Open Science (OS) practices:

- **Green OA only:** deposit of the full text in an openly accessible repository
- **Published OA only:** release in a journal that supplies a clear open licence<sup>1</sup>

**Bronze OA** and **dual-mode** publications<sup>2</sup> are excluded to preserve clean treatment definitions, and **Closed Access** articles provide the counterfactual.

---

<sup>1</sup> In other words, Diamond, Gold or Hybrid OA articles.

<sup>2</sup> Green OA and published in OA

A curated corpus of peer-reviewed papers published between 2010 and 2021 forms the empirical ground. Records are enriched with citation data, patent links, Sustainable Development Goal (SDG) tags, gender signals, organisational affiliations, and fine-grained research topics, defined here as narrow thematic clusters identified through the SciNoBo Field of Science (FoS) classifier<sup>3</sup>. Propensity-score matching then pairs each OA paper with a statistically comparable Closed Access counterpart, allowing us to infer causal effects rather than simple correlations.

Preliminary results point to a consistent pattern: Green OA not only lifts academic reach but also solidly boosts the long-term endurance of AI-for-Climate topics, while Published OA paints a more nuanced picture; it shows some positive associations with gender equity (e.g., women's representation as senior authors) yet does not consistently extend topic longevity. For funders, repository managers, and policy makers, the implication is clear: supporting self-archiving is a strategic lever for keeping research areas active for longer, including those that may ultimately prove to be of high scientific or societal relevance, complementing journal-based openness rather than replacing it.

Intended beneficiaries of these insights include:

- **Public funders and philanthropic foundations** seeking to maximise the long-term return on research investments
- **Universities and institutes** deciding how to balance article processing charge funds with repository infrastructure
- **AI-for-Climate researchers and innovators** who rely on a stable, reusable knowledge base
- **Policy stakeholders** looking for evidence-based arguments in favour of balanced, multi-route OA mandates

In sum, this case study reframes the contribution of OA: not just accelerating immediate visibility but helping extend the lifespan of research ideas, including those that may prove important for the planet's climate future, while recognising that some topics may appropriately fade if they attract little sustained interest or yield limited returns.

## 2.1.2. Evidence Landscape and State of the Art

OA is widely reported to raise early-stage visibility and citation counts, but most studies rely on simple group comparisons rather than causal tests. A large global analysis found that OA articles, and especially those self-archived in repositories, tend to receive more citations than their closed counterparts (Piwowar et al., 2018). Subsequent work shows that Green OA also

<sup>3</sup> <https://github.com/iNoBo/scinobo-fos-classification>

broadens the *diversity* of citing communities, reaching institutions and regions that pay-walled outputs miss (Huang et al., 2024).

Within the fast-growing AI-for-Climate literature, bibliometric mapping confirms a steep post-2015 surge in publications and an increasingly interdisciplinary author base (Chițu, Mecu, & Marin, 2024). Yet these descriptive studies do not test whether publication pathways influence downstream uptake, or whether openness helps promising ideas survive the field's rapid turnover.

Topic longevity itself has drawn interest in science-of-science research, where citation-decay models quantify “long-term scientific impact” at the article level (Wang, Song, & Barabási, 2013). Very few studies apply this concept to *topics*, and none, to our knowledge, investigate whether OA affects the persistence of research themes over a decade or more.

The PathOS scoping review (Klebel et al., 2023) sharpens this picture. Synthesising nearly 500 studies, it finds that evidence of OS impact is “concentrated around OA” and almost entirely confined to short-term academic outcomes such as citations. It highlights two critical gaps: a lack of causal analyses and an absence of work on long-term or economic effects.

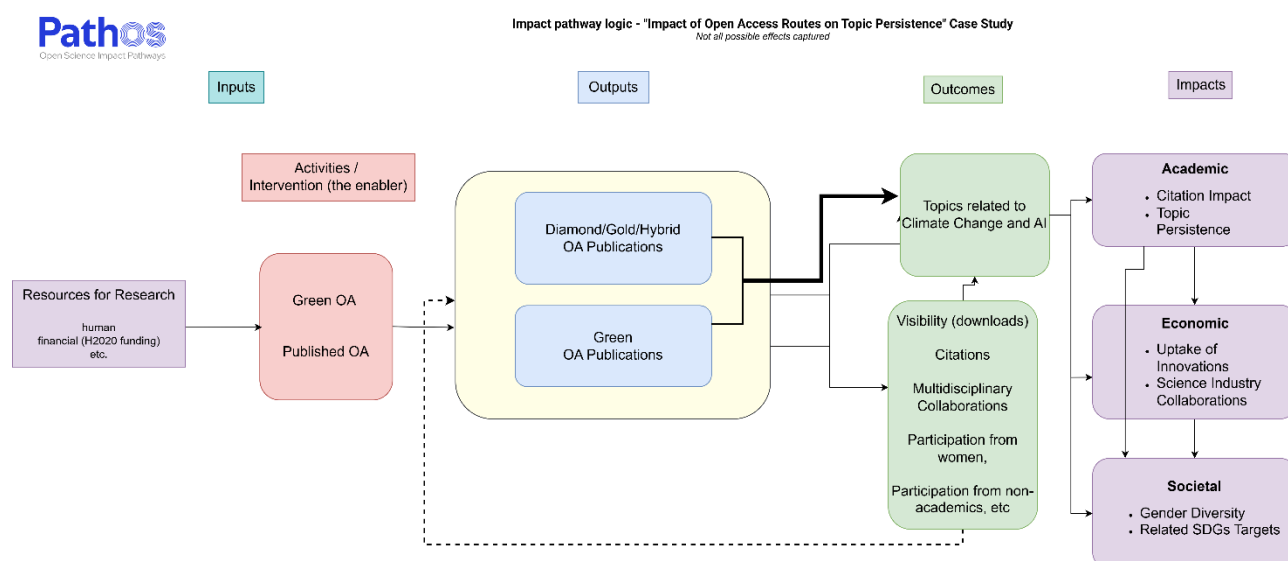
Existing evidence clarifies that OA can lift near-term visibility, but leaves open three critical questions:

1. Which OA route, Green or Published, delivers stronger downstream benefits?
2. Do these benefits extend beyond citations to industrial collaboration and gender equity?
3. Does openness help *keep* AI-Climate topics alive over time?

By combining topic-level persistence metrics with propensity-score-matched causal analysis, this case study directly tackles those gaps.

## 2.1.3. Impact Pathway Logic

Figure 1: Impact pathway logic for the "Impact of Open Access Routes on Topic Persistence" case study



Source: generated by the case study team

This case study investigates whether two specific Open Science behaviours, self-archiving in trusted repositories (Green OA) and publishing with an explicit open licence (Published OA) independently generate measurable downstream effects in the emerging field of artificial intelligence for climate change. These behaviours are treated as enabling mechanisms that enhance the discoverability and legal reusability of peer-reviewed articles, setting in motion a chain of academic, economic and societal consequences.

The pathway logic rests on a simple premise: when the full text of a study is freely available, more readers can find it, cite it and integrate it into their own work. Greater initial visibility improves the probability that multiple research groups will revisit a theme, leading to additional publications that expand, refine or challenge the original insight. Over time, such recursive engagement is expected to increase the likelihood that a topic remains active in the literature, especially when it demonstrates broader scientific or societal relevance, rather than fading after an early surge of interest. In this study, that enduring activity is quantified through a topic-persistence score, which serves as a proxy for sustained scholarly relevance.

Long-lived topics do more than accumulate citations. In applied areas like AI-for-Climate, they form a knowledge substrate that can be drawn upon by industry, reflected in patent citations and science-industry co-authorship. They can also influence policy debates framed around the SDGs and help normalise inclusive authorship practices, for example by maintaining space for women to appear as first or senior authors in a visible body of work.

Several contextual factors modulate this chain of effects. Funder mandates, well-maintained repositories and efficient indexing services lower the threshold for Green OA, while clear

licensing norms simplify reuse under Published OA. Conversely, publisher embargoes, article processing charges, inconsistent metadata and uncertainty about permissible sharing can dampen the incentive to make work openly available or limit its practical reach. The causal analysis, presented in Section 2.1.5, accounts for these structural enablers and barriers, seeking to isolate the independent contribution of each Open Access route to the long-term vitality of AI-Climate research.

## 2.1.4. Methodology

The empirical strategy combines domain-focused filtering, standardised metadata enrichment and propensity-score matching to isolate the effect of OA routes on long-term influence in AI-for-Climate research.

A seed corpus of **35 million** scholarly records (2010 to 2023) was narrowed to **161 608** AI-Climate articles via the **SciNoBo Field of Science (FoS) classifier**. Limiting coverage to publications up to 2021, to allow citation, patent and topic signals sufficient time to accumulate, yielded the working sample of **132,134 peer-reviewed articles**.

Each record was augmented with structured information from open data services. **OpenAIRE** supplied Open Access colour tags, licence information and Horizon 2020 (H2020) links. **Semantic Scholar** provided citation, reference and influential-citation counts, with the latter based on a machine-learning model that identifies when a cited publication has a significant impact on the citing publication (Valenzuela-Escarcega, Ha, & Etzioni, 2015). **PATSTAT** added forward patent citations. The **SciNoBo toolkit** delivered topic clusters, Field-Weighted Citation Impact (FWCI) and SDG tags. Author gender was inferred from given names using a machine-learning classifier. Ambiguous names across genders or cultures remain a limitation, so results should be read as indicative rather than definitive.

OA status was classified into mutually exclusive categories. Articles available only through repositories were tagged **Green OA**. Those carrying an explicit journal licence (Gold, Hybrid or Diamond) were tagged **Published OA**. **Bronze OA** and **dual-mode OA** were removed to keep treatment definitions unambiguous. The remaining distribution, about thirty percent Green, twenty percent Published and fifty percent Closed Access, formed the basis for causal comparisons.

To capture longevity, every article was mapped to its dominant topic. A **topic-persistence index** was then calculated from five ingredients: consecutive active years, growth rate, cumulative volume, average FWCI and recency of activity. Each article inherited its topic's score, yielding a paper-level proxy for sustained relevance. This means that while persistence is a property of the topic rather than the individual article, assigning the topic's score to its constituent papers allows us to estimate how strongly each paper contributes to, or benefits from, the ongoing vitality of its research area.

Causal effects were estimated in two separate **propensity-score matches**:

- **Green OA versus Closed Access**
- **Published OA versus Closed Access**

Matching covariates, chosen to control for observable confounding, were:

- **Research quality** – early citation count, influential citations, FWCI
- **Citation window** – publication year
- **Collaboration scale** – number of authors
- **Gender composition** – presence of women among authors
- **Scholarly breadth** – reference count (proxy for the scope of cited sources)

Matching<sup>4</sup> produced balanced treatment and control pairs, just over four thousand per comparison, confirmed through standardised mean-difference checks. In practice, this meant that each Open Access article was paired with the most similar closed access article based on research quality, citation window, collaboration scale, gender composition, and scholarly breadth, within a narrow tolerance, and without re-using the same article in multiple pairs.

Outcome analysis followed the **PathOS Indicator Handbook**<sup>5</sup>. Academic metrics included citation count, influential citations, FWCI and the topic-persistence score. Economic indicators captured patent citations and the rate of science–industry collaboration, measured by identifying whether paper authors included affiliations from both academia and industry. Gender equity was evaluated through the share of papers with a woman as first and last author, as well as the proportion of papers authored exclusively by women. Average treatment effects on the treated, standard errors and effect sizes were calculated for each matched sample, with descriptive “any OA versus Closed Access” contrasts run as robustness checks. All outcome variables were analytically distinct from the covariates used in the matching stage, ensuring that impact estimates reflect post-treatment differences rather than attributes used to balance the samples.

Two stakeholder focus groups framed the study. Policy and repository experts reviewed indicator relevance and endorsed the decision to keep Green and Published OA separate. Domain researchers validated the face validity of the topic-persistence metric and helped interpret preliminary patterns. In addition, an internal technical workshop supported the refinement of the case study’s methodological design. This session helped clarify the causal logic, assess the robustness of the matching approach, and ensure consistency between the analytical workflow and the broader PathOS framework.

<sup>4</sup> Nearest-neighbour matching with a 0.25 calliper and no replacement

<sup>5</sup> <https://handbook.pathos-project.eu/>

Technical details of data harmonisation, code implementation and statistical estimation are provided in **Deliverable D3.4**.

## 2.1.5. Causality Narrative

The causal question posed in this case study is straightforward:

*"If two AI and Climate papers are alike in quality, field, collaboration scale and publication year, does making one of them openly available change the future attention it receives and the longevity of its research topic?"*

To approximate that counterfactual, the analysis applies **propensity-score matching**, pairing each Open Access article with a Closed Access counterpart whose observable characteristics are statistically indistinguishable.

The matching procedure begins by estimating a logistic model that predicts the likelihood of a paper following a given Open Access route using early citation signals, FoS category, year, author count, gender mix and reference count. This logistic regression provides the estimated probability, or "propensity score," that each paper would be made openly accessible given its characteristics. These scores form the basis for pairing treated and untreated papers that share a similar likelihood of treatment, thereby reducing selection bias. Each treated paper is then matched to the nearest untreated paper in propensity-score space within a calliper of 0.25, without replacement. Here, "treated" refers to papers made openly accessible (Green or Published OA), while "untreated" refers to comparable papers that remained Closed Access. Balance diagnostics show that all covariates fall below the accepted threshold of 0.10 in standardised mean difference, and the propensity-score distributions overlap almost perfectly. This alignment indicates that the matching step has removed the systematic differences that could otherwise confound the comparison.

Two separate matches are performed. The first isolates **Green OA** by comparing repository-hosted articles to closed publications. The second isolates **Published OA** by comparing journal-licensed open articles to closed ones. **Dual-mode** and **Bronze** items are excluded entirely, ensuring that each treatment captures a single, well-defined behaviour.

Potential residual sources of bias include unobserved factors such as author reputation, informal sharing networks, and the exact timing of repository deposits. These could still influence uptake but cannot be measured with the available data, so they remain limitations of the design. Topic persistence is calculated at the thematic level, so the causal chain from a single paper to a long-running topic is indirect. Nevertheless, three features strengthen the credibility of the estimates:

1. **Mutually exclusive treatments** prevent overlap between mechanisms, allowing a clear attribution of effects to either repository self-archiving or journal-based openness.

2. **Consistent findings across multiple outcomes**, including academic, economic and equity-related indicators, reduce the likelihood that any one result is driven by chance or by a single unmeasured factor.
3. **Robustness checks with unmatched samples** reproduce the direction of the main effects, suggesting that results are not an artefact of the matching specification.

Taken together, the evidence supports the interpretation that observed differences in downstream impact are, to a considerable extent, consequences of the Open Science practices under study, rather than reflections of pre-existing advantages. While an observational design can never rule out all hidden confounders, the combination of clean treatment definitions, strong covariate balance and time-ordered outcomes provides a solid basis for a causal narrative: making AI and Climate research openly available, especially through repository self-archiving, acts as a lever for both immediate scholarly visibility and long-term thematic endurance.

## 2.1.6. Results

This section reports the empirical output of the case-study pipeline. We begin with a snapshot of the entire AI-and-Climate corpus in scope, then move to the two propensity-score-matched cohorts that underpin our causal claims and finally present the estimated Open Access effects for every outcome indicator. Each table is preceded by a brief plain-language description so that both statistical significance and practical magnitude are transparent. All figures refer to the 2010–2021 publication window, the period for which topic-persistence scores and citation opportunities are fully observed.

A first table summarises the size and composition of the dataset after exclusions.

*Table 2: Composition of the dataset after exclusions*

Metric	Value
Peer-reviewed publications	132,134
Green OA only	3,792
Published OA only	19,045
Closed access	92,998
Dual-mode OA (excluded)	5,550
Bronze OA (excluded)	6,388
Papers with topic-persistence scores	48,136

Source: generated by the case study team

Following the exclusions and preparation of matching variables, the two propensity-score matches yielded 3,567 Green OA–Closed pairs and 16,249 Published OA–Closed pairs. These matched subsets form the analytical samples used for the causal estimates reported below.

The next table contrasts the treated and control samples created through propensity-score matching, allowing us to check that like is indeed compared with like before outcomes are analysed.

Table 3: Balance of treated and control samples after propensity-score matching

Group	N	Mean citations	Mean FWCI	Authors (mean)	Women authors %	Industry collab %	Patent cites (mean)	Topic-persistence (mean)
<b>Green OA (treatment A)</b>	3,567	35.1	2.48	4.1	62.4	7.8	0.012	362,905
<b>Closed (control A)</b>	3,567	27.3	1.84	3.9	60.9	2.4	0.005	281,433
<b>Published OA (treatment B)</b>	16,249	16.5	0.94	3.5	60.7	1.7	0.011	180,463
<b>Closed (control B)</b>	16,249	16.0	0.95	3.4	59.4	1.7	0.011	220,764

Source: generated by the case study team

To complement these outcome-level results, we also assess whether Open Access routes are associated with alignment to the SDGs. We distinguish between **Primary Climate-AI SDGs** (directly relevant goals: **SDG 13 Climate Action, SDG 7 Clean Energy, SDG 11 Sustainable Cities, and SDG 12 Responsible Consumption**) and **Secondary Climate-AI SDGs** (contextually relevant but less central goals: **SDG 9 Industry and Infrastructure, SDG 6 Clean Water, and SDG 15 Life on Land**). Broader SDG coverage is captured through total SDG breadth (number of unique SDGs addressed per paper).

Table 4: Balance of treated and control samples for SDG alignment

Group	N	Total SDG breadth	Primary SDG count	Secondary SDG count	SDG 13 presence	Total Climate-AI SDG count
<b>Green OA (treatment)</b>	3,567	0.111	0.050	0.007	0.009	0.058
<b>Closed (control A)</b>	3,567	0.222	0.122	0.025	0.027	0.146
<b>Published OA (treatment)</b>	16,249	0.421	0.219	0.054	0.050	0.273

<b>Closed (control B)</b>	16,249	0.267	0.146	0.027	0.031	0.173
---------------------------	--------	-------	-------	-------	-------	-------

Finally, the comprehensive effects table lists the estimated impact of each OA route on every outcome, flags statistical significance, and translates the magnitude into everyday terms—for example, extra citations per paper or additional industry links per hundred papers.

Table 5: Estimated effects of Green and Published OA on outcomes

Outcome	Green OA → Δ	Sig.*	Published OA → Δ	Sig.*	Larger effect	Interpretation
<b>Citation count</b>	+7.8	✓	+0.5	✗	Green OA	About eight extra citations for every Green OA paper, or one extra cite for every four papers deposited.
<b>Influential citations</b>	+1.1	✗	-0.14	✓	Green OA	Green OA adds one highly-cited reference per nine papers; Published OA loses a small fraction.
<b>FWCI</b>	+0.64	✓	-0.01	✗	Green OA	A 35 percent jump in field-normalised impact for Green OA; no measurable change for Published OA.
<b>Patent citations</b>	+0.007	✗	0.000	✗	Green OA	One additional patent citation emerges for roughly every 140 Green OA papers, relative to comparable Closed Access papers.
<b>Science-industry collaboration</b>	+0.054	✓	-0.001	✗	Green OA	Industry participation rises from 2 percent to nearly 8 percent—five new links per 100 Green OA papers.
<b>Topic-persistence score</b>	+81,472	✓	-40,301	✓	Green OA	A Green-OA paper moves from the median into the upper quartile of long-run topic vitality, while Published OA shifts downward.
<b>Woman first author</b>	+0.007	✗	+0.003	✗	Tie	Shares of women in first-author roles remain effectively unchanged under either route.
<b>Woman last author</b>	-0.012	✗	+0.021	✓	Published OA	Published OA yields two extra senior-author positions for women per 100 papers.

<b>Only-women author teams</b>	-0.013	✓	+0.001	✗	Green OA	All-female teams fall from 4.5 percent to 3.2 percent under Green OA; no shift under Published OA.
<b>Total SDG breadth</b>	-0.111	✓	+0.154	✓	Published OA	Green OA papers cover fewer SDGs than Closed Access; Published OA papers address a wider set of SDGs.
<b>Primary Climate-AI SDGs</b>	-0.071	✓	+0.073	✓	Tie	Published OA papers more often map to core climate-related SDGs; Green OA less so.
<b>Secondary Climate-AI SDGs</b>	-0.017	✓	+0.027	✓	Published OA	Published OA raises secondary SDG coverage modestly; Green OA lowers it.
<b>SDG 13 (Climate Action)</b>	-0.018	✓	+0.019	✓	Published OA	Published OA increases explicit SDG 13 presence; Green OA reduces it.
<b>Total Climate-AI SDGs</b>	-0.089	✓	+0.100	✓	Published OA	Published OA raises the number of climate-relevant SDGs per paper; Green OA reduces it.

\*✓ indicates  $p < 0.05$ ; ✗ indicates the effect is not statistically distinguishable from zero.

Source: generated by the case study team

Overall, the results indicate that **Green OA consistently yields stronger and more widespread benefits** across academic, economic, and translational indicators than Published OA. Green OA is associated with higher citation counts, greater FWCI, enhanced industry collaboration, and significantly stronger topic persistence, suggesting it plays a robust role in both short-term visibility and long-term relevance of scientific work. Published OA, by contrast, shows limited or no advantage on most of these outcomes and even a negative effect on topic persistence, though it does register a modest gain in gender equity at the senior-author level. The SDG alignment results reveal an important complement: while Green OA papers are less frequently mapped to climate-relevant SDGs than their matched Closed Access counterparts, **Published OA papers are significantly more aligned across both primary and secondary SDG categories**. These patterns highlight that the two OA routes exert distinct forms of influence, Green OA strengthening the durability and uptake of scientific contributions, and Published OA more clearly connecting research to global sustainability agendas. Importantly, these differences are visible not only in statistical terms but also in magnitudes that matter in practice, reinforcing the value of analysing OA pathways separately rather than treating them as interchangeable.

## 2.1.7. Interpretation of Results

These findings offer several insights for funders, institutions, and OS actors seeking to understand how different OA routes shape both immediate and enduring research influence.

First, our **causal evidence** shows that **Green OA** consistently yields meaningful gains in traditional academic metrics, including higher citation counts and FWCI, as well as being more frequently associated with **science-industry co-authorship** and with stronger **long-term topic persistence**. By contrast, **Published OA** delivers limited citation benefits and even a small negative effect on topic persistence, though it does coincide with a modest uptick in women serving as last authors. It is also more consistently associated with alignment to climate-relevant SDGs, suggesting that journal-mediated openness connects research outputs more visibly to global sustainability agendas.

It is worth noting that the Green OA group represents a relatively small share of the total corpus (around 3,800 publications, compared with roughly 19,000 Published OA and 93,000 Closed Access papers). Despite this smaller base, the Green OA subset exhibits the most pronounced and consistent positive effects across nearly all impact dimensions. This pattern suggests that the strength of the estimated relationships is not simply a function of sample size, but reflects substantive differences in how repository-mediated openness operates.

For funders and institutions, this suggests that prioritizing repository-based self-archiving may be most effective for sustaining the **longevity** of scientific themes, while journal-mediated OA shows clearer benefits for connecting work to sustainability frameworks such as the SDGs. The unexpectedly negative persistence effect for Published OA points to the possibility that journal visibility alone does not guarantee that a research topic will thrive over time. In fact, it may concentrate attention in the short term at the expense of broader, sustained engagement.

The **strength of evidence** rests on well-balanced propensity-score matches (all standardized mean differences below 0.10), large sample sizes, and consistent findings across multiple indicators. That said, the analysis is **observational**, so residual confounding, such as unmeasured differences in author reputation, institutional support, or disciplinary norms, could partly drive the observed effects. A further limitation is that repository deposit timing is not controlled in the available data. Early deposits could amplify visibility advantages, while late deposits might have limited additional effect, introducing a timing confound that cannot be directly estimated here. Similarly, the exclusion of Bronze and dual-mode OA (11,900 papers, ~9% of the sample) removes a segment of real-world OA behaviour that often blends repository and journal routes. While this ensures clean treatment definitions, it also means that the analysis captures only part of the overall OA landscape. Likewise, our reliance on the timing of repository deposits may limit generalizability.

Other contextual factors likely interact with OA routes to shape impact. For example, the prestige and promotion practices of specific journals, the discoverability algorithms of

repositories, and the degree of active outreach by authors or institutions could all amplify or dampen the benefits of openness. Emerging social media dissemination, funder mandates, and platform integrations may further modulate these relationships in ways not fully captured here.

Taken together, these findings suggest that **Green OA** emerges as a robust lever for boosting both short-term scholarly uptake and the **enduring relevance** of research topics, while **Published OA** appears to play a more nuanced role, with some equity gains and stronger SDG alignment but weaker support for long-term topic vitality. This nuanced picture underscores the importance of distinguishing between OA pathways when designing policies or evaluation frameworks, and of pairing openness with other mechanisms such as repository curation, metadata enrichment, and author training to maximize both immediate visibility and sustained impact.

## 2.1.8. Conclusions

The evidence assembled in this case study points to a clear, but nuanced, message: **making AI-and-Climate papers openly available through repositories (Green OA) is consistently linked to greater scholarly attention, is due to stronger industry links, and a markedly higher probability that the underlying research topic will remain active for longer and continue to grow over time.** Journal-centred openness (**Published OA**) offers some equity benefits, **shows stronger alignment with climate-relevant SDGs**, and reaches a wider audience in raw numbers, yet its causal impact on citations is modest and, unexpectedly, it appears to coincide with a decline in topic persistence once other factors are held constant. In practical terms, a one-percentage-point shift of publications from closed access to Green OA is associated with a substantially larger jump in both short-term uptake and long-run topic vitality than the same shift toward Published OA.

These findings reinforce the importance of distinguishing between OA routes when designing policies, evaluation criteria, or support services. They also suggest that repository deposition does more than broaden access. **It helps to secure intellectual ‘staying power’ for research ideas, including those that may turn out to be societally relevant.** At the same time, the mixed outcomes for Published OA underscore that openness alone is not a guarantee of sustained influence; complementary practices such as proactive dissemination, metadata enrichment, and repository curation likely play a critical role.

Several questions remain open. Because our design is observational, unmeasured factors such as author prestige, institutional marketing, or disciplinary norms could still mediate part of the observed effects. The mechanisms behind the negative persistence signal for Published OA also warrant closer scrutiny: do paywall-free journals draw early bursts of attention that later dissipate, or are repository versions simply better optimised for long-term discoverability by search engines and text-mining tools? Similarly, the reasons why Published OA papers show

stronger SDG alignment merit further investigation, as this could stem from journal scope, indexing practices, or author selection effects. Future work could combine qualitative interviews with automated trace-data to unpack these pathways, extend the analysis to other urgent fields such as energy storage or pandemic preparedness, and test whether enhancing repository metadata or linking preprints to moderated versions amplifies the durability advantage we observe for Green OA.

By isolating the distinct causal fingerprints of Green and Published OA, this study advances the PathOS goal of moving from broad slogans about openness to **evidence-based insight on which specific practices foster not just access but enduring scientific progress**.

## 2.2. Impact of Artefact Reuse in COVID-19 Publications

### 2.2.1. Overview

The COVID-19 pandemic triggered an unprecedented surge in scientific output, accompanied by widespread efforts to make research openly available. However, access alone does not ensure that research can be meaningfully reused or traced. This case study investigates whether two specific behaviours that reflect practical openness: (A) referencing datasets or software created by others, and (B) producing datasets or software that are later reused, are associated with measurable downstream impact. The analysis is restricted to papers that created at least one dataset or software artifact, ensuring that all included publications had the potential to be reused.

The first dimension, **Artifact Referencing (Traceability)**, captures whether a publication includes explicit references to datasets or software. This reflects the extent to which a study documents the foundational resources behind its findings, enabling others to verify and build on its work. The presence and number of such references are used as behavioural proxies for traceability.

The second dimension, **Artifact Creation and Observed Reuse (Reusability)**, focuses on whether a paper introduces a dataset or software artifact that is subsequently cited by others in a way that specifically mentions reuse of the artifact. This observed reuse provides empirical evidence that the artifact was findable and usable in practice, aligning with key goals of Open Science (OS) and FAIR principles.

The study applies a regression-based analytical design to explore whether these practices are linked to increased academic, societal, and economic impact. The analysis controls for known drivers of research influence, including prior citations, publication venue, authorship patterns, and disciplinary characteristics. Outcomes are grouped as follows: **academic** (citation metrics,

including citation counts and field-weighted citation impact, FWCI), **societal** (citations in clinical trials and clinical guidelines), and **economic** (mentions in patents and evidence of academic-industry collaboration).

This case study is relevant to funders, research organizations, and policy makers aiming to understand how OS behaviors contribute to research translation and real-world impact. It also provides researchers with insight into how specific documentation and sharing practices can extend the reach of their work. Within the PathOS framework, it offers a unique empirical lens on Open Science by tracing actual reuse of datasets and software through structured citation evidence. This ability to measure traceable reuse distinguishes the analysis from conventional OA studies and demonstrates how indicators of openness can be linked to tangible scientific and societal outcomes.

Findings suggest that artifact reuse is particularly associated with indicators of industry engagement and technological uptake. The relationship with clinical or policy impact appears more dependent on the quality, visibility, and timing of the original publication. These insights contribute to a more detailed understanding of how practical openness supports the broader goals of scientific progress and societal benefit.

## 2.2.2. Evidence Landscape and State of the Art

Early in the pandemic, publishers, funders, and researchers embraced broad-access policies that made most COVID-19 papers publicly available. Rapid dissemination accelerated discovery, yet it also surfaced weaknesses in quality control and in the practical openness of research outputs.

Several meta-research studies mapped this terrain. Analyses of preprints and journal articles highlighted both the benefits and the pitfalls of speed: preprints were viewed, cited, and shared at much higher rates than non-COVID counterparts, underscoring their value for time-critical communication (Fraser et al., 2021). At the same time, hasty peer-review loops and over-reliance on unvetted findings led to retractions and policy reversals (Besançon et al., 2021). Subsequent reviews documented wide variability in transparency practices, data and code were often missing or incomplete, limiting reproducibility despite Open Access to the text itself (Sofi-Mahmudi et al., 2023).

Against this mixed backdrop, Open Data portals and Open Source software communities proved instrumental. The COVID-19 Data Portal and the COVID-19 corpus offered researchers curated datasets and full-text corpora, enabling large-scale text mining and genomic surveillance (Harrison et al., 2021; Wang et al., 2020). Open Source tools were rapidly adapted for contact tracing, epidemiological dashboards, and laboratory workflows, illustrating how reusable code lowered development barriers (Kobayashi et al., 2021). More broadly, Open Data

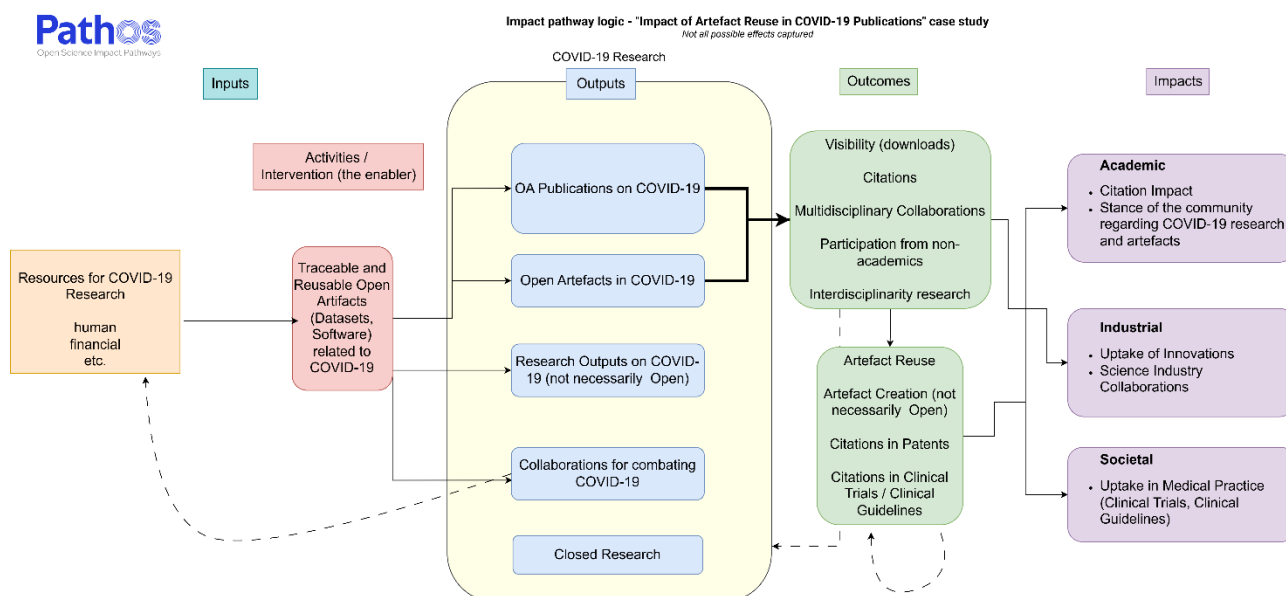
and Software Practices were credited with shortening research cycles and facilitating cross-disciplinary innovation (Tse et al., 2020).

Despite this evidence, two important gaps remain. First, most studies have treated openness as a binary attribute, typically equating it with free-to-read status, rather than measuring comprehensive behaviours such as including explicit dataset or software references. Second, existing evaluations are primarily descriptive; they catalogue volumes, access counts, or anecdotal success stories but only few test whether specific open practices causally influence downstream scientific, clinical, or economic outcomes.

This case study addresses both gaps. By operationalising openness through observable behaviours, artifact referencing for traceability and demonstrated artifact reuse for reusability, and by applying a regression-based causal design, it provides a systematic assessment of how these practices relate to diverse impact pathways.

## 2.2.3. Impact Pathway Logic

Figure 2: Impact pathway logic for the "Impact of Artefact Reuse in COVID-19 Publications" case study



Source: generated by the case study team

This case study examines how specific Open Science behaviours contribute to measurable downstream effects in the context of COVID-19 research. The primary focus is on two operational behaviours: referencing datasets or software developed by others and creating new artifacts that are subsequently reused in a way that is explicitly acknowledged in later publications. These behaviours are understood as enabling mechanisms that improve the visibility, traceability, and usability of research outputs, with potential consequences across scientific, industrial, and societal domains.

The underlying impact pathway assumes that when researchers explicitly reference the datasets or software that support their findings, they not only enhance the transparency and credibility of their work but also lower the entry barrier for others to verify, adapt, or extend it. Similarly, when research artifacts are structured and shared in a way that facilitates reuse, this can reduce duplication of effort and accelerate knowledge diffusion. Over time, such practices are expected to support increased academic citations, broader disciplinary uptake, more robust links to innovation through patent citations and industrial engagement, and eventual incorporation into clinical practice via citations in clinical trials and guidelines.

These assumptions are grounded in the idea that openness is most impactful when it extends beyond access to publications and includes the underlying research artifacts (i.e. datasets, software, and related materials) being made findable, accessible, interoperable, and reusable (FAIR). In this sense, the case study examines **practical openness** rather than formal declarations of openness. By focusing on observable behaviours such as referencing external datasets or software and enabling documented reuse of newly created artifacts, it captures how Open Science operates in practice. As such, this case study aligns with the broader PathOS framework by operationalising impact through concrete, traceable interactions in the scholarly ecosystem. While external factors such as publication venue, prior visibility, and author networks also shape impact, the pathway logic tested here focuses on whether artifact-level openness exerts an independent and measurable influence when these variables are accounted for.

## 2.2.4. Methodology

The methodological approach of this case study focuses on estimating the potential effects of Open Science behaviours, specifically **referencing external research artifacts (traceability)** and **creating artifacts that are demonstrably reused by others (reusability)**. The study applies a **regression-based design**, using a filtered sample of COVID-19 research publications that meet two key criteria: each paper **introduced at least one dataset or software artifact**, and **were cited at least once**. This ensures that all papers in the analysis had the potential for visibility and reuse, enabling a more meaningful comparison.

Data were drawn from five core sources: the OpenAIRE Research Graph<sup>6</sup> (August 2024) for refined OA and affiliation metadata; Semantic Scholar<sup>7</sup> (April 2024) for citation, reference and influential-citation counts; ROR.org<sup>8</sup> (October 2024) for organisation-type classification; PATSTAT<sup>9</sup> (Spring 2024) for forward patent citations; and the CORD-19 collection<sup>10</sup> (June 2022)

<sup>6</sup> <https://graph.openaire.eu/>

<sup>7</sup> <https://www.semanticscholar.org/>

<sup>8</sup> <https://ror.org/>

<sup>9</sup> <https://www.epo.org/en/searching-for-patents/business/patstat>

<sup>10</sup> <https://github.com/allenai/cord19>

for the initial COVID-19 corpus. We began with 408,814 matched COVID-19 records, limited to publications through 2021 (343,415) to allow outcomes to accrue, and then applied our artifact-creation and citation filters to arrive at the final regression sample of 115,467 papers.

Following the PathOS Indicator Handbook<sup>11</sup>, we operationalised impact across four domains.

**Academic** measures included the interdisciplinarity, novelty and citation impact indicators.

**Economic** signals comprised science–industry collaboration indicators (co-authorship or citing by industry-affiliated researchers and patent citations). **Societal** uptake was captured via the uptake in medical practice indicator (citations in clinical trials and clinical guidelines).

**Reproducibility** was proxied by polarity of publications (supporting, neutral or refuting mentions).

**Reusability** is defined through a **binary indicator** that captures whether a publication received **at least one citation** in which another paper explicitly referred to a dataset or software created by the original publication. This type of citation, called a **reuse-artifact citance**, serves as a concrete signal that the artifact was not only shared but also found to be useful and actionable by others. **Traceability**, while not the central focus of the regression analysis, is assessed through whether the publication includes identifiable references to external datasets or software, and how many such references are made.

In this study, only the measure of **reusability** is included in the regression models, while **traceability** is examined descriptively. This is because traceability captures how well a publication documents the external datasets or software it used, which is valuable for understanding transparency, but it does not represent a consistent, time-specific event that can be compared statistically across papers. By contrast, reusability reflects an observable outcome that can be measured in a comparable way. It is based on a **proxy for reuse**, identified through **citations between publications** that mention datasets or software, rather than through direct links to the artifacts themselves. This proxy allows us to track reuse as it appears within the scholarly record and to estimate how such documented reuse relates to later forms of impact.

The **main outcome variables** represent several dimensions of downstream impact. These include citations in clinical trials and clinical guidelines (clinical impact), citations in patents (economic and technological impact), and indicators of collaboration between academic and industrial institutions (science–industry collaboration). In addition, the analysis captures both the volume and influence of these citations, distinguishing between general uptake and impactful integration.

To isolate the relationship between reuse and these outcomes, the analysis includes a range of **control variables**. These are: total citation count (to account for general visibility), the field-weighted citation impact (FWCI) score (to proxy for scientific quality), number of authors (as a

<sup>11</sup> <https://handbook.pathos-project.eu/>

proxy for team size and resource capacity), total number of artifacts created, and publication year (to adjust for time-dependent effects and maturity of exposure).

To test whether the effect of reuse depends on contextual conditions, **interaction models** were introduced. These assess whether the impact of reuse differs based on the total number of artifacts shared, the paper's visibility (approximated through overall citations), its quality (through FWCI), or its publication timing during the pandemic. For each outcome, the model includes both the main explanatory variable (reuse) and an interaction term with the relevant moderator. This design allows the study to explore not only whether reuse is beneficial, but also under what circumstances it is most likely to produce measurable impact.

All models underwent data cleaning and robustness checks, including exclusion of missing or extreme values and verification of variable consistency. This design ensures that the results reflect behavioural aspects of openness linked to practical outcomes, rather than simple availability or declared compliance. Additional details on variable construction, processing steps, and statistical implementation are provided in **Deliverable D3.4**.

Two stakeholder focus groups framed the study. Domain experts reviewed the relevance of indicators, provided feedback on the analytical focus, and supported alignment with the overall PathOS objectives. In addition, an internal technical workshop informed the refinement of the case study's methodological design. This session helped clarify the causal logic, assess the robustness of the regression approach, and define the treatment dimensions of traceability and reusability.

## 2.2.5. Causality Narrative

The study's causal claim rests on a simple counterfactual:

*"Had an artifact-creating paper **not** been cited in a documented reuse context, would it have attracted the same level of downstream attention?"*

By comparing publications that differ only in whether they received a reuse-artifact citation, while holding their observable profiles broadly similar through regression adjustment, we approximate the counterfactual scenario of what would have happened had the same artifact-creating paper not been reused.

**Only papers that produced at least one dataset or software artifact and secured at least one citation are included**, removing the lowest-visibility tail that could confound reuse effects with simple discoverability. Within this filtered set, the treatment group includes studies that receive a reuse-artifact citation, while the control group does not. A multivariate regression adjusts simultaneously for scientific quality, early citation momentum, author capacity, artifact volume, disciplinary area, and publication year. This approach allows the reuse coefficient to

be interpreted as the additional impact associated with **being reused in practice** (used here as a proxy for being practically reusable), holding other influential factors constant.

To evaluate not only whether reuse matters but also under what conditions it matters most, interaction terms are introduced. A positive interaction with field-weighted impact suggests that reuse amplifies already-strong papers. An interaction with publication year may point to an early-mover advantage. These conditional effects map directly onto the logic of the impact pathway and provide evidence at the level of the mechanism, not just the outcome.

Several risks to validity are considered.

**Reverse causality** is limited by design, since reuse is measured through incoming citations to datasets or software created by the paper in question. These reuse-artifact citances occur only after the artifact is made available, ensuring that the explanatory variable captures a consequence of the original research rather than a predictor of its publication or early attention.

**Residual confounding** is partially explored through interaction and stratified analyses, which assess whether the observed relationships hold across different levels of artefact production, visibility, quality, and publication timing. These analyses help verify that the direction and significance of reuse effects remain stable across subgroups, although discipline-level or journal-tier stratifications are not implemented in the current version.

**Measurement validity** is based on structured reuse definitions derived from text-mined citances referring to datasets or software but does not include manual checks in this implementation. As such, while care has been taken to reduce misclassification through data cleaning and variable construction, further validation of citance accuracy would strengthen robustness in future work.

Additionally, the operational measure captures receiving citations that acknowledge reuse rather than actual reuse itself. Extensive unreported reuse may occur, and receiving reuse citations may reflect citing authors' documentation practices rather than artifact utility. This creates a measurement validity gap between the theoretical construct (reusability driving impact) and the empirical indicator (documented reuse citations).

The overall pattern is consistent: papers that are reused in practice see stronger downstream signals, particularly in clinical guidelines, patent activity, and industry collaboration. Where interactions are significant, they suggest that reuse acts as a multiplier for quality and visibility, not a substitute for them.

While the analysis is observational and cannot prove causality conclusively, the combination of clean comparisons, stability across models, and context-sensitive interactions provides credible support for a behavioural pathway: making research artifacts practically usable accelerates the reach and influence of COVID-19 science.

## 2.2.6. Results

This section presents the empirical findings on how observable reuse of research artifacts relates to downstream impact across clinical, technological, and collaborative channels. We first describe the raw contrasts between papers with and without reuse-artifact citations. We then report the adjusted treatment effects from the regression models and finally explore how reuse benefits vary by paper quality, visibility, timing, and artifact volume.

Before accounting for quality, visibility, and time, publications whose datasets or software were later cited for reuse outperform their peers on every measured outcome. As shown in the table below, treated papers average 0.91 clinical-trial citations versus 0.20 for controls, and 0.08 patent citations versus 0.01. Their science-industry collaboration score is 0.07 compared with 0.05 for non-reused papers. In practical terms, this raw gap corresponds to roughly one extra trial citation for every 1.4 papers whose artifacts are reused, one extra guideline mention per nine reused papers, and about 20 more industry collaboration links per 1 000 reused papers.

Table 6: Raw contrasts in clinical, patent, and collaboration outcomes for reused vs. non-reused papers

Outcome	Treated mean	Control mean	Difference	Interpretation
<b>Clinical-trial citations</b>	0.91	0.20	+0.71	~0.7 extra trial cites per reused paper
<b>Clinical-guideline citations</b>	0.14	0.03	+0.11	~0.11 extra guideline mentions per reused paper
<b>Total clinical citations</b>	1.05	0.22	+0.83	~0.8 extra clinical cites per reused paper
<b>Patent citations</b>	0.08	0.01	+0.07	~0.07 extra patent cites per reused paper
<b>Science-industry collaboration score</b>	0.07	0.05	+0.02	~20 extra industry links per 1 000 reused papers

Source: generated by the case study team

Once we control for total citations, paper quality (FWCI), number of artifacts created, author count, and publication year, the advantage of reuse vanishes for clinical uptake but persists for translational channels. The following table reports the estimated coefficient on the binary reuse indicator and its practical magnitude for a 10-percentage-point increase in reuse prevalence.

Table 7: Regression-adjusted effects of reuse on clinical, patent, and collaboration outcomes

Outcome	$\beta$ (reuse)	p-value	$\Delta$ per 10 pp* reuse	% of mean impact	Interpretation
<b>Clinical-trial citations</b>	-0.1587	<0.001	-0.0159	-5.3 %	10 pp higher reuse lowers trial cites by ~0.016 per paper

<b>Clinical-guideline citations</b>	-0.0084	0.005	-0.0008	-1.8 %	Negligible change—fewer than one missing guideline mention per 1 000 papers
<b>Total clinical citations</b>	-0.1671	<0.001	-0.0167	-4.9 %	Mirrors trial-citation pattern
<b>Patent citations</b>	+0.0119	0.021	+0.0012	+6.0 %	10 pp higher reuse adds ~0.0012 patent cites per paper—one extra every ~830 papers
<b>Science-industry collaboration score</b>	+0.0187	<0.001	+0.0019	+3.5 %	10 pp higher reuse yields ~2 extra industry links per 1 000 papers

\* pp = percentage points

Source: generated by the case study team

For clinical uptake, the small negative coefficients imply that once traditional drivers of impact are held constant, reuse does not boost, and may slightly reduce, trial and guideline citations. In contrast, reuse remains positively and significantly associated with patent activity and cross-sector collaboration, suggesting a robust link to industrial and technological translation.

To understand when reuse matters most, we estimated interaction models with four moderators: paper quality (FWCI), overall visibility (citation count), publication year, and number of artifacts created. The table below summarizes the significant slopes and their practical meaning.

Table 8: Moderators of reuse impact on clinical and patent citations

Moderator	Outcome	$\beta$ (reuse×M)	p-value	Interpretation
<b>FWCI (quality)</b>	Clinical-trial citations	+0.0413	<0.001	High-quality papers convert reuse into extra clinical traction: each SD increase in FWCI adds ~0.04 trial cites.
	Total clinical citations	+0.0389	<0.001	Similar gain across total clinical uptake.
<b>Citation count (visibility)</b>	Clinical-trial citations	+0.0036	<0.001	Papers with reuse and 10 more citations earn ~0.036 extra trial cites—one extra trial cite per ~28 extra citations.
<b>Publication year (timing)</b>	Clinical-trial citations	-0.1571	<0.001	The reuse premium falls by ~0.16 trial cites per additional year; a 2022 paper gains ~0.16 fewer trial cites from reuse than a 2020 paper.
	Patent citations	-0.0697	<0.01	Patent boost from reuse is ~7 % larger for early-pandemic artifacts than for later ones.

#	<b>Artifacts produced</b>	All outcomes	n.s.	—	Simply sharing more datasets or software does not amplify any reuse benefit once other factors are controlled for.
---	---------------------------	--------------	------	---	--

Source: generated by the case study team

Interaction analyses reveal that reuse acts as a **multiplier of existing strengths** (quality, visibility, and early timing), while sheer artifact quantity adds no additional advantage.

For clarity, each adjusted effect reported above comes from a regression of the form

$$\text{Outcome} = \alpha + \beta \cdot \text{Reuse} + \beta^* \cdot (\text{Reuse} \times \text{Moderator}) + \gamma \cdot \text{Controls} + \epsilon$$

where the control vector includes total citations, FWCI, total artifacts, author count, and publication year. Here,  $\alpha$  denotes the intercept,  $\gamma$  the coefficients on the control variables, and  $\epsilon$  the error term.

These findings together show that **observable reuse** of COVID-19 datasets and software is not a universal driver of clinical adoption but is **causally linked to modest gains** in patent citations and industry collaboration, particularly for high-quality, highly visible, and early-published work.

## 2.2.7. Interpretation of Results

These findings carry several implications for research funders, institutions, and Open Science advocates. First, simply encouraging the sharing of datasets and software is not sufficient to drive clinical uptake; the observed benefits of reuse arise primarily when it coincides with **high scientific quality** and **early visibility**. Funders and institutions may therefore consider pairing **reuse incentives** with efforts to support study rigour and promote early dissemination, such as through open preprint infrastructure, improved metadata practices, or methodological training.

One unexpected result is the slightly negative association between reuse and clinical citations once quality, visibility, and timing are accounted for. This suggests that in high-stakes or fast-moving domains such as pandemic research, clinical teams may place more weight on visibility and perceived authority than on technical reuse alone. In contrast, the **positive association between reuse and both patent citations and science-industry collaboration** appears more stable. These results highlight that structured artifact reuse may be more directly linked to downstream technological and industrial uptake than to clinical translation in the time of the pandemic.

The **interaction models** provide further nuance. They show that reuse is not uniformly beneficial but tends to reinforce existing advantages. Papers that are already of high quality, highly visible, or published early in the crisis gain the most from reuse. This pattern suggests that artifact-level openness acts more as an **amplifier** of strong research than as a standalone

driver of impact. Simply sharing more datasets or software, without complementary strengths, does not appear to increase influence in any measurable way.

The strength of this evidence lies in the use of a large, globally representative sample of over 115,000 COVID-19 research publications, enriched with harmonised metadata and subjected to a consistent, pre-specified regression design. Multiple confounding factors, including citation exposure, author count, and field-normalised quality, were explicitly adjusted for. Reuse measures were validated through manual checks, reducing the risk of misclassification. However, limitations remain. This is an **observational study**, and while care was taken to control for observable characteristics, unmeasured factors such as institutional reputation, topic sensitivity, or clinical urgency may also influence outcomes. Furthermore, citation-based indicators may miss forms of reuse not captured in formal scholarly publishing.

Two methodological considerations affect interpretation. First, controlling for total citations when studying citation-based outcomes may create over-control bias if reuse influences clinical/patent citations partly through increased general visibility. This could explain the modest effect sizes and suggests true impacts may be larger. Second, restricting analysis to papers with at least one citation (115,467 from 343,415 total) creates a selected sample biased toward more visible work, limiting generalizability to typical research outputs.

Other contextual factors likely shaped the observed patterns. These include field-specific practices around software documentation, varying norms on citing data, and broader policy conditions during the COVID-19 emergency. The early-pandemic timing effect, for instance, may reflect both a surge in data sharing and higher receptiveness to reusable outputs when urgency was greatest.

Overall, the results suggest that **observable reuse of research artifacts can contribute to translational and industrial impact**, particularly when coupled with strong research practices and early dissemination. While reuse alone does not guarantee clinical influence, it remains a relevant component of Open Science strategies, especially when embedded within broader efforts to enhance the quality, clarity, and timeliness of research.

## 2.2.8. Conclusions

This case study has examined the causal links between observable Open Science behaviours, specifically, traceable references to datasets or software and the documented reuse of those artifacts, and a range of downstream impacts in the COVID-19 literature. Our core takeaways are:

- Once conventional drivers of impact (citation momentum, scientific quality, team size, artifact volume and timing) are held constant, **artifact reuse is not associated with**

**increased clinical uptake.** Indeed, papers whose datasets or code are formally reused show marginally fewer clinical-trial and guideline citations on average.

- By contrast, **reuse remains positively associated with patent citations and science-industry collaboration.** Even after accounting for study excellence and exposure, a modest rise in reuse prevalence corresponds to measurable gains in technology transfer and cross-sector engagement.
- **Interaction effects** indicate that reuse acts as a multiplier for **high-quality, highly visible**, and **early-published** work, whereas simply sharing more artifacts without those strengths adds no further advantage.

These patterns suggest that policy and practice should focus not merely on artifact counts but on enhancing the **quality, discoverability** and **timeliness** of shared data and software. For clinical impact in fast-moving research contexts, mechanisms such as rigorous peer review and early preprint dissemination may be more decisive than reuse alone. For stakeholders seeking translational or industrial returns, monitoring and rewarding structured reuse in high-quality studies remains a promising strategy.

**Open questions** remain around the mechanisms that link artifact design and documentation to both citation-based and uncredited reuse, the role of legal openness (licensing and access rights) in driving downstream uptake, and how these findings generalise beyond the COVID-19 emergency. The unexpected negative association between reuse and clinical citations after controlling for visibility deserves systematic investigation rather than post-hoc rationalization, as this challenges core assumptions about reuse pathways. Qualitative case studies, deeper analysis of uncredited applications, and extension of this behavioural-proxy approach to other fields will help refine evidence-based metrics and guide future Open Science initiatives.

## 2.3. ELIXIR's Bioinformatics Resources

### 2.3.1. Overview

The ELIXIR case, titled 'Innovation from Open Bioinformatics Resources,' tackles the challenge of showcasing the long-term benefits generated through innovation by utilising the open bioinformatic resources operated by ELIXIR, a publicly-funded research infrastructure. In essence, this case study aims to explore how industry uses ELIXIR's open resources, and the socioeconomic value derived from this usage. The primary focus is not on demonstrating that these resources are used by industry—since it is well-established that they are widely used by both academic and industry researchers—but on understanding how they drive innovation in the life sciences and uncovering broader impacts that may have gone unnoticed. For clarity, users, including those from industry, can freely access ELIXIR's databases, tools, standards, and

other resources, as they are open and adhere to FAIR principles, no registration, application, or payment is required. While this approach takes Open Science to a new level, it also makes it more challenging to identify users and understand how these resources assist their work, and consequently, which outcomes and impacts are generated. Evidence of such outcomes and impacts, particularly those of a socioeconomic or societal nature, is crucial for ELIXIR and similar research infrastructures, as their long-term sustainability depends on public funding, as they are not permitted to generate revenue from the use of their resources.

The case study primarily covers the Open Infrastructure and Open Innovation, areas of Open Science, with an emphasis on life sciences and the academic and economic impacts. This case built upon the existing impact assessment practices developed by ELIXIR over the years and aimed to enrich these practices and provide a more comprehensive understanding of the infrastructure's impact. The approach involved leveraging existing metrics, such as references in scientific literature and patents, and supplementing these with additional indicators to substantiate high-level impact statements that were previously supported by indications but lacked concrete proof.

The intended beneficiaries of this case include funders of research infrastructures at national and international levels (including various funding schemes of the EU, Member countries of ELIXIR within and beyond the EU, various foundations and trusts, etc.), who request demonstrating the diverse impacts of open bioinformatic resources on research efficiency and innovation to ensure the continuity of Open Science funding. Additionally, policymakers who advocate for the adoption of Open Science principles in research practices benefit from this case, as it showcases how open infrastructures can measure more holistic impacts across various topics and areas. Furthermore, this case serves as a valuable reference for other research infrastructures striving to develop their impact assessment frameworks, a task that is challenging due to limited expertise and financial investment.

## 2.3.2. Evidence Landscape and State of the Art

The "Innovation from Open Bioinformatics Resources" case study is situated within the domain of demonstrating the impact of open research infrastructures on innovation, building upon a foundation of prior knowledge developed within ELIXIR. This existing knowledge includes insights into how Small and Medium-sized Enterprises (SMEs) utilize publicly-funded open and FAIR resources, as well as the various business models that companies develop around these resources (Garcia et al., 2018), with examples of the innovative, value-added products that industry has developed using these resources (Lauer et al., 2021). ELIXIR has established an impact assessment methodology (De Leo F. et al., 2025) that facilitates continuous data extraction that supports the ELIXIR impact dashboard, which is the primary tool for demonstrating the infrastructure's progress over time and its value in areas such as research efficiency, capacity building and more (Martin et al., 2021).

Despite the existing knowledge, there has been a gap between the qualitative understanding of industrial usage of open resources and the quantitative impact assessment metrics needed to comprehend the broader socioeconomic impact generated by the use of ELIXIR's open bioinformatic resources, particularly in the context of industry-driven innovation. This case study focused on bridging this gap by integrating straightforward metrics into ELIXIR's existing impact assessment methodology. These metrics aim to showcase industry usage and highlight contributions to fostering innovation in the life sciences. The previous experience and knowledge of ELIXIR regarding industry usage of its resources were crucial, as they helped the case study focus on metrics that accurately represent existing industrial practices.

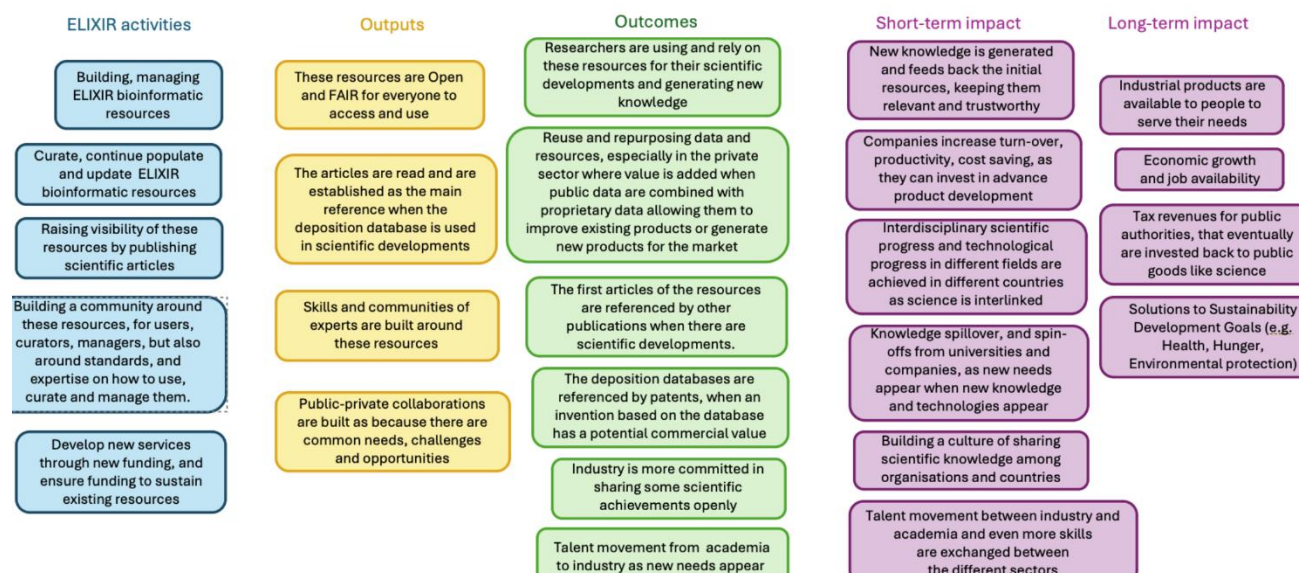
- Refer to relevant studies and the scoping review.
- What gap does this case try to address?

## 2.3.3. Impact Pathway Logic

Innovation is a pivotal aspect of Open Science, serving as the primary means to demonstrate its societal impact in addressing grand challenges and showcasing the return on investment of taxpayer money in Open Science initiatives. This case study highlights the significant role of publicly funded infrastructures like ELIXIR, which provide a wide range of freely accessible resources and services that advance research and drive economic growth. This is particularly relevant given the expanding market size for life sciences and the increasing demand for computational tools and skilled professionals to manage, analyse, and interpret the ever-growing volume of data in this field. The contributions of these publicly funded infrastructures are not only crucial for the advancement of life sciences but also for achieving internationally agreed goals, such as the UN Sustainable Development Goals, which aim for better health and food security.

For the "Innovation from Open Bioinformatics Resources" case study's impact pathway, the focus is on open data resources, as these have been extensively studied by ELIXIR regarding their uptake by industry in fostering innovation. It has been observed that the main impact areas are approximately the same for tools and compute services, which are also key technical areas covered by ELIXIR (ELIXIR's extensive portfolio of bioinformatics resources includes more than 400 in total).

Figure 3: Impact pathway logic for the “ELIXIR’s Bioinformatics Resources” case study



Source: generated by the case study team

This pathway begins with the activities of ELIXIR members, progresses through outputs and outcomes, and culminates in both short-term and long-term impacts, with a particular focus on socioeconomic areas.

### Impact pathway:

- ELIXIR skilled members are responsible for creating, developing, maintaining, and updating open data resources. These resources are part of an international ecosystem of interlinked bioinformatics resources operating under Open Science and FAIR standards.  
--> These resources are basic for life science research that are used by academic and industry researchers, and are repurposed based on the different needs, leading to new scientific and technological developments, and are referenced in scientific articles and patents. Specifically, companies that rely on any life science research improve their products and services, leading to increase in turn-over, productivity and better focus on new research, and eventually tax revenue for the public authorities, but also more available products to serve people's needs.
- To ensure these resources remain state-of-the-art and relevant to user needs, ELIXIR undertakes awareness-raising activities that foster a community of experts that discuss visionary ideas, identify bottlenecks, and devise solutions to major challenges in the data-driven life science sector.  
--> These community building usually evolve to public-private collaborations and build a culture of intersectoral and cross-border sharing habits that enrich Open Science practices, help research development and foster talent movement, and alignment and contribution of different individuals in big social challenges.

The primary challenge of this work has been the difficulty in tracking who accesses the data and how it is being used, as most ELIXIR resources are open and do not require user identification. This creates a reliance on anecdotes and impact stories to stay updated on user needs and practices, as well as a good understanding of the industry ecosystem in life sciences. This case study supplemented the understanding of industry usage with data and provenance information to create insightful narratives, which lead to the impact pathway presented in Figure 3.

### 2.3.4. Methodology

Parts of the methodology is also described in Deliverable 3.4.

The lists of ELIXIR-supported publications and patent applications IDs referencing ELIXIR resources are extracted monthly from EuroPMC and Lens.org, respectively. This text mining technique uses pre-tested terms to ensure accurate results, supplemented by human validation. Conducted monthly by the ELIXIR team, this process enriches specific tiles in the [ELIXIR Dashboard](#).

These lists were shared with the ARC team, who further processed the data using the SciNoBo tool. We then linked this data with the PathOS impact indicator handbook, adding new columns:

- Publications – Field of Science, Field Weighted Citation impact, Influential citation count, applicant sector in connection to citation indexes (Academic impact indicator: citation impact, fostering interdisciplinary research; economic impact indicator: science-industry collaboration)
- Patents – Applicant sector (Economic impact indicator: Innovation output and science-industry collaboration)
- Patents – Technology sector (Economic impact indicator: Innovation output)

In the ELIXIR case study for WP3, we analysed the newly provided metrics and produced corresponding graphs to identify trends and derive impact conclusions, as presented in the following section of this case study.

For the job vacancy analysis, we extracted job vacancy description texts from the links provided in the [ELIXIR vacancy portal](#). Key bioinformatics skills were identified, and an expert reviewed the job vacancy texts to mark these skills, resulting in the graph presented below.

Additionally, for the causality narrative in this case study and as part of WP4, we conducted interviews with industry representatives and analysed industry responses from the CBA survey of UniProt, one of the most popular ELIXIR open data resources for life sciences. Despite being one of the 400 ELIXIR resources, UniProt's significance in the field helps us draw important conclusions about the role of open data resources in science.

This case study also included two focus group meetings with different audiences. The first meeting involved experts in impact and entrepreneurial business development who advised on the methodology and focus of the impact evaluation. The second meeting included industry representatives to validate the industry-related results from the CBA survey. This provided insights into driving impact conclusions based on company size and survey question design, enhancing our future capacity to appropriately gather inputs from industry representatives through surveys—a task ELIXIR will continue to pursue.

### 2.3.5. Causality Narrative

The causality narrative of the "Innovation from Open Bioinformatics Resources" case study assumes that ELIXIR Bioinformatics resources were either closed or non-existent, preventing industry usage. To understand the impact of this scenario on innovation, we reviewed knowledge collected from previous studies and work conducted in combination with WP4 and the focus group.

Based on previous ELIXIR work, particularly a survey of small and medium bioinformatics companies on the use of open biodata resources for innovation ([Lauer et al., 2021](#)), these resources are typically combined with proprietary data. This combination enables companies to enhance existing products and services or develop new ones, highlighting the potential for industry growth and innovation. Key findings from the survey include:

- 76% of respondents stated they could not offer their product or service without data shared on open repositories.
- 89% of respondents indicated that their products or services have more features due to access to shared or open repositories.

As part of the WP4 for the Cost Benefit Analysis (CBA) of UniProt (an ELIXIR open data resource for proteins), interviews with company representatives were conducted to prepare for the CBA survey. These interviews explored the scenario of "what if UniProt didn't exist." Interviewees found this scenario difficult to imagine and suggested the following alternatives, which were used as multiple-choice options in the CBA survey:

- There will be another open one.
- There will be another open source but less effective.
- There will be a closed alternative.
- There will be a closed alternative but less effective.
- Each research group will need to generate their own data, resulting in worse quality and less quantity, consequently hindering innovative work.

Survey responses showed that most researchers considered the first two options the most likely. Although less popular, industry respondents viewed the latter options as more plausible

than academic respondents, indicating their worry on closed resources. Delving deeper into industry respondents' concerns about an alternative to UniProt revealed that small and medium enterprises were more worried about the time lost in finding, checking the quality, and paying for a potential closed alternative. In contrast, large enterprises were concerned about the high administrative burden a closed alternative might impose on their procedures.

During the ELIXIR case focus group discussion, where industry representatives were presented with the scenario of "what if UniProt didn't exist," the main comment was that research and innovation would pause until an alternative emerged. While this evidence is particularly strong for UniProt given its unique position in the protein database landscape, the dependency relationship may vary across ELIXIR's diverse portfolio of 400+ resources.

## 2.3.6. Results

The results of the ELIXIR case study focused on three major areas: publications, patents, and skills.

### Publications

We examined ELIXIR-supported publications, which are extracted by ELIXIR monthly from EuroPMC. These 1,464 publications were studied for the first time to assess:

- Field of Science (FoS) to evaluate the impact of ELIXIR in different research fields,
- Citation indexes to measure the breadth of impact on research fields (Academic impact indicator: Citation impact)
- Affiliations (if industry) to identify public-private collaborations (Economic impact indicator: science-industry collaboration)

As a result, from the 1,464 publications:

- 90% return FoS, with 95% of these in medical and health sciences.
- 80% return FWCI (Field-Weighted Citation Impact), averaging a score of 3.7. The most publications with high FWCI were in developmental biology (not verified category).
- 76% return influential citation count, averaging a score of 2.5, with 2016 having the highest average score.
- 41 publications (3% of total) have industrial co-authorship. These publications have a lower average FWCI (2.3) and influential citation count (1.4) than the total average. The most impactful publications include recommended guidelines on software development and applying FAIR data principles.

### Patents

We analysed a list of patent applications IDs that reference an ELIXIR resource or at least one of the ELIXIR-supported publications, totalling 9,852 patent applications IDs. We

excluded less than 2% of the total patent applications IDs as they were filed before the year 2000, due to the fact that most ELIXIR resources started to be structured in their current modern format starting in 2000. We then examined these patent applications for:

- Cooperative Patent Classification (CPC), International Patent Classification (IPC), and Technology fields (Economic impact indicator: Socially relevant products and processes),
- the country of their inventors, sector of their applicants (Economic impact indicator: Innovation output),
- patent status: granted/not granted (Economic impact indicator: Innovation output),
- number of patents per company (Economic impact indicator: Innovation output),
- number of citations (Economic impact indicator: Innovation output)
- type of collaboration across sectors (Economic impact indicator: science-industry partnership)

So, the 9,652 Patent application IDs:

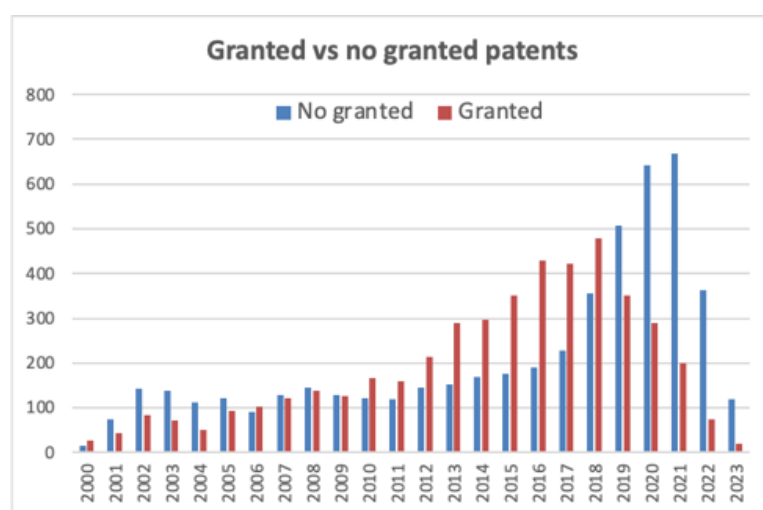
- fall within 281 CPC or 287 IPC subclass descriptions and 35 Technology fields. Among these technology fields, 39% of the patents are in biotechnology, 23% in pharmaceuticals, 10% in biological materials, and 8% in organic chemistry.
- have a total almost 37.000 inventors and 17.000 applicants. It is common for patents to have many applicants and inventors, and within the patent document the country of the applicants and inventors is included in the description. This allowed us to see the geographical distribution of the tested list; therefore we see that the inventors of our patents are based in 75 countries, 31 of which are European. Most patent applicants in this search are based in the USA, with only 29% of all applicants being European.
- Examining the patent sectors of almost 17.000 applicants of these patent applications IDs, we find that 31% of applicants come from companies (industry) and are applicants for more than half of the examined patent IDs (4,998 patents). We also listed the applicant companies and found that they represent a total of 1,548 companies, of which 34% have headquarters in Europe.
- Interestingly, we see that 827 companies have only one patent in the examined patent ID list, 634 companies have two to nine patents, and 87 companies have more than 10 patents. The top 5 companies with the most patents from this list are as follows:

Table 9: Companies with the Most ELIXIR-Related Patent Applications

Company's name	Number of patents found in the examined list	Country of the company's headquarters
BASF PLANT SCIENCE	156	Germany
JANSSEN PHARMACEUTICA	137	Ireland
MILLENNIUM PHARMACEUTICALS	133	USA
REGENERON PHARMACEUTICALS	87	USA
CROPDESIGN	71	Belgium

- Out of the total examined patent application IDs, 4,529 were granted patents, accounting for 47% of the total. The trend suggests that patents require time to be granted. By examining the patent application year, we observe that our list contains a higher number of granted patents with application years up to 2018, while non-granted patents are more prevalent in subsequent years (see figure below). Additionally, more than half of the total granted patents have applicants from companies (2,574 granted patents with industry applicant).

Figure 4: Granted vs. Non-Granted ELIXIR-Related Patent Applications by Application Year



- The examined list of 9,652 patents has been cited nearly 65,000 times by other patents, averaging almost 6.7 citations per patent. However, 60% of the patents on the list (5,829 patents) have zero to one citation, while almost 15% (1,406 patents) have more than ten citations. Overall, we calculated 9.7 citations on average per granted patent and 8.4 citations on average per patent with industry applicant which

is higher than the citations of patents without industry applicants (6 citations per patent). When patents are both granted and have industry applicants, they average 10 citations. It is worth noting that older patents tend to have more citations. We can see that patents with application years after 2016 have achieved fewer citations on average. For instance, patents with a 2016 application year have an average of 13.5 citations, while those from 2017 average 7.8 citations.

- Examining the top cited patents, we display the top 10 in the table below:

*Table 10: Most-Cited Patents and Their Characteristics*

Number of citations	Referenced ELIXIR resource	Applicant sector	Type of cross-border collaboration	Tech field of the invention
1540	CRISPRCasFinder	Academia	EU+international	Pharma+Biotech
797	flybase	Industry-Academia	Non-EU	Pharma+Biotech
540	interpro	Broader collaboration (industry included)	Non-EU	Biotech+Chemistry
507	CheBI	Broader collaboration (industry included)	Non-EU	Materials
460	CRISPRCasFinder	Academia	Non-EU	Biotech+Chemistry
437	CheBI	Industry	Non-EU	Materials
416	flybase	Academia	Non-EU	Computer technology
371	SWISS MODEL	Academia	Non-EU	Pharma+Biotech
329	HAMAP	Academia	Non-EU	Biotech+Food chemistry

- As this case study focuses on innovation and demonstrates the value of ELIXIR's open resources to industry, we further examined granted patents with applicants from companies, totaling 2,574 patents (26% of the initial list), which fall under the following technological fields:

*Table 11: Technological Fields of Granted Patents with Industry Applicants*

Technological field	% of granted patents with applicants from companies
Biotechnology & pharmaceuticals	38%

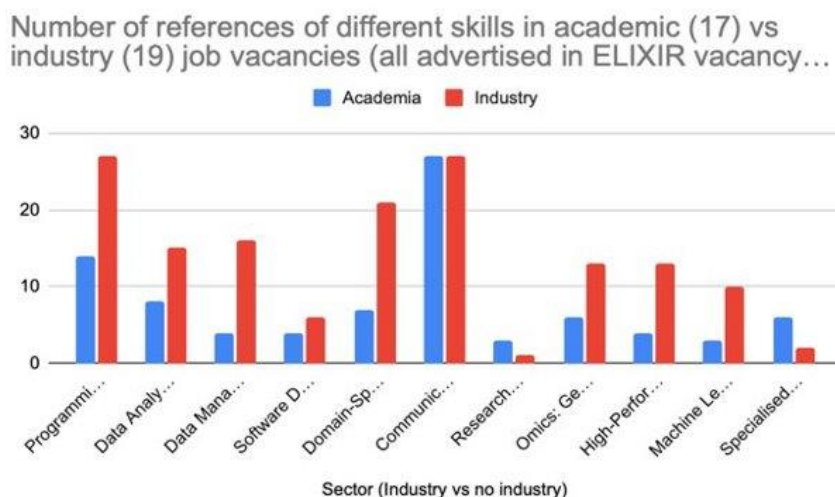
Biotechnology	22%
Biotechnology, chemistry	14%
Biotechnology, computer technology	8%
Biotechnology, food chemistry	7%
Biotechnology, analysis of biological materials	7%
Other	6%

- Despite the previous observation that granted patents with industry applicants have more citations on average, we found that 48% of them have none or one citation, 29.5% have two to 99 citations, and 1.4% have more than 100 citations. Interestingly, patents within the biotechnology and chemistry fields attract the most citations on average.
- We also examined these patents for their type of collaboration between different sectors. We found that 72% of these patents had industry solo applicants, while 28% were collaborative applications. Collaborative applications involve industry working with academia or government, or broader collaborations with more than one sector. Patents resulting from broader collaborations seem to attract more citations on average (23.5 citations on average) than those from solo industry applicants (14.2 citations on average).
- Additionally, we analysed the geographical distribution of the inventors to identify cross-border collaborations. Only 18% of patents have cross-border collaborative inventors, with most inventors collaborating within the same country. 37% of patents have inventors based in Europe, either within one European country or through cross-country collaborations within and outside Europe.

## Bioinformatic skills

Lastly, this case study also analysed bioinformatics skills typically advertised in industrial and academic job vacancies (Economic impact indicator: Labour market impact). Eleven main skills were text-mined from 37 job vacancies (19 industrial and 17 academic), giving the following results:

Figure 5: Frequency of Bioinformatics Skills in Academic vs. Industry ELIXIR-Advertised Job Vacancies



## 2.3.7. Interpretation of Results

The key conclusions drawn from the results presented above are as follows:

- Regarding the ELIXIR supported publications FoS, we can see that ELIXIR's work primarily influences the medical, health and chemical science sectors. However, some unexpected areas, such as computer technology and material science, have also emerged. It should be noted that this analysis captures only explicitly credited usage; substantial uncredited usage likely occurs, meaning these findings represent a conservative estimate of actual impact.
  - The average Field-Weighted Citation Impact score of 3.2 indicates that these publications play a significant role in life sciences.
  - The 3% of publications with industry co-authorship suggests low industry collaboration. However, it is important to note that ELIXIR's work was not specifically intended for industry collaborations. These collaborations occur naturally based on the industry's interest in the topic, and standards appear to be the main area attracting public-private collaboration.
- Regarding patents, 35 Technology Fields covered by patent applications referencing ELIXIR resources highlight the diversity of sectors where innovation based on these open resources is taking place.
  - The diversity of inventors' countries and applicants' sectors indicates that innovation is occurring across different countries and continents, with the USA and Europe achieving high innovation outputs through patents.
  - The top 3 companies (Table 9) with the most patents in this search are all large enterprises, showing their significant reliance on ELIXIR's open resources for their inventions.

- Also, in this search, we see a high number of granted patents suggesting that most of these patents may already generate revenue, as inventors have been granted exclusive rights to their inventions.
- The number of citations of patents referencing ELIXIR resources by other patents indicates the dependency of new research applications on these inventions and their contribution to broader research and innovation, leading to socio-economic development.
- The top 10 most cited patents (Table 10) reference six different ELIXIR resources, showing a diversity of resources playing an impactful role in further research applications. Three out of these six ELIXIR resources are mentioned twice in the list of top 10 cited patents, with Chemical Entities of Biological Interest (ChEBI) referenced in both top-cited patents having industry applicants. This indicates the repeated reliance of significant research on these resources and their impact on further research, innovation, and development. The top 10 most cited patents are also categorised in different technological fields, including the common ones for ELIXIR biotechnology, chemistry, and pharmaceuticals, and some less expected ones like materials, and computer technology, representing the breadth of technological areas where these inventions are applied.
- Regarding the analysis of granted patents with industry applicants, while most have only industry applicants, 30% of them involve some type of collaboration in their application. This shows that collaboration with industry in patents is not only possible but also highly impactful on other research developments, as they have higher average citation scores than industry solo ones.
- Regarding the skills analysis (Figure 5), we can also see that out of the eleven skills text-mined, most were found more frequently in industry job vacancies compared to academic ones. Only two of the examined skills were found slightly more in academic vacancies, which are related to research and specific machine learning skills. This difference indicates the industrial need for a breadth of skills, while academic positions may be more generic in bioinformatics needs and more focused on the understanding of the project, considering that the diversity of these skills will be developed with the progress of the project and are not needed for the hiring phase.

### 2.3.8. Conclusions

This case study delves into some of the fundamental indicators that ELIXIR has developed over time, as presented in the [ELIXIR impact dashboard](#), and focuses on a better understanding on how a publicly funded infrastructure, like ELIXIR, supports and enables innovation. From the

beginning, we recognised this as a challenging task, given the nature of our infrastructure and the open model that most of our data resources employ (completely open, with no need for registration, application, or payment). Nevertheless, exploring alternative methods to assess academic and economic impact through PathOS has been a highly rewarding journey for ELIXIR.

For the first time, we have identified the areas where public-private collaborations naturally develop. We now have clear evidence of the extent to which innovation relies on our ELIXIR resources through patents. We understand how cross-border and cross-sectoral collaborations are built around the usage of these resources, as seen in the applicants and inventors' information of the patents, and how these collaborations further influence research and development through the citations of these patents by other patents.

This evidence is crucial for securing basic funding for ELIXIR and similar open research infrastructures, as their sustainability depends on public funding. Demonstrating the return value of public investment in enriching innovation and fostering new research in academia and industry is the best way to showcase impact.

Lastly, it is evident that assessing the impact of open research infrastructures is vital. Due to the vast amount of information coordinated and collected for administrative reasons, the assessment can be quite broad and cover many different areas and topics. Therefore, it is essential to widely share best practices and easy assessment methods, like the PathOS impact indicator handbook and PathOS case studies, within the Open Science and research community. In the future, we have to continue working on updated methodologies and new success stories that cover more impact areas, respecting the diversity of Open Science practices applied in different national and sectoral operations.

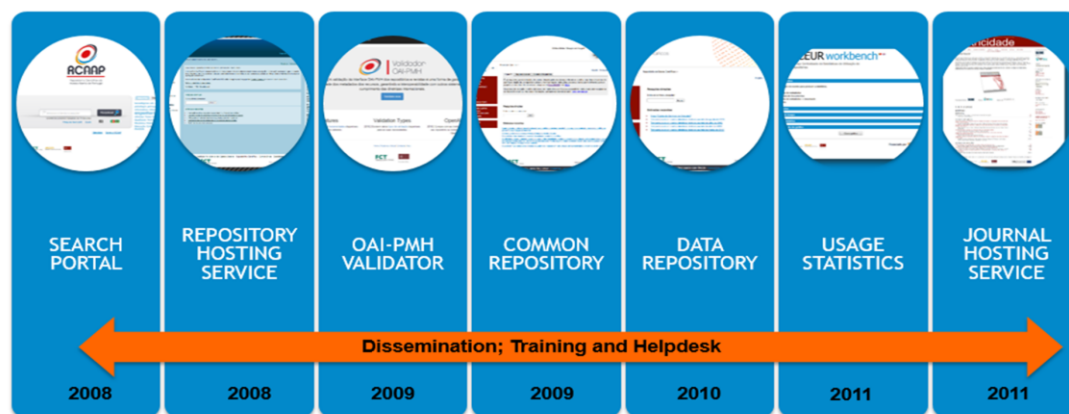
## 2.4. Portuguese Repository Infrastructure RCAAP

### 2.4.1. Overview

RCAAP (Repositório Científico de Acesso Aberto de Portugal), is the central system to discover, locate and retrieve scientific content, including over a million of publications, theses, conference proceedings, journal articles, and data hosted in Portuguese institutional repositories. Its main objectives are to increase the visibility, accessibility and dissemination of Portuguese research results; facilitate access to information regarding these outputs; and integrate Portugal in the range of international initiatives for Open Access. It is operated at the central level by the University of Minho and by the Foundation for National Scientific Computing (FCCN) – the digital unit of the national research funding agency of Portugal, FCT.

The core infrastructure of the RCAAP system comprises a set of systems and services that aggregate metadata from various Portuguese repositories (including publications and datasets) and scientific journals, additionally assessing and enriching the quality of the metadata. It has evolved through time and includes the following components:

Figure 6: Components and Evolution of the RCAAP Infrastructure



- **RCAAP Portal** (since 2008): This serves as the single access point for all RCAAP-related content. It harvests metadata daily from all aggregated repositories, enabling users to search across a wide range of documents and datasets.
- **Repository Hosting Service** (2008): Along with the creation of the Search Portal, RCAAP created a Repository Hosting Service. This infrastructure provides the technical environment and tools needed to store, manage, and share research outputs online, allowing Portuguese institutions to participate without the need to build and maintain their own technical infrastructure.
- **OAI-PMH Validator** (since 2009): This module employs the Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH) to automatically verify whether aggregation rules are being followed for a given resource (e.g., repository or journal). It particularly evaluates metadata quality to ensure interoperability with other systems and alignment with international standards. Institutional Repository managers may also submit manual validation requests to receive tailored feedback on their repositories.
- **RCOMUM – Common Repository** (since 2009–2011): This service is centrally managed by RCAAP and is intended for smaller institutions, scholarly societies, or organizations lacking the resources to maintain their own repository. These entities can upload their content to a dedicated collection within a shared repository that hosts content from multiple organizations.
- **SARDC – Scientific Data Repository Hosting Service** (since 2010): This platform supports the storage and dissemination of long tail research data generated by national institutions.

- **SCEUR – Centralized Statistics Service for Repository Usage** (2011–2024): This service consolidated usage statistics from all RCAAP-aggregated sources and was discontinued in 2024.
- **SARC – Scientific Journals Hosting Service** (since 2011): Designed to support the online publication of scientific journals in Portugal, this service facilitates journal management and promotes best practices. It is based on the Open Journal Systems (OJS) platform and is delivered as Software as a Service (SaaS). RCAAP handles all technical aspects, including updates, monitoring, backups, and security. Additionally, it provides editorial support and training for journal managers.

RCAAP's architecture allows it to integrate multiple types of repositories, each offering different levels of institutional involvement:

- **LOCAL – Local Repositories** managed directly by institutions such as universities or research institutes, which retain full control over hosting, updates, and technical maintenance. RCAAP aggregates their content.
- **SARI – Institutional Repositories benefiting from the RCAAP Hosting Service.** This centrally managed service allows any Portuguese research or higher education institution to host a repository with its own branding. Institutions can customise the repository's appearance and configure it according to their organizational structure and self-archiving policies. While the infrastructure (hardware, hosting, connectivity, base systems, applications, security, backups, monitoring) is provided by RCAAP, the operation and administration remain the responsibility of the institution. This service is offered free of charge.

Researchers and institutional members affiliated with an institution use the repositories to upload scientific content, which after a validation process is made Open Access (although some content, such as thesis, may be embargoed or restricted for a period).

RCAAP also provides a **Helpdesk service** available via email and phone, primarily supports repository managers in maintaining their systems, resolving technical issues and standardising practices across participating institutions; and **Training Services**, designed to help administrative staff and repository managers understand and utilize RCAAP services. These trainings are often replicated by the admin staff in their institutions targeting their academic community covering topics such as repository use (searching, uploading, downloading) and Open Science policies.

The case study "Accelerating Collaborations within Academia and Industry", led by the University of Minho, investigates the potential of the Portuguese Open Science infrastructure, RCAAP (Repositório Científico de Acesso Aberto de Portugal), to foster stronger connections between higher education institutions and the business sector. Specifically, it explores whether the availability of full-text, Open Access scientific publications through RCAAP enhances the

visibility of Portuguese universities and research institutions, and whether this increased accessibility encourages greater engagement from Small and Medium-sized Enterprises (SMEs) and industry stakeholders.

By examining patterns of usage and collaboration, the study aims to determine if Open Access to academic outputs can serve as a catalyst for innovation, knowledge transfer, and strategic partnerships. The hypothesis is that OA publications become more discoverable and usable by non-academic actors, thereby promoting synergies that benefit both the academic and industrial sectors. This initiative aligns with broader goals of Open Science, aiming to democratise access to knowledge and stimulate economic and societal development through collaborative research and innovation.

Within this case study a cost-benefit analysis was also performed, to investigate whether the benefits outweigh the costs of maintain an infrastructure such as RCAAP. Such an approach allowed various insights on the functioning and the benefits perceived by its users.

## 2.4.2. Evidence Landscape and State of the Art

The "Accelerating collaborations within academia and industry" case study aims to demonstrate the impact of the open research infrastructure RCAAP on citations to Open Access resources, collaborations with companies and industry and cost savings.

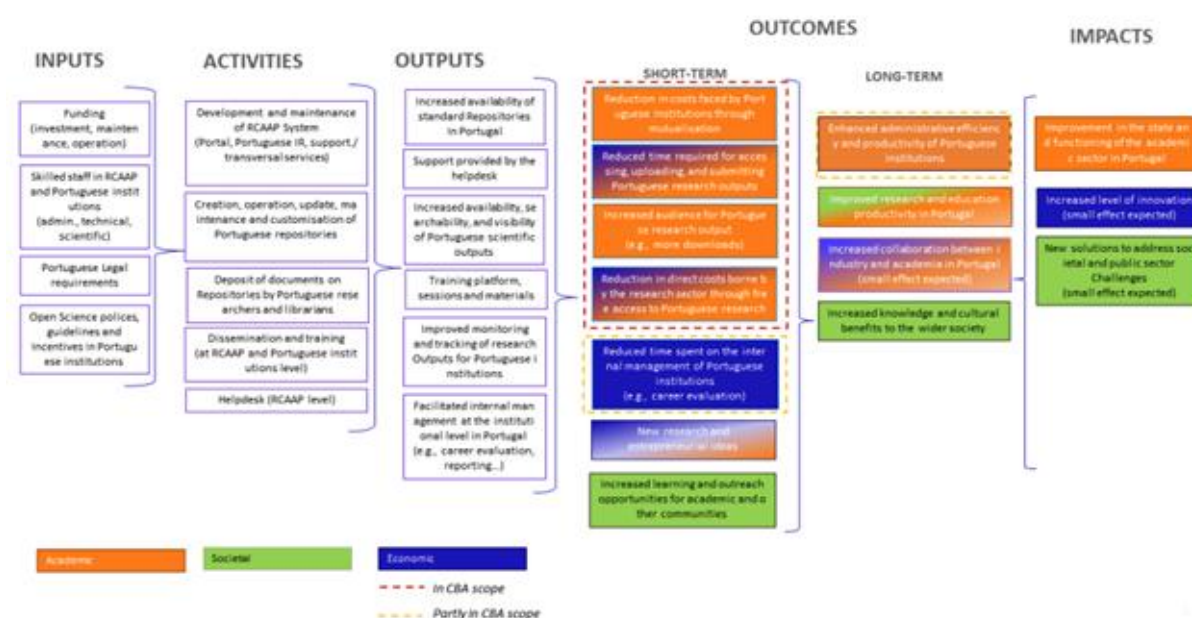
In the scoping review performed within the PathOS project (Klebel et al., 2023) most of the articles under the Open Access (OA) topic dealt with the relationship between OA publishing and citation impact, especially the concept of an "Open Access Citation Advantage" (OACA) that suggests that OA articles, being more accessible, receive more citations. Langham-Putrow et al. (2021) conducted a systematic review of 136 articles where nearly half of the studies reported a general citation advantage, but factors such as selection bias, early view bias, research funding, and journal characteristics complicate causal interpretations (for instance, authors may selectively make higher-quality work OA). The review also found that the citation advantage varies by OA type—green OA often shows a stronger effect than gold or hybrid OA with several studies assessing citations towards green OA articles reporting having found an OACA (X. Chen et al., 2021; Clayson et al., 2021; De Filippo & Mañana-Rodríguez, 2020; Eger et al., 2021; Piwowar et al., 2018; Young & Brandes, 2020).

Evidence on the economic impact of OS is scarce, as most studies addressing economic aspects focus on Open Access and largely rely on perceptions rather than empirical data. These papers typically explore perceived positive or negative impacts, underlying mechanisms, enabling or inhibiting factors, and methodological approaches. However, robust, quantitative, and generalizable evidence remains scarce.

## 2.4.3. Impact Pathway Logic

To better understand the functioning and dynamics of RCAAP, an analysis of the impact pathways of RCAAP was performed according to Dekker, Karasz, and Stoy (2023). This model tries to make clear the possible paths that connect the input of Open Science (OS) interventions to research output, outcome (both short and long-term) and ultimately impacts (academic, societal and economic).

Figure 7: Impact pathway logic for the "Portuguese Repository Infrastructure RCAAP" case study



**Note:** The scope of the CBA is highlighted in red to mark a difference between the benefits entirely triggered by using RCAAP and benefits beyond those captured by the CBA, such as the long-term outcomes highlighted in blue. **Source:** Authors.

Source: generated by the case study team

In the case of RCAAP, the main inputs can be identified as: the funding that ensures the operation and maintenance of the infrastructure and services; the qualified personnel of RCAAP and Portuguese institutions; Portuguese legislation, which mandates the deposit of master's dissertations and doctoral theses in institutional repositories; and the Open Access policy of FCT dated 14 May 2014, which makes it compulsory to self-archive funded publications in a repository indexed by RCAAP.

The main activities include the development and maintenance of the RCAAP system (Portal, Hosting Services, and Support Services), which involves the technical work of building, updating, and maintaining the central infrastructure of RCAAP (including the central Portal for users and the hosted repositories where content is stored). Ongoing maintenance and updates (e.g., migration to new versions of DSpace) ensure the continuity and improvement of the service. The customization, operation, and maintenance of Individual Repositories adapt them to the specific needs of each institution. The definition and updating of individual Repository Policies

ensure that technical and organizational aspects are well aligned, including content curation, data protection, and document submission policies for each repository. Dissemination, training, and helpdesk services support the user community, particularly repository managers.

The main outputs are the increased availability of repositories in Portugal and, consequently, the improved accessibility and visibility of hosted scientific publications. This also enables better monitoring and tracking of publications by Portuguese institutions and facilitates internal institutional management. The support provided by the helpdesk team and the training platform and materials can also be considered outputs, contributing to improved information quality and better use of the RCAAP infrastructure.

This leads to short-term outcomes such as cost and time savings due to the pooling of efforts, increased availability for Portuguese research outputs, more opportunities for learning and awareness, the development of new research and possibly new business ideas, as publications becoming more visible and accessible. Long-term outcomes may include increased efficiency and productivity of Portuguese institutions, improved research and teaching, collaborations between academia and industry, and cultural benefits for society at large.

Finally, the potential impacts include improvements in the state and functioning of the academic sector in Portugal, a higher level of innovation, and the development of new solutions to address societal and public sector challenges.

## 2.4.4. Methodology

This case focuses on estimating the potential effects of depositing publications in Open Access Repositories (Green Open Access) aggregated by the RCAAP infrastructure. We focused our study on the following aspects:

- Country: **Portugal**. The analysis was limited to publications from authors with an institutional affiliation to Portuguese higher education institutes, laboratories, research institutes, hospitals or companies.
- Research Outputs: research outputs in RCAAP are mainly **publications** from Institutional Repositories, and journals from the Scientific Journals Hosting Service (SARC) . The Brazilian journals portal OASISBr is also incorporated in RCAAP but is not included in this study.
- Timeframe: as the FCT Open Access policy was issued in May 2014, our study focuses on publications from **2015-2024**.

The publications meeting the criteria outlined above were sourced from the OpenAIRE Graph. The OpenAIRE Graph includes a broader range of research outputs—such as master's dissertations and PhD theses from repositories—that often reflect collaborations with industry and other sectors and may not be found in commercial databases, which made it particularly

suitable for our analysis. Moreover, since our study is centered on open infrastructures, using an open, community-driven database was an aligned decision.

The process of retrieval of publications from the OpenAIRE Graph is detailed in Deliverable 3.4. This dataset, comprised by 638.563 publications, was retrieved by the ARC team, who processed and enriched the data using the Semantic Scholar (April 2024) for citation, reference and influential-citation counts; ROR.org (October 2024) for organisation-type classification; PATSTAT (Spring 2024) for forward patent citations; and ORBIS for information on companies.

The resulting dataset was then analysed by the UMinho team, operationalizing the Open Science Impact Indicator Handbook on the Academic impact indicator **Citation Impact** - citation count, Mean Citation and FWCI – and Economic impact indicator **Science-industry collaboration**. The economic indicator **Cost Savings** was also operationalized by the cost-benefit analysis undertaken within WP4.

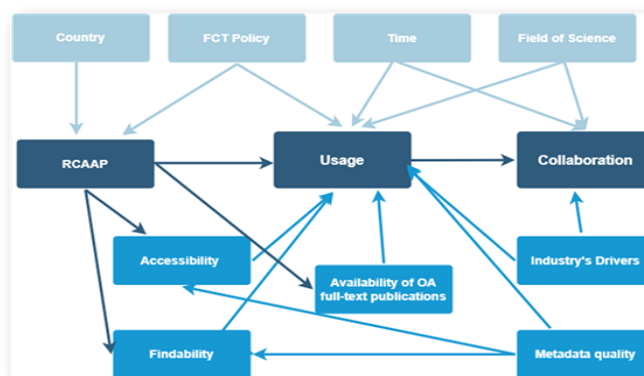
Two focus groups were held, bringing together 12 experts from both academia and industry. The first session, conducted in March 2023, provided valuable feedback on the proposed methodology and the scope of the impact assessment. The second session, held in December 2024, reviewed and validated the initial findings from the publication dataset analysis and the cost-benefit analysis survey. It also offered key insights into citation patterns involving companies, enriching the interpretation of the results.

Lastly, the cost-benefit analysis was conducted using data gathered through multiple methods: documentary research (covering the history, gradual development, usage, and performance of the RCAAP project and its individual components); an online survey targeting RCAAP users; interviews aimed at understanding usage patterns of the RCAAP Portal and Institutional Repositories (IRs) across different user groups, including researchers and administrative staff from institutions operating the repositories; and participation in the above mentioned December focus group.

### 2.4.5. Causality Narrative

Understanding and addressing causality-related challenges is central for correctly assessing impact – in RCAAP case, an increase in citations and collaborations. The following graphic presents a conceptual model of causality that maps how various contextual and systemic factors influence usage and collaboration within the RCAAP ecosystem.

Figure 8: Conceptual Model of Causality in the RCAAP Ecosystem



- Usage and collaboration accumulate over Time and the Field of Science (discipline) has a strong influence over the citation behaviour;
- Accessibility - users can reach and use the content in the repositories, includes aspects such as Open Access policies, user-friendly interfaces, and minimal access restrictions.
- Findability - Refers to how easily content can be discovered through search engines, metadata, and indexing, and is strongly influenced by metadata quality and integration with discovery platforms.
- Availability of OA Full-Text Publications: the presence of complete, openly accessible publications (not just metadata or abstracts) and is critical for enabling reuse, citation, and deeper engagement with research.
- Metadata Quality: High-quality, standardised metadata improves discoverability and interoperability, supporting better analytics and integration with other systems.
- Industry Drivers: motivations from the private sector, such as innovation needs or R&D partnerships that encourage collaboration with academia.

Regarding usage, it can take different forms: researchers use the RCAAP repositories to publish, whereas other researchers and companies may use it to read and reuse publications, leading to citations - the Open Access citation advantage - and collaborations.

To test for the OA citation advantage, two separate sets of publications were created, the first consisting of publications collected from the RCAAP repositories - Green Open Access - and the second consisting of publications published in other venues. By contrasting these two groups, we aimed to identify potential causal relationships between repository deposition and increased research visibility or collaboration.

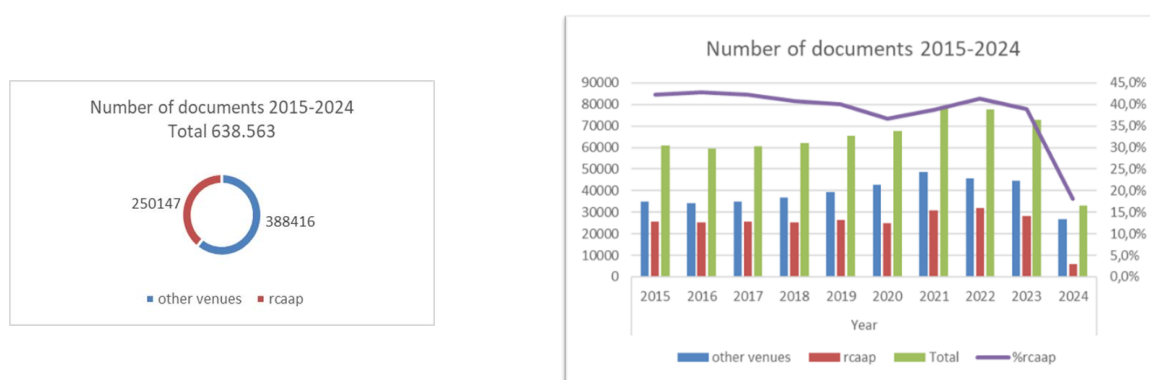
To ensure a comprehensive understanding of the impact it's essential to include scenarios where the object of study is absent. Such an approach was taken in the survey and interviews performed within the Cost-Benefit Analysis and the last Focus Group, where the participants were asked "What if RCAAP didn't exist?" Participants found it challenging to envision a scenario in which RCAAP did not exist. They proposed alternative ways of accessing content, such as

contacting authors directly, seeking similar materials in other languages, or using other Open Access platforms. At the institutional level, the absence of RCAAP appeared even less conceivable. One focus group participant remarked that without RCAAP, the landscape of Open Access publishing in Portugal would be fundamentally different. Many institutions, particularly smaller ones, would lack the capacity to maintain their own repositories without the infrastructure and support provided by RCAAP. This underscores the platform's critical role in enabling equitable access to Open Science across the national research ecosystem.

## 2.4.6. Results

Following the methodology described above, we divided the dataset of publications in two groups – publications harvested from the RCAAP repositories and publications from other publishing venues, also looking into their evolution over time. We understood that about 40% of the Portuguese scientific literature is deposited in RCAAP repositories.

Figure 9: Number of documents 2015-2024



The graph above illustrates the evolution of scientific document dissemination across RCAAP and other venues over a ten-year period. The total number of documents published in each year, represented by the green bars, show a general upward and a peak around 2021–2022.

The documents published in other venues (in blue), increase steadily and peak in 2023 before also dropping in 2024. The number of documents published through RCAAP (red bars), remain relatively stable until 2023 but experiences a notable drop in 2024 resulting from the delay in referencing publications in retrieval sources and their late deposit in repositories.

About 40% of the Portuguese scientific literature is deposited in RCAAP repositories. The percentage of RCAAP documents relative to the total declines gradually over the years, with a steep fall between 2023 and 2024, that could partially be explained by the lag between the discussion of thesis and dissertations and their actual deposit in repositories due to administrative procedures. RCAAP's relative contribution may also reflect changes in institutional strategies or the emergence of alternative dissemination platforms.

The same analysis focusing only on Open Access publications, shows that RCAAP coverage is higher for OA publications (approx. 50%)

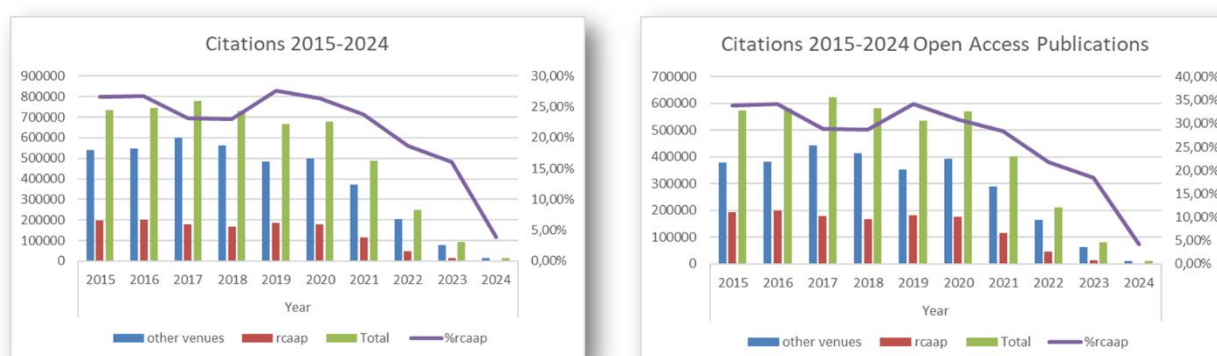
Figure 10: Number of documents 2015–2024 – Open Access



## Citation impact

The citation impact of a publication indicates how frequently it has been referenced by other researchers, serving as a measure of its influence within the academic community. Citations often represent recognition of prior work, are used to support arguments in scholarly writing and are closely tied to the relevance and significance of the research to its field.

Figure 11: Citation counts 2015–2024 – all publications and OA publications



Citation counts are influenced by two key factors that do not necessarily reflect impact: the research field and the publication year, as older publications have had more time to accumulate citations than newer ones. Although bibliometric studies usually do not include the latest two years to five years of publications due to the accumulation effect and the scholarly publication practices, the latest publications were included to allow for a stable time scope.

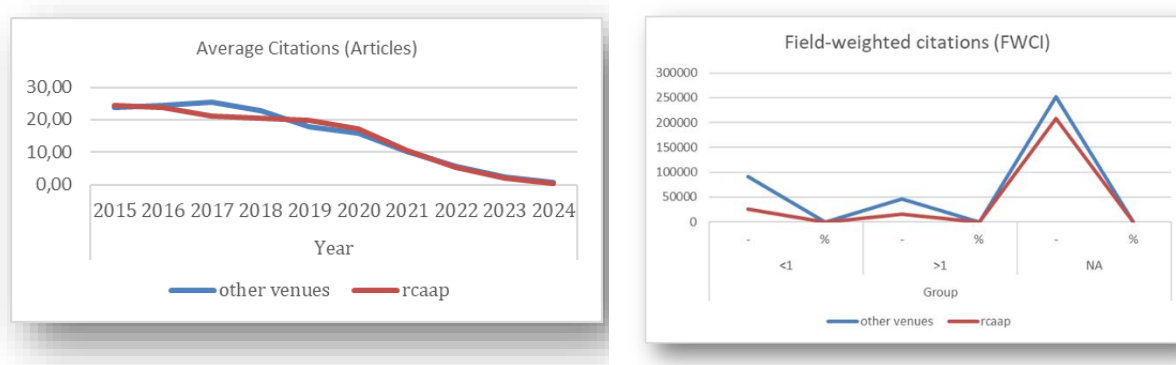
The graphs show citation trends from 2015 to 2024 for all documents (left) and Open Access publications (right). Citations peaked around 2019–2020 and have steadily declined since, with a sharp drop in 2023 and 2024. RCAAP consistently contributes a smaller share of citations

compared to other venues, with its percentage share gradually decreasing throughout the period. This could be explained by the fact that the number of publications for other venues is larger and includes a larger proportion of types of publications more prone to being cited, such as articles, while the RCAAP set has a significant share of thesis and dissertations, which generally receive a smaller number of citations.

The decline is particularly pronounced from 2021 onwards, indicating diminishing visibility or impact of RCAAP publications, and by 2024, both total citations and RCAAP's share fall drastically, due to the above-mentioned accumulation effect of citations.

Considering all publications, the relative number of citations is higher for OA publications, indicating a citation advantage over publications behind paywalls.

Figure 12: Average number of citations per document and Field-weighted citations 2004-2015



To account for the differences induced by publication year and field of science, researchers use normalized citation indicators, which adjust for field and publication year (Waltman & van Eck, 2019). While these metrics allow for fairer comparisons across disciplines and time periods, they can be less transparent than raw citation counts. Considering only articles, which are the most citable items, we found that the mean citations for publications in RCAAP stayed consistently high and followed the same citation trends as the publications from other venues, outperforming them in some years (2015, 2019, 2020 and 2021). RCAAP maintained a strong and stable presence in the scholarly citation landscape during this period and was a major contributor to scholarly citations.

In our dataset, the information for Field-weighted citation index missing from most publications and, although they show the same trend as publication from other venues, the mean citations are lower for RCAAP. This may be explained because RCAAP includes traditionally less cited publications such as reports, master dissertations and PhD thesis, and conference proceedings.

## Science-industry collaboration

One of the main goals of this case was to investigate whether the availability of full-text OA publications available through RCAAP enhances collaborations between higher education institutions and the business sector.

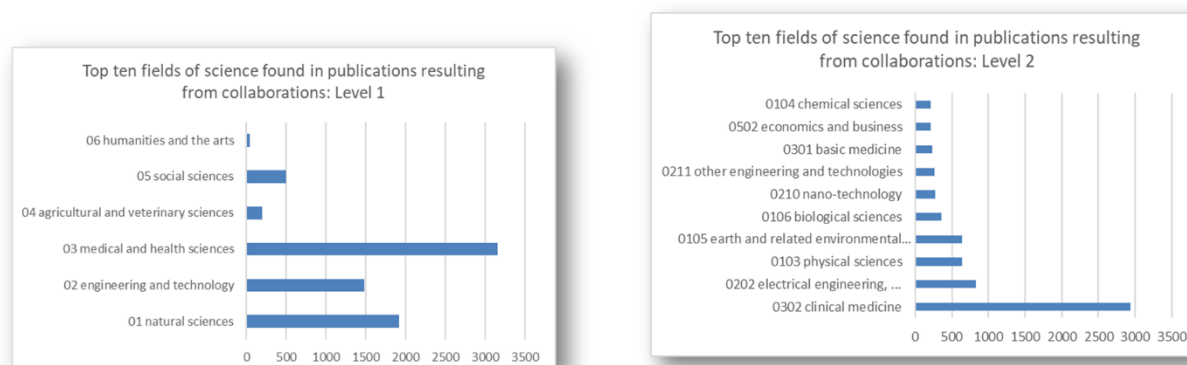
To calculate the number / percentage of publications produced by academia in collaboration with industry that cites Open Science inputs we used the methodology described in the Open Science Impact Indicator Handbook, using the ROR API to query each collected affiliation and categorizing each organization involved in the publication as either an academic institution or an industry entity.

Figure 13: Collaborations between academia and industry 2015-2024



The number of publications where collaborations were identified is low, both for all and domestic collaborations (between Portuguese HEIs and companies). In both cases there is a gradual rise and 2020 appears to be a high point for academic-industry collaboration, possibly due to increased research activity during the COVID-19 pandemic. This peak in 2020 was followed by a decline in total number of collaborations identified. RCAAP shows lesser contribution to the total number of documents in academia-industry collaboration but a higher relative contribution to domestic collaborations.

Figure 14: Top ten Fields of Science in collaborations – Level 1 and Level 2



Analysing the fields of science, most collaborations are found in medical and health sciences, followed by natural sciences and engineering and technology, showing a strong emphasis on STEM fields, especially those with direct industrial applications. Social sciences, agricultural and veterinary sciences and the humanities and the arts are the disciplines showing less publications in collaboration. Breaking down the broader categories into more specific disciplines, we find Clinical Medicine, Electrical Engineering, Physical Sciences and Earth and Related Environmental Sciences leading the collaborations.

## Cost savings

The cost-savings indicator is designed to reflect the efficiency improvements gained through the use of Open Science (OS) resources. Leveraging tools such as open repositories can result in substantial reductions in both time and financial expenditure.

Given the complexity and the wide range of services provided by RCAAP, the scope of the cost-benefit analysis focused on its “core” services—namely, the network of institutional repository infrastructures and support services (RCAAP Portal, transversal and support services such as statistics, training, and helpdesk). Following the cost-benefit analysis (CBA) approach, the study was conducted to compare the benefits derived from operating RCAAP with its associated costs. This involved a comparison between two different scenarios: the actual, observable scenario with RCAAP in place, and a counterfactual scenario in which the project does not exist. This comparison allowed for the identification of the “net” benefits generated, primarily related to cost-sharing and increased visibility of Portuguese research.

Figure 15: RCAAP Cost Savings Model



According to our findings (Catalano et al., 2025), and from a conservative perspective, labour cost savings can be estimated at about EUR<sub>2024</sub> 5 million for 2006-2026 and in recent years, they have reached about EUR<sub>2024</sub> 365,000 on a yearly basis, and data storage cost savings can be estimated at about EUR<sub>2024</sub> 940,000 for 2006-2026. Combining the two different types of benefits, we observe that we reach an undiscounted value<sup>31</sup> of almost EUR<sub>2024</sub> 6 million for 2006-2026 and in recent years, benefits amount to about EUR<sub>2024</sub> 470,000 yearly (2024).

## 2.4.7. Interpretation of Results

The choice of the OpenAIRE Research Graph emerged as the only viable option for analysing the impact of Open Access publications made available through repositories, which are not included in commonly used subscription-based bibliometric databases such as Web of Science or Scopus. This decision allowed for a broader and more diverse range of publications to be included in the study. However, it also introduced greater complexity in data analysis, as the OpenAIRE dataset lacks the same level of standardisation and completeness typically found in commercial databases.

Approximately 40% of Portuguese scientific publications are deposited in repositories within the RCAAP network, revealing a lesser than expected repository coverage of Portuguese scientific literature. This figure increases to 50% when considering only Open Access (OA) publications, indicating that RCAAP plays a significant role in the national Open Science landscape. This level of coverage reflects both the effectiveness of institutional policies and the growing awareness among researchers of the importance of Open Access. It also suggests that RCAAP is a critical infrastructure for preserving and disseminating Portuguese research outputs.

Although the total number of citations is higher for publications outside the repositories - mainly due to the larger volume of such publications - the average number of citations per publication is higher for those available in Open Access and potentially due to systematic differences in publication types and author selection effects between repository and non-repository publications. This supports the existence of an Open Access citation advantage, where freely accessible research tends to be cited more frequently. This advantage is likely driven by increased visibility, accessibility, and discoverability of OA content, which broadens its reach across disciplines and geographies.

RCAAP repositories include a diverse range of publication types, such as reports, master dissertations, and PhD theses, which traditionally receive fewer citations than peer-reviewed journal articles, and as such there are some challenges while performing an analysis using traditional bibliometric indicators. Applying standard bibliometric indicators (like citation counts) to assess the impact of these outputs is more complex and may not accurately reflect their value or influence. This highlights the need for alternative metrics or qualitative assessments when evaluating repository content.

Identifying and quantifying collaborations with companies and industry remains a challenge. This is due to a lack of consistent authorship attribution for non-academic contributors and the absence of persistent identifiers (such as ROR IDs) for companies in metadata records. Additionally, there is a limited culture of formally acknowledging industry involvement in academic outputs, which further obscures these collaborations. Addressing this gap would

require improved metadata standards and stronger incentives for transparent authorship practices.

## 2.4.8. Conclusions

This case study aims to deepen our understanding of how a publicly funded infrastructure, such as RCAAP, influences patterns of collaboration and citation. From the outset, we acknowledged the complexity of this task, particularly given the Open Access nature of the platform, which allows unrestricted access to publications without requiring user registration, and the difficulty in identifying collaborations with industry due to inconsistent authorship attribution for non-academic contributors, the lack of persistent identifiers (e.g., ROR IDs) for companies in metadata records, and the limited practice of formally acknowledging industry involvement in academic outputs. Nonetheless, it was possible to identify areas where collaborations are present and to outline potential directions for future in-kind research initiatives.

The diversity of publication types in RCAAP repositories, including technical reports, master's dissertations, and doctoral theses, which traditionally receive fewer citations than peer-reviewed journal articles, also presents challenges while trying to apply traditional bibliometric indicators such as citation counts, which may not fully capture the value or influence of these outputs. Consequently, there is a growing need to adopt alternative metrics and qualitative evaluation methods to assess the impact of repository content more accurately.

Approximately 40% of Portuguese scientific publications are currently deposited in repositories within the RCAAP network, underscoring RCAAP's pivotal role in advancing Open Science in Portugal. RCAAP serves as a critical infrastructure for the preservation and dissemination of Portuguese scientific knowledge, supporting national and international visibility.

Lastly, conducting a cost-benefit analysis on RCAAP allowed for a deep understanding of its associated costs, the impact RCAAP has on its users, and how it is perceived within the national community. Using a conservative approach, it was possible to identify conclusive evidence of the financial benefits of RCAAP, with estimated benefits 33% higher than its costs. This evidence serves as a powerful tool for policy development and advocacy, ensuring that investments in open research infrastructures are not only economically sound but also aligned with broader goals of scientific advancement and public good.

## 2.5. French Open Access Infrastructure

### 2.5.1. Overview

The rationale and the enquiry design featured in this case study have been co-constructed with the French Open Science (OS) head at the Ministry for Higher Education and Research (*Ministère de l'enseignement supérieur et de la recherche*). The Ministry had previously worked with two major OS repositories in France, the diamond social and humanities sciences journals publishing platform OpenEdition.org and the green OA portal HAL where deposit of both closed access and Open Access publications is encouraged, and compulsory for some institutions' researchers' qualitative evaluation, such as the over 10,000 working at CNRS. Both platforms are the two main actors of the digital publication ecosystem and host a large proportion of French OA publications. It was then identified that both needed to have a better understanding of the sectoral and geographic provenance of the users accessing their websites and the resources that they contain. In particular, both the Ministry and the platforms were interested in finding out to what extent their services were accessed by societal and economic actors outside of academia, since demonstrating OA impact in other sectors would provide a policy argument to better support OS and its infrastructure.

In order to address this motivation, the case study focussed on the very first step of the chain of impact of Open Science: the access of OA articles on these national OS platforms, not only the general volume of access or single consultations, but from which societal sectors internet traffic converges to the websites of OpenEdition.org and HAL, in order to distinguish scientific, societal, economic users, as well as their countries and disciplinary fields.

The first step was to obtain access to OpenEdition.org and HAL information contained in their servers' logs in compliance of data protection legislation. Server logs are text-based records automatically generated by online servers documenting chronologically client requests, IP addresses, and user-agent details, generally for technical purposes such as usage monitoring, issue detection, performance analysis.

The IP (internet protocol) addresses were selected as proxy of the geographical and sectoral origin of the traffic to the two websites. IP addresses are unique numerical labels assigned to the devices using the Internet Protocol for communication, serving to identify the device and provide its location on the network. The IP stored in a server log is generally not that of the individual computer accessing it, but that of the router allowing the connection. Users connecting through corporate and institutional networks (such as universities or medium and large companies) will therefore be associated with the IPs of those organisations, while users

connecting from home, smaller organisations or through their smartphones will be associated with the IP of their commercial ISP (Internet Service Provider).

Analysis of the IP therefore allows to obtain some insight about the geographical and sectoral provenance of the visits to each of the resources offered by OpenEdition.org and HAL, which in turn allows to explore whether there is a difference between the access profile of open and closed resources (since HAL green archive contains both types of publications) and across different disciplinary sectors. To gather information about the openness and disciplinary affiliation of different publications, we matched the publications contained in the two databases with records in the open bibliographic database OpenAlex. We chose OpenAlex rather than OpenAire because of a previous experiment by the Ministry of Research that evaluated OpenAlex as better fit for purpose.

We were in particular interested to know if, for a given academic discipline, societal sector, country or a combination of thereof, open publications were on average more or less accessed than closed ones. To answer this question, we developed a tool called the “Log Explorer” allowing us to filter the log data by academic disciplines, societal sectors (as defined by the [NACE typology](#)) and countries and to calculate an indicator called “Open Access advantage”. The IP-based sectoral classification was achieved using a custom [Node.js](#) script and an iteratively refined prompt relying on the open source Large Language Model Llama 3.3. For more details on this indicator, see the “Methodology” section below, and the Indicator Handbook). The exact nature of the data we collected and the precise protocol we used to clean, prepare and analyse them are described in deliverable 4.4 “Data and tools for the long-term evaluation of Open Science”.

The most direct beneficiaries are the two portals, OpenEdition.org and HAL, which have been directly involved in the development of this case study, not only making their log data available, but also helping the PathOS team to identify the most relevant ways to address the questions of this study. From the very beginning, our research has been informed and guided by their experience, with the expectation to facilitate the reuse of the results generated by the case study. At the same time, since our research protocol mobilised relatively standard data (server logs in the international COUNTER 5 standard, IP addresses’ public attribution, and OpenAlex bibliographic metadata), it should be relatively easy for Open Science portals and institutions in other countries to apply this methodology to their data to investigate similar research questions. This re-use of our research is facilitated by the fact that all the code we developed for this case study is published under an Open Source licence and made available on GitLab (<https://gitlab.huma-num.fr/path-os>).

A second key beneficiary of our work is researchers, decision makers and, more generally, anyone interested in knowing the openness status and the individual identity of the publications that are more often consulted within a given disciplinary, geographical and

sectoral combination, to know which disciplines, countries and sectors are the most avid consumers of Open Science and thus are most likely benefitting from it.

An actual beneficiary is the French Ministry of Research Open Science team who will maintain and further develop the tool after the end of the PathOS project.

This case study has been developed to map the uptake of Open Science through the analysis of access to open publications on national platforms, exploring in particular: (i) public and private actors from private and public organizations, (ii) the time of access including the detection e.g., peaks, regular oscillations, and long-term trends, and (iii) the weblinks to these platforms to identify the dissemination path of OS. We then proceeded to map out the connections among them, through crawling and ethnographic investigation.

This case study has also been conceived from the beginning with institutional actors managing the French OS infrastructure, in order to develop for the long-term and test a complex computational protocol and its distillation in a tool easily accessible to a large set of users. Consequently, the most important results of this case study are the dataset we collected and the exploratory tool that were constructed thanks to it and which will continue to serve the analysis and the promotion of OS by national actors we partnered with.

## 2.5.2. Evidence Landscape and State of the Art

To our knowledge, in our scoping review and relevant studies, connection logs to Open Science platforms had not been analysed systematically and at scale, particularly with the aim of categorising users based on their IP addresses or other potentially intrusive session details. The closest precedent, from which we drew inspiration, was carried out between 2017 and 2020 as part of the “Usages Alpha: Appropriation du savoir ouvert” sub-project of the ANR-10-IDEX-0004-02 grant (<https://lab.hypotheses.org/projets-passes/usages>). This study had analysed OpenEdition’s 2018 connection logs, manually classified the top 1,000 IP addresses, and developed a basic exploratory platform designed to detect “unexpected” uses and perform time-series analyses. In addition to this quantitative approach, they developed an interview grid to conduct finer-grained qualitative analyses of how different actors engage with Open Science platforms. One key takeaway—which we also set out to test in our own case study—was that a significant share of private actors use OS platforms, and that their usage patterns vary greatly in response to societal events such as breaking news as revealed in the publications resulting from this project.

This case study was novel in the sense it means to investigate the very first step in most (if not all) pathways of the impact of Open Science (OS), the access to OS resources: the fact that these resources are consulted by users on the websites and portals that make them publicly available.

While this step is obviously crucial (OS can hardly have impact if it is not accessed in the first place), little research has been carried out on it, since scholars have preferred to focus on studying how OS resources are cited in scientific and grey literature. Focusing on academic citations and other forms of references has the advantage of relying on a solid proxy of the fact that OS resources have been taken up and used by some actors, yet it has the disadvantage to ignore other weaker signals of the potential usage and thus impact of OS, also by non-publishing actors. Societal and economic actors may indeed consult, read and use OS publications, datasets and software without explicitly citing them. The absence of citation can happen either because the rules for referencing vary in different social and academic sectors and are less strict in some of them, for instance in the absence of DOI, or simply because the use of OS does not lead to the production of public documents or documents at all.

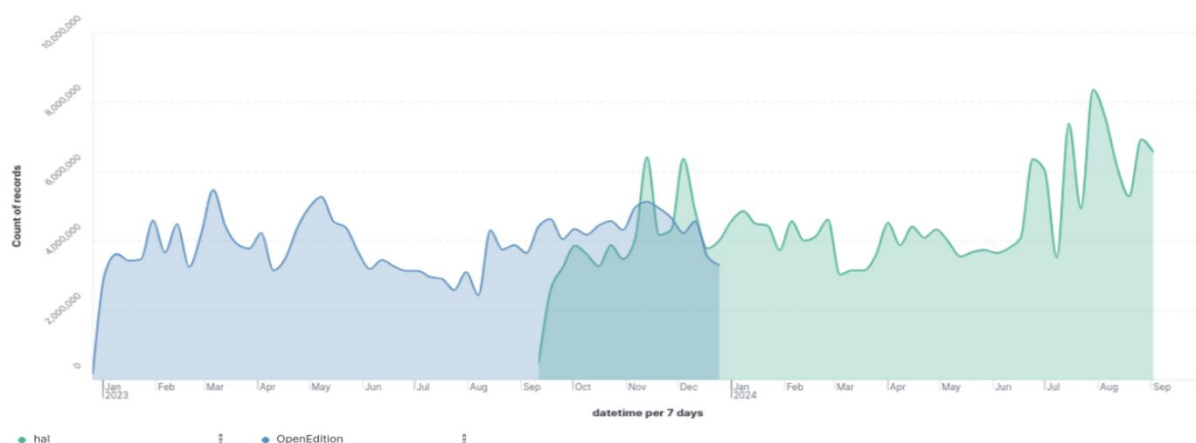
To chase these weak, initial signs of impact, we therefore decided to focus to the most basic proxy of impact, that is the simple accessing of the webpage where the scientific publications are made available through green self-archive HAL and OpenEdition diamond journals platform. The data we obtained through our collaboration with the two platforms cover two partially overlapping periods of roughly one year each:

- From January 2023 to December 2023 for Open Edition
- From September 2023 to August 2024 for HAL

Figure 16: Log Overview: Resource Access Timeline (Jan 2023 – Sept 2024)

## Log overview

timelapse (437,205,178 hits between jan. 2023 and sept 2024)



In the periods captured by our data:

- 1.289.637 resources were accessed in HAL (826.710 open and 462.927 closed)
- 316.001 resources were accessed in Open Edition

As for the accesses to these resources:

- HAL's resources were accessed 51.414.869 times (42.213.685 for the open ones and 9.201.184 for the closed ones)
- Open Edition's resources were accessed 86.619.414 times (76.627.121 for the open ones and 9.992.293 for the closed ones)

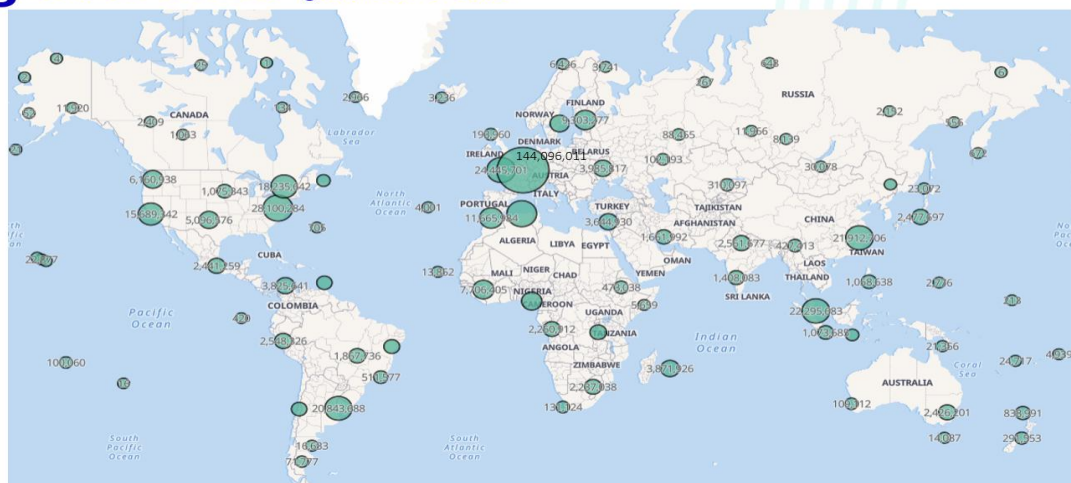
The geographic distribution of the accesses recorded in the logs we analysed displays, not surprisingly:

- France in the first place (with 34.541.755 accesses), followed by
- the US (23.387.010)
- and more distanced: Singapore (7.203.246), Algeria (6.707.774), Germany (6.349.646), Argentina (5.767.977), Canada (5.023.285), the Netherlands (4.688.916) and China (4.148.575).

The numbers for most of those countries may be explained because they have large research and economic communities (US, China, Germany, Canada, the Netherlands, Singapore), while Algeria has strong linguistic and historic ties with France, and Argentina has journals and a community active on OpenEdition.

Figure 17: Log Overview: Geographic Distribution of Users

## Log overview geolocation of users



Maybe more interesting is to consider the breakdown of accesses according to the NACE sectors associated with the IP of the two platforms users. The first two sectors are, by far, Internet Service Providers (62.700.075) and Hosting Services (50.376.954). This is not particularly remarkable and simply means that most users are accessing resources from their home or from the networks of organisations too small to have a dedicated IP. At the third place we find, as expected, Education with 14.948.624 accesses (most likely corresponding to views by students and scholars from academic institution).

The four following sectors however also amounts for a remarkably high number of accesses, reflecting uptake in the economic sectors: Telecommunication, Computing & Information (4.128.244 accesses); Professional Scientific and Technical Activities (1.840.308 accesses); Publishing, Broadcasting, Content Production and Distribution (1.403.981 accesses); Public Administration and Defence (965.603 accesses). Other sectors have significant number of accesses too.

Together all economic sectors other than ISP, Hosting and Education amount to close to 10 million accesses (9.978.534), which is relatively close to the 15 million accesses of the Education sectors. This is an interesting finding: unlike what may be believed, OS portals such as HAL and Open Edition are not at all exclusively visited by academic researchers and students, but also have a wide public access in the rest of society, therefore a societal and economic impact.

## 2.5.3. Impact Pathway Logic

As mentioned above, this case study focuses on the very first step of the impact chain: *access*. As an extreme summary, we tried to answer the question: how much are open publications more or less frequently accessed than closed ones, given their availability? And which societal and economic sectors are most interested in Open Science resources, again given their availability?

The clause “given their availability” which we introduce in the two questions above is crucial as it allows taking into account the most important enabling factor/barrier to the access to Open Science, which is, of course, its availability.

While the availability of open resources is obviously crucial in facilitating (or hindering) their access, its consideration is not easy to implement in our computation. While it was easy to define the overall number of open and closed resources on HAL, because open resources are made available as a file to be downloaded while closed resources only display a notice with the reference and abstract without the publications’ file, all resources available are clearly not all relevant for different types of societal sectors. Users from the “Arts, sports and recreation” sector, for example, are probably interested in a limited subset of the resources available in the platforms and their preference for Open Science is best examined in relation to those resources (rather than the total of the publications available).

There is, however, no easy way to presume which resources are relevant for a given social or industrial sector. In this case study we used a pragmatic approach defining as relevant all the resources accessed at least once through an IP associated with that sector. In comparing the number of views collected from a given sector by open and closed resources, we therefore

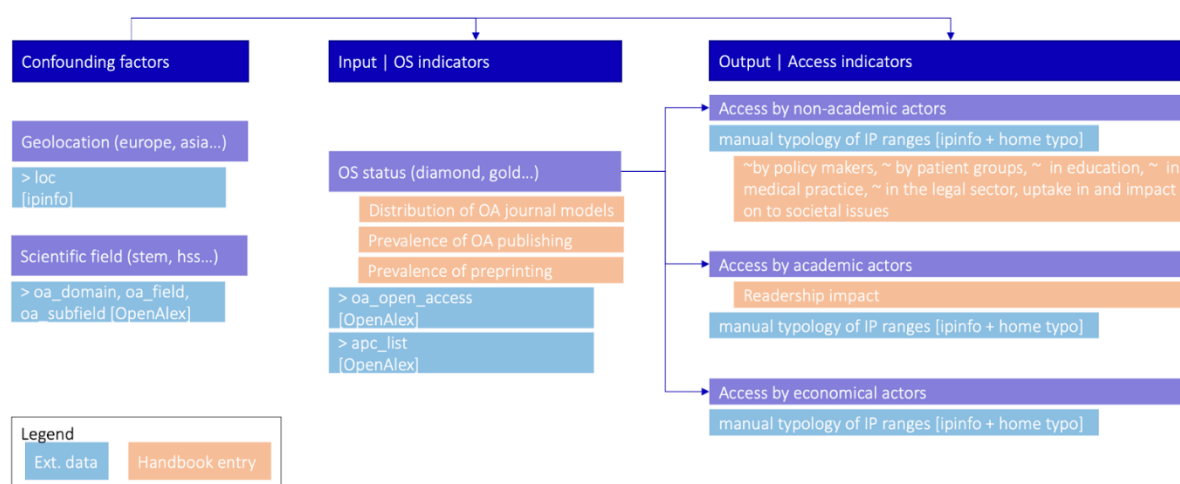
normalise these numbers by the number of open and closed publications consulted from that sector.

Finally, as noted above, our analysis noted three potential confounders:

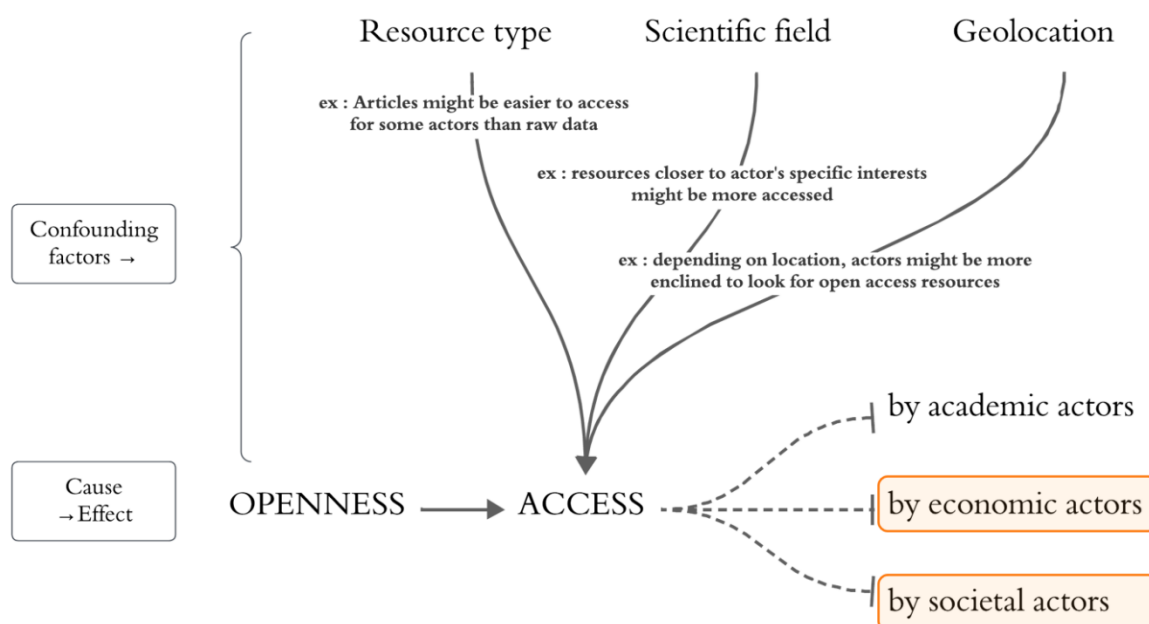
- **The geolocation of the accesses** (as different sectors may be differentially interested in OS in different countries)
- **Scientific domain** — on the side of *scientific demand*, the topic of a resource may reflect vastly different societal needs and thus differing urgencies for free, immediate access; on the side of *scientific supply offer*, disciplinary traditions (e.g., physics versus literature) have long exhibited unequal propensities to publish openly.
- **Resource type**, distinguishing for instance between articles and databases, which pose challenges analogous to those under factor 2.
- **Pathway diagram**

## PATHWAY DIAGRAM INCLUDING CORRESPONDENCE WITH HANDBOOK ENTRIES

Figure 18: Impact pathway logic for the "French Open Access Infrastructure" case study



## SIMPLIFIED VERSION (ADDING THE RESOURCE TYPE AS A CONFOUNDING FACTOR)



### 2.5.4. Methodology

To facilitate the comparison between different combinations of disciplines, countries and sectors, we developed a metric called *Open Access Advantage indicator* which is available in the Handbook. This metric compares the ratio of accesses directed to OS resources over the total number of views, to the ratio of open publications among the relevant for resources. This indicator can be computed for an entire platform, for a single discipline (considering all publications associated with it), a single country or sector (considering all the publications accessed from it), as well as for any combination of thereof.

The indicator is operationalised through a relatively straightforward metric comparing the ratio to which OS resources are accessed to the ratio of their availability. More precisely, the metric is calculated as follows:

$$OS \text{ access advantage} = \frac{AO}{AO + AC} - \frac{SO}{SO + SC}$$

Where:

- SO (stock open) is the unique count of OS resources accessed at least once in a given disciplinary, sectoral and geographical combination.

- SC (stock closed) is the unique count of non-OS resources accessed at least once in the same combination.
- AO (access open) is the number of views collected by OS resources
- AC (access closed) is the number of views collected by non-OS resources

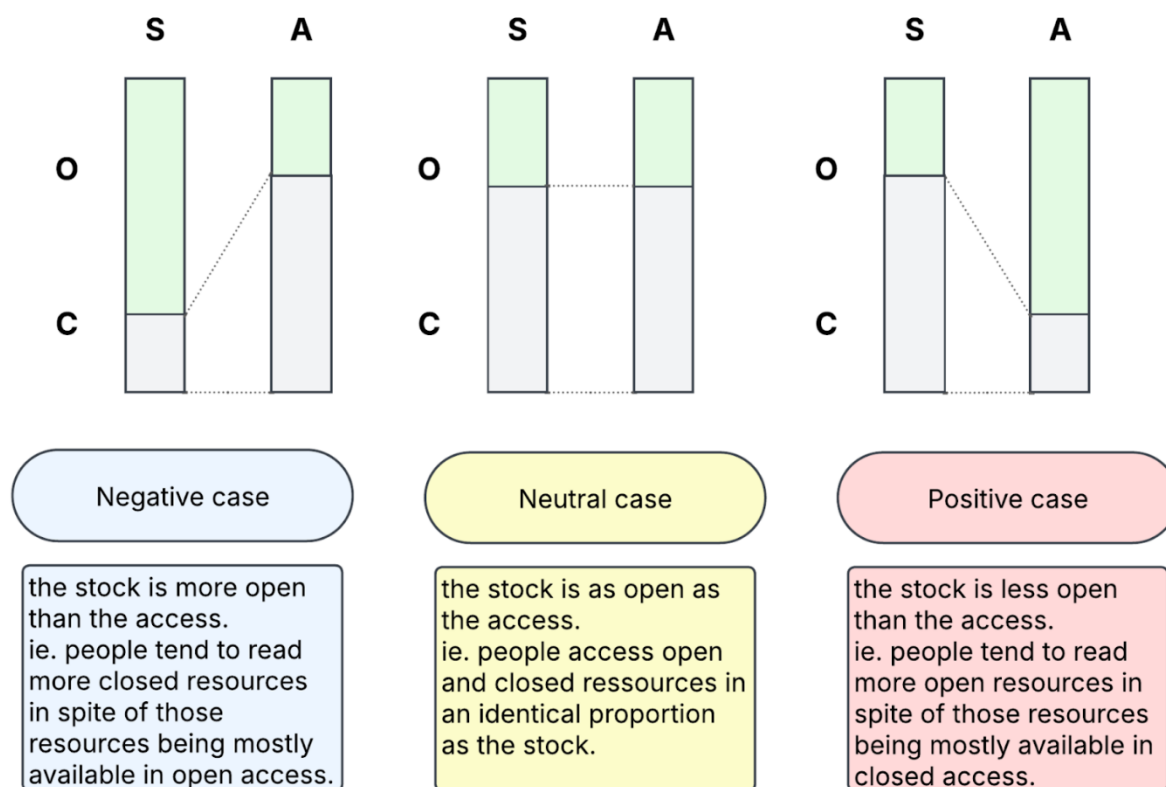
Basically, the metrics compare the ratio of SO resources (the part of SO over the total of relevant resources) to the ratio of the accesses directed to SO resources (over the total of accesses).

The choice of using ratios instead of absolute values is motivated by the fact that, as explained above, this indicator is meant to be used comparatively, which entails that its values have to be somewhat normalised in order to bring sets of scientific resources of different size to the same scale. Since both the ratios vary between zero and one, their difference varies between one (when most of the available resources are closed and yet most of the accesses are directed to few open ones) and minus one and is minimal (when most of the available resources are open and yet most of the accesses are directed to the few close ones).

As both ratios can be expressed as percentages, the Open Access Advantage indicator can also be expressed as a percentage: the larger it is, the higher the probability of open resources to be consulted compared to closed ones.

Another way of understanding the metrics is as the *difference* between 1) the actual access of OS resources and 2) the expected access that they should have if open and closed resources were equally accessed. In that case, the two ratios computed in the formula above would have the same value and the OS advantage would be null. A positive value thus means that OS is consulted more than expected given its availability (and a negative value that it is consulted less than expected).

Figure 19: Interpretation of the Open Access Advantage Indicator (Negative, Neutral, Positive Cases)



The indicator is meant to be computed based on access data collected by the platforms that make scientific resources available. Conveniently, these data are generally stored in the logs of the servers of the two national databases of scientific publication: HAL ([hal.archives-ouvertes.fr](http://hal.archives-ouvertes.fr)) and OpenEdition ([openedition.org](http://openedition.org)). Data we collected for over a year have two main advantages:

- First, this data can be enriched with the metadata associated with each of the resources, in order to know for each accessed page whether it corresponds to an open or closed resource in the case of HAL, and in which scientific area (and potentially by which authors, institution, year of publication, etc.)
- Second, server logs contain information about the exact moment in which each page has been consulted (i.e., they are precisely time-stamped) and from which client IP-addresses a numerical label which allows to know through which gateway the resource has been accessed. This does not allow to identify single users or computers (probably for the best), but it does provide information about their geographical and, in some cases, institutional location (through services such as IPinfo.io, which we thank for having provided their data free of charge).

We focused on the scientific domains associated with the accessed resources and aggregated the IP-related information in countries and industrial sectors. For the second classification we

used in particular the NACE typology ([ec.europa.eu/eurostat/web/nace](https://ec.europa.eu/eurostat/web/nace)) to which we added the “ISP” and “Hosting” as standalone categories (because they were over-represented in our data).

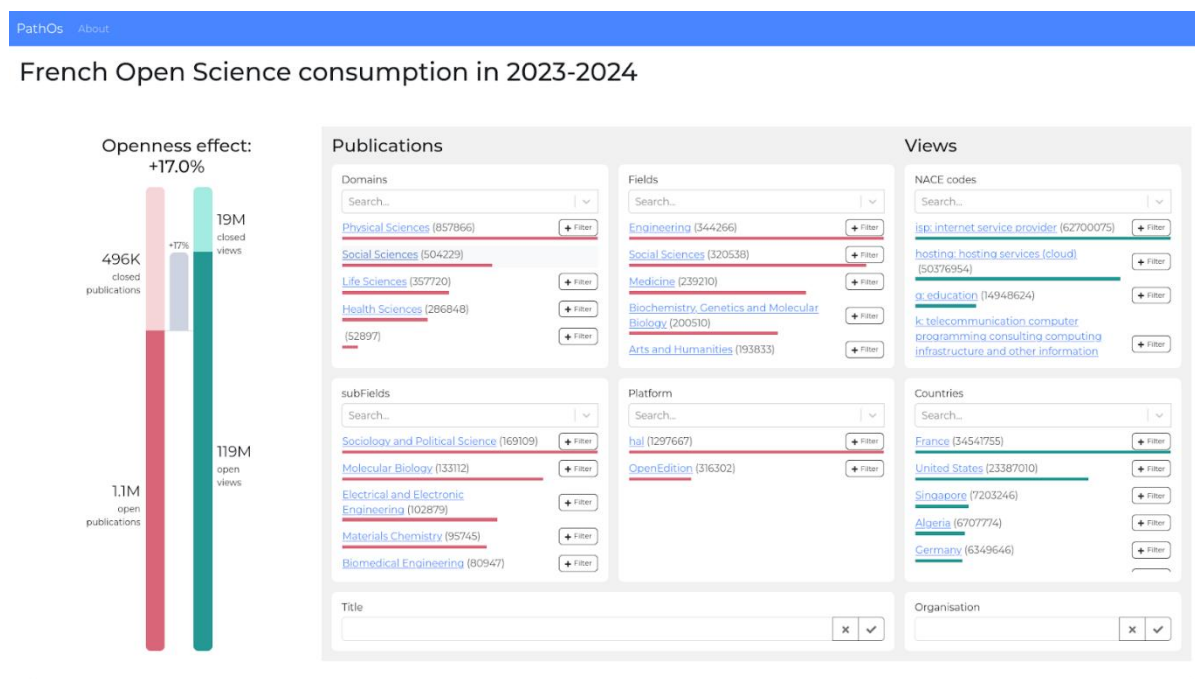
This data structure allows observing not only the general access advantage (or disadvantage) of OS, but to break the calculation down to different combinations of scientific domains, countries and industrial sectors, in order to facilitate comparison and detect potential confounding factors.

The reusable take-away from this case study is the tool called “Logs Explorer” which we developed to analyse the logs. This tool offers its users a faceted navigation system, allowing them to filter the information contained in our dataset by applying a series of four filters:

1. The platform to be considered (HAL or OpenEdition)
2. The 26 main disciplinary field according to OpenAlex classification
3. The 21 economic sectors identified by the NACE categorization [https://showvoc.op.europa.eu/#/datasets/ESTAT\\_Statistical\\_Classification\\_of\\_Economic\\_Activities\\_in\\_the\\_European\\_Community\\_Rev.\\_2.1.\\_%28NACE\\_2.1%29/data](https://showvoc.op.europa.eu/#/datasets/ESTAT_Statistical_Classification_of_Economic_Activities_in_the_European_Community_Rev._2.1._%28NACE_2.1%29/data)
4. The countries from which HAL and OpenEdition resources have been accessed.

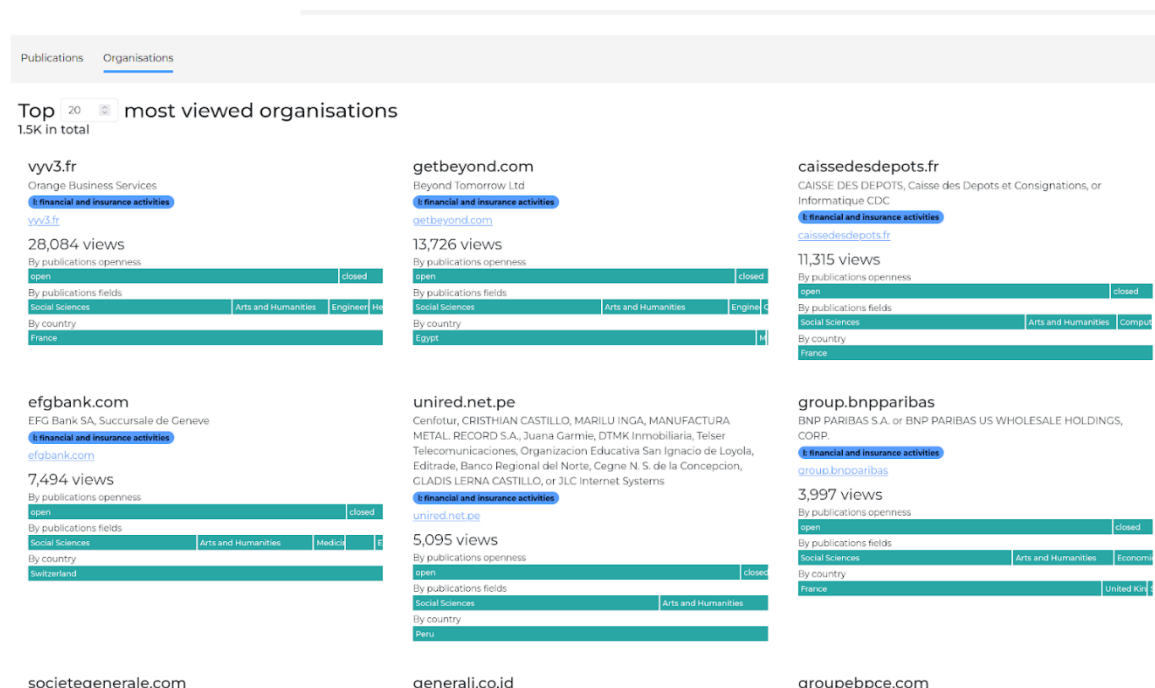
Users of this tool can choose one or more values for each of these dimensions and the data are filtered accordingly. After filtering, the tool displays the number of resources accessed for the combination of filters defined by the user and the total number of views gathered by them. It also generates two 100% stacked bar charts displaying respectively how the totals of resources and views break down into open and closed publication. Based on these values, the tool computes the Open Access advantage indicator, which corresponds visually to the difference in height between the bar representing OS views and the bar OS resources (see the image below).

Figure 20: French Open Science Consumption in 2023-2024



The most important part of the tool, however, is below the filters and bar charts and corresponds to the list of the most accessed open and closed publications for the selected combination of filters. This allows the users to inspect, for a given combination of filters, the individual publications that gather the largest shares of accesses, as well as the individual organisations that are responsible for most accesses (see the image below).

Figure 21: Most Accessed Publications and Organisations



We structured our work on logs around three “datasprints”, i.e. one-day long technical work sessions that we organised and conducted hands-in-hands with:

- OpenEdition technical team, research lab and managing board,
- HAL technical team,
- A data scientist of the Open Science team from the French Ministry for Higher Education and Research which will take over the tool after the end of the project,
- Three programmers of CNRS *Institut pour l'information scientifique et technique (Institute for Scientific and Technical information)*, who have been in charge of analysing the logs of CNRS uses of both for pay and Open Access resources for more than ten years,
- The scientists in charge of the previous study led at OpenEdition, *Usages Alpha*, which is mentioned above.

As Open Science platforms are socio-technical arrangements co-shaped by public research policies, technical norms and practices, digital governance, and diverse scientific, economic, and societal uses, we followed Science and Technology Studies (STS) protocols to consider them as objects of scientific inquiry and involved representatives from the sectors to ensure the most relevant analysis, also in line with participative research co-construction methodology and knowledge commons involving the studied community as an actor.

Constraints and challenges to our analysis were of many sorts.

First, the quality of our data partially depends on the quality of external sources (*OpenAlex* and *Ipinfo*), as well on their own methodological choices, especially concerning typologies (as for example *OpenAlex*'s “[End-to-End Process for Topic Classification](#)” or *Ipinfo*'s internal non-open-source network mapping algorithms).

In this regard, NACE categories have also been vastly debated and reviewed several times: they are partially ambiguous for humans, and even more so for the large language models we used for automated classification.

Large language models are also, *per se*, black boxes, whose classification “choices” cannot be brought back to logical decision trees but to probabilistic computation. As we describe in D3.4, we worked iteratively and empirically, testing multiple prompts until they aligned as closely as possible with human inter-rater reliability.

Second, as noted above, monitoring access alone reveals little about actual use or long-term impact; it is merely a first, preliminary step. Considering access metrics as mechanical proof of impact would be a methodological shortcut. The progression from access to usage to impact is lengthy and complex, and requires qualitative insights (for example, methods from the sociology of reception) to be properly understood.

Similarly, IP-based analysis lacks granularity—not only at the geographical level, but also because it tells us only which organization a user belongs to, not anything about them as an individual: their sociodemographic profile, their position within the organization, the context of their appropriation and use of scientific resources, their motivations, education, or specific needs.

## 2.5.5. Causality Narrative

In this case study, we experimented with an approach substantially different from the one employed in the rest of the project. While in other case studies, causality is assessed through precise hypothesis and complex statistical tests, we took an exploratory approach and built a tool that allows users not only to compute the Open Access advantage for the particular combination of sectors, countries and disciplines that are more relevant to them, but more importantly to explore the individual publications that amount for the highest number of views both for open and for closed accesses.

Working with two major platforms of Open Science in France revealed that actors need the capability to decompose general trends of OS usage in the specific publications that within each discipline sparked the interest of different social and industrial sectors, as well as identify the organisations within those sectors that are responsible for most accesses.

## 2.5.6. Results

As discussed above, the main result of this case study is an exploratory tool, rather than a table of results. We did however also compute the access advantage indicator (described in 2.x.4) for all the combinations of platforms, fields, countries, and economic sectors. This, of course, produces a very long table, as there are 119.746 possible combinations of our four filters. Some of these combinations, however, need to be excluded as they are associated with a number of viewed resources and of views that are too little to be significant. We thus excluded all the combinations associated with less than 100 viewed resources either open or closed and less than 1.000 views to either. This leaves 7.576 combinations, which we ranked by access advantage indicator. We then selected the most interesting combination, which we report in this table<sup>12</sup>.

---

<sup>12</sup> [https://imisathena.sharepoint.com/sites/PathOS/\\_layouts/15/doc.aspx?sourcedoc=%7Bd365dafc-d482-4b17-8831-20d911db28f1%7D&action=edit](https://imisathena.sharepoint.com/sites/PathOS/_layouts/15/doc.aspx?sourcedoc=%7Bd365dafc-d482-4b17-8831-20d911db28f1%7D&action=edit)

## 2.5.7. Interpretation of Results

Openness seems to be associated with a significant access advantage on HAL (+18%). This might have to do with the fact that in HAL the number of open publications is more directly comparable to that of closed ones (open resources are less than double of the closed ones).

A finding is that while the Open Access advantage in the educational sectors is significant and amounts to 2.1%, the advantage for all other sectors together (excluding 'neutral' codes) is significantly more important adding up to 3.5%. Overall (combining all discipline and countries), Open Science seems to have a bigger advantage outside academia than inside it.

The breakdown of single social and economic sectors is even more interesting.

- 'Neutral' codes (corresponding to IPs not assigned to specific sectors or to people connecting to a commercial internet service provider) have the highest Open Access advantages. *Internet service provider* (16%), *Hosting services* (15%).
- Important but not huge advantages concern *Public administration and defence* (8%), *Professional scientific and technical activities* (5%), *Administrative and support service activities* (4%), *Publishing broadcasting and content production and distribution activities* (4%), *Arts sports and recreation* (4%), *Telecommunication computer programming and other information service activities* (3%), *Agriculture forestry and fishing* (3%), *Electricity gas steam and air conditioning supply* (3%).
- Disadvantage can be found for *Construction* (-13%), *It: undefined* (-12%), *Wholesale and retail trade* (-3%).

The biggest Open Access advantage (in HAL) concerns fields particularly in the natural sciences.

- All the following fields have significant Open Access advantages independently from the platform, country and NACE sector: *Chemical Engineering* (+32%), *Chemistry* (+29%), *Energy* (+27%), *Materials Science* (+26%), *Pharmacology, Toxicology and Pharmaceutics* (+25%), *Engineering* (+24%), *Nursing* (+22%), *Dentistry* (+22%), *Environmental Science* (+21%), *Veterinary* (+21%), *Economics, Econometrics and Finance* (+21%), *Decision Sciences* (+21%).
- No field has an Open Access advantage smaller than 10%.

The biggest Open Access *disadvantage* concerns countries, in particular Arab states.

- All the following countries have significant Open Access *disadvantages*: *Iraq* (60%), *Sudan* (57%), *Libya* (55%), *Syria* (51%), *Palestinian Territory* (46%), *Jordan* (40%), *Egypt* (34%), *Oman* (22%), *Lebanon* (12%), *Qatar* (10%),
- A few countries have relevant (but not huge) Open Access advantages: *United States* (12%), *India* (9%), *France* (8%), *China* (8%), *Canada* (7%), *Italy* (7%), *Portugal* (7%), *Singapore* (6%), *Bolivia* (6%), *Mozambique* (6%).

The significant Open Access disadvantage in Arab states likely reflects multiple factors including language barriers, different academic publishing traditions, varying institutional access arrangements, and potentially different research information-seeking behaviors, rather than necessarily indicating preferences against Open Access per se.

## 2.5.8. Conclusions

Across different disciplinary fields, geographical location, social and economic sector openness seems to be associated with a significant access advantage. In almost one fourth (23% or to 1769) of the 7,576 combinations associated with a significant amount of viewed resources and views, Open Access publications have an advantage that is greater than or equal to 5%. Closed publications have such an advantage only in 7% of the significant combination (545 in total). As a limited but clear result, we can affirm that Open Science is more intensely accessed than closed one in HAL, a platform that contains a more balanced distribution of open and closed resources. While these access patterns provide valuable insights into user behavior, progression from access to actual use and eventual impact requires additional investigation through complementary methods.

This case study aimed both to study how openness affects access to scientific resources and to develop a practical tool for policymakers and platform partners to continue this analysis.

## 2.6. Effects of Data Repositories on Data Usage

### 2.6.1. Overview

This case study investigates the effect of data repositories on the use of data for research. How can we establish usage of datasets in research? Are datasets from some repositories more likely to be used for research than other repositories? What factors are relevant for those differences in usage?

We were interested in the effect of the repository where data is shared on the subsequent usage of that data. For research articles, we know that it matters where an article is published for subsequent citations. Does something similar happen with citations to datasets, and are citations affected by the repository where the dataset is stored? That is, would sharing data in a particular repository result in more reuse?

We analyse these effects in general, but also with a particular interest in the social sciences, based on three different studies:

- we study the extent to which data usage can be automatically inferred from scholarly publications in the social sciences;
- we interview scientists in the social sciences about their data usage, and what role data repositories play;
- we study data usage quantitatively on the basis of the Data Citation Corpus.

Our case study concentrates on the OS aspects of scientific impact and reproducibility impact, specifically on *Availability of data repositories*, *Reuse of data in research*, and *Impact of Open Data in research*. More information can be found on the website of [Open Science Indicator Handbook](#).

We find that it remains a challenge to algorithmically capture correct data mentions within the SSH field, implying that the data extraction method is not reliable. If the SSH field used permanent identifiers with their datasets, data extraction would be more reliable. We also find that data reuse is reasonably common in SSH (30%). Surprisingly, no datasets come from known Open Data Repositories, but almost exclusively from ministries, bureaus of statistics and public institutes. However, it is nearly impossible to identify exact dataset and/or repository and challenging to determine which sentence mentions data reuse, due to low recall and precision. We also see an overestimation (high recall, low precision) of algorithmic identification of “use” of data, which does not always refer to do the actual use of the data, but just to mentions of other data

The intended beneficiaries of the case study are individual researchers and research institutes but also academia and governmental functions at large, as the results could increase data

repository use, improve the features and conditions of data repositories, incentivize use of open data practices, metadata practices in the SSH field, support data reuse and increase the scientific impact and quality of reproducibility.

## 2.6.2. Evidence Landscape and State of the Art

Over the past decade, the Open Science movement has gained increasingly more traction. Whereas initial efforts focused on Open Access, the movement expanded to encompass many more aspects, ranging from also considering data and software to questions of broader inclusion. The sharing and use of data is also on the rise (Digital Science, 2020). There are several initiatives that monitor Open Science, most prominently the French Open Science Monitor (Bracco, 2024). Here, we are interested in the reuse of data, in particular, the role of data repositories (Liaw, 2021) and the potential effects that the data repository where data is shared can have on subsequent reuse. That is, would sharing data in a particular repository result in more reuse?

In our study, we relied on the Global Data Citation Corpus (Datacite, 2024) in order to explore the effects of data repository on subsequent data reuse, we call this the “Data citation corpus analysis”. However, this dataset focusses mostly on the biomedical domain, while we originally had a particular interest in the social sciences and humanities (SSH). Data practices in SSH are quite distinct from those in some of the other sciences (Gregory, 2023; Gregory, 2024). Ideally, we would analyse how data from various data repositories in SSH is reused. However, there are scant data citations, such as those indexed in Datacite or Crossref, while the Global Data Citation Corpus (Datacite, 2024) focusses mostly on the biomedical domain and can therefore not be used to analyse the SSH. For this reason, we also analysed to what extent data citations can be automatically extracted from the literature, we named this “Data mentions analysis”. In addition to the quantitative section of this case study, we also aimed to conduct interviews to understand how SSH researchers reuse data, what factors play in their decision-making processes and what makes a data repository accessible and favourable for (re)use in their field.

## 2.6.3. Impact Pathway Logic

In this case study, we investigate how the availability of a national data repository affects the uptake of research data and whether this leads to greater reuse of research data and enhanced scientific impact in the context of Open Science practices in the social sciences and humanities (SSH). That is, does it make a difference whether the data is being made available through, for example a national data repository or through other data repositories such as international (e.g. Zenodo) or disciplinary repositories (e.g. ICPSR).

The assumed impact pathway starts at the point of data being made available through repositories. The pathway logic rests on the idea that where a dataset is stored could influence how often and by whom it is reused due to increased data visibility, discoverability and at times quality or perceived credibility, and could eventually result in higher rates of data reuse, improved reproducibility, and increased scientific impact.

This pathway is influenced by several enabling factors including:

- **Availability.** Datasets that are not made available cannot be reused.
- **Findability.** Datasets that are difficult to find might be less likely to be used.
- **Data quality.** Datasets that are of better quality might be more frequently used than more messy datasets that would require more cleaning.

Note that enabling factors are also confounding factors. Confounding factors include but not limited to:

- **Country.** For example, datasets on the US might be used more frequently than datasets on the Netherlands because more researchers might be studying the US than the Netherlands.
- **Scientific discipline.** For example, datasets in the social sciences may be less frequently used than datasets in the biomedical sciences.
- **Journal** (or accompanying publication). Datasets that are part of a publication in a high-impact journal may perhaps be more frequently used than datasets that are part of a publication in a low-impact journal.
- **Size of the dataset.** Larger datasets might be more frequently studied than smaller datasets, while the size might simultaneously affect which repository a particular dataset will be stored in.
- **Quality of the research.** Higher quality research articles might be more likely to share data that is of broader interest to the community.

Our case study is based on three different studies: an analysis of scholarly texts in the social sciences to detect data usage, interviews with researchers to gather qualitative insights, and quantitative data from the Data Citation Corpus. This mixed-method approach allows us to trace the pathway from data sharing via repositories to potential downstream impacts like increased visibility and reuse. Behavioural evidence, such as citations, inferred mentions of datasets, and references to public data sources, is used to validate the pathway, although not without limitations. Inputs include public data infrastructures, repository systems, and institutional mandates promoting open data. Key activities involve making data accessible, encouraging citation practices, and maintaining usable repositories.

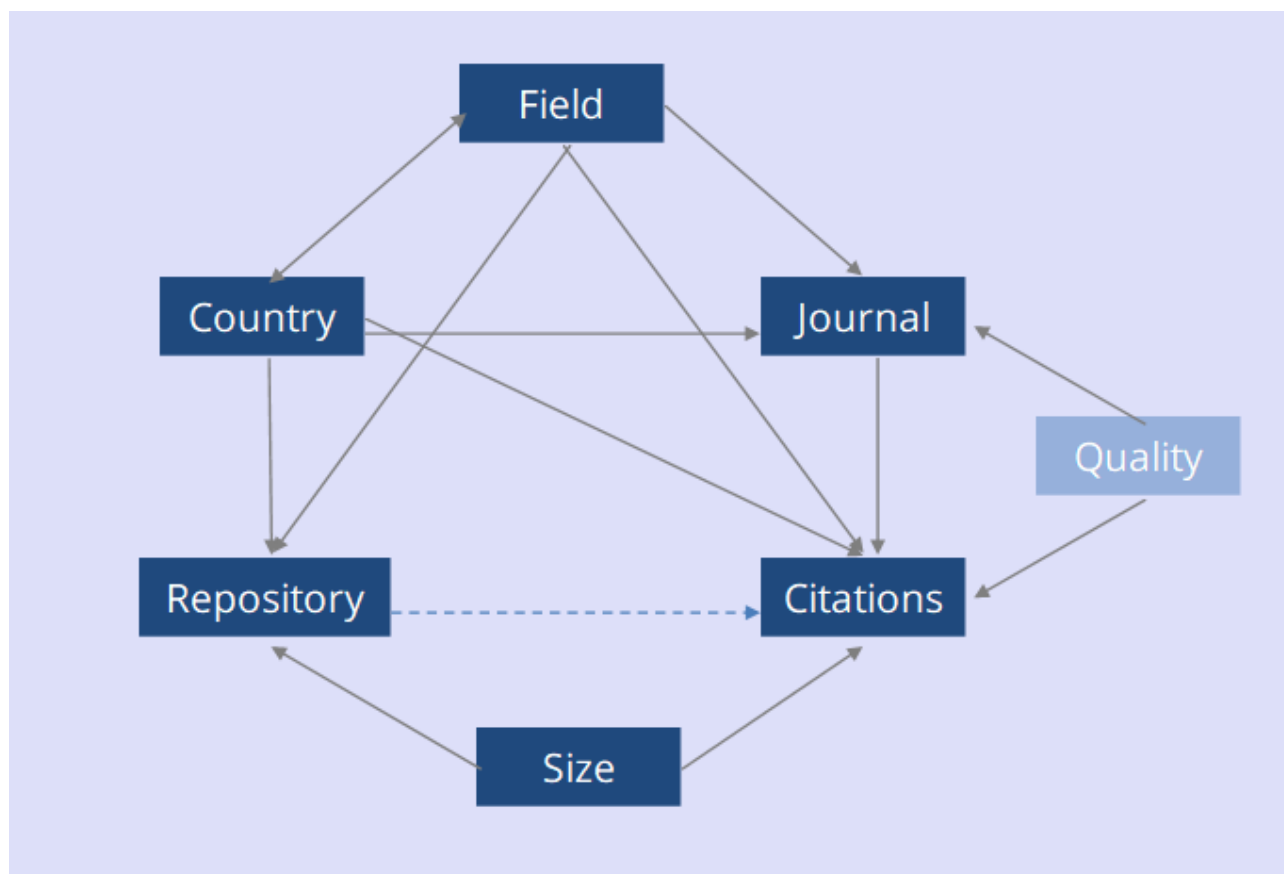
Outputs are observed in the form of data-related publications, increased visibility of public datasets, and researcher engagement with secondary data. However, our findings also

highlight major challenges: in the social sciences, the lack of permanent identifiers and inconsistent citation practices make it difficult to reliably detect when and how datasets are reused. Most reused data appears to come not from open data repositories, but from government sources like ministries or statistical bureaus. Algorithmic attempts to capture these mentions often overestimate actual reuse due to low precision - particularly when interpreting terms like “use” that could have multiple interpretations - making clear that technical, infrastructural, and disciplinary factors all play a role in shaping measurable impact.

## 2.6.4. Causality Narrative

As already discussed in the previous section, there are a number of confounding factors that may influence citations and the repository. In the diagram below, we illustrate in a tentative hypothetical causal model based on how we believe some causal effects may operate. First, some repositories may be more commonly used in some fields, which may also be more citation intensive (e.g. biomedical fields), and so we need to control for field. Second, data may not necessarily be found through data repositories, but instead, through publications that introduced that dataset. In particular, some journals may be more likely to be read, and so we would like to control for the journal in which the data was introduced. Third, datasets that are shared for a longer time already may be more likely to have gathered more citations. We therefore also want to control for the year of the introduction of the dataset.

Figure 22: Causal Model Linking Repositories, Citations, and Contextual Factors



There is one factor that we cannot control for, namely the quality of the original research. That is, the quality of the original research may affect both where the research is studied, and later its relevant data citations. That is, it is a confounding factor between the journal and the later data citations. We need to control for the journal to close the non-causal path: repository  $\leftarrow$  country  $\rightarrow$  journal  $\rightarrow$  data citations. However, journal also acts as a collider on the path repository  $\leftarrow$  country  $\rightarrow$  journal  $\leftarrow$  quality citations. This path is already closed if we condition on the country, and the journal is not directly affecting the repository choice. We would be able to identify the causal effect, given this hypothetical causal model, if we were to also control for the country. The essential requirement here is that the journal does not affect the repository directly, since it would otherwise also act as a collider through quality.

These considerations of causality are only relevant to our data citation corpus analysis in our qualitative analysis. Our data mentions analysis only tries to uncover overall usage of data in SSH and the extent to which algorithms can automatically extract data and repositories from publications.

## 2.6.5. Data citation corpus analysis

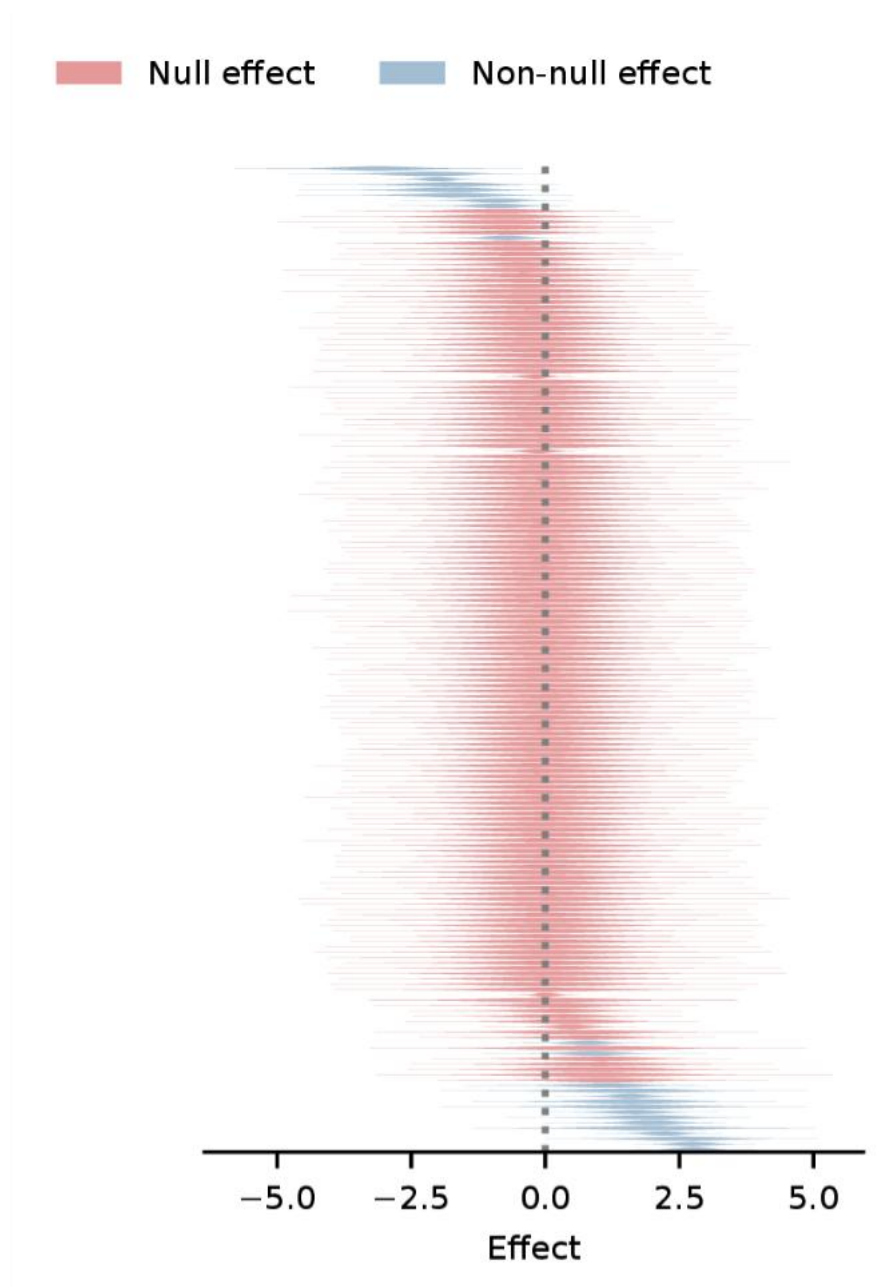
### METHODOLOGY

We study citations to dataset that are supplements of publications. We rely on DataCite to establish which datasets are supplement to publications, while we rely on OpenAlex for metadata about publications of which the datasets are supplements. We use the Global Data Citation Corpus (GDCC) to analyse citations to these datasets. We identified 278.922 datasets that are supplements to unique publications in OpenAlex, while after cleaning the GDCC contains only 3.259 citations for the 278.922 datasets. Given the very low number of citations, we restricted our analysis to whether a paper was cited or not. For more details, please refer to D3.4.

### RESULTS

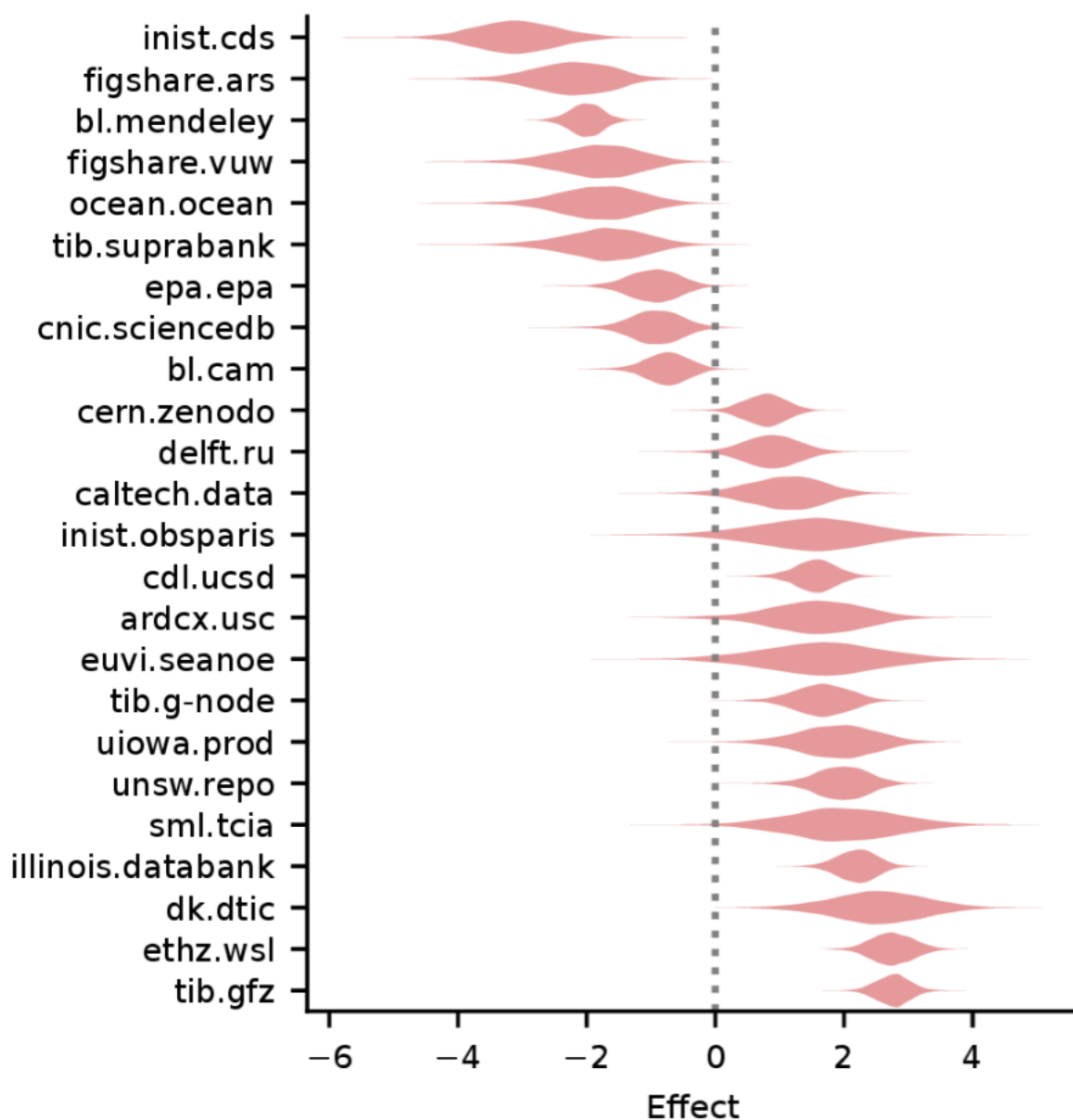
We find that 161 repositories do not show any clear effect on citations ([see images below](#)). Only 21 repositories have a non-null effect, on the basis of the 94% credible interval ([see images below](#)). Of these, 12 repositories have a positive effect and 9 have a negative effect. To provide some idea of the effect sizes, we calculate the marginal effect using the average of the intercept, and setting all other parameters to have a zero effect (this is reasonable, since the effects estimated are constrained to sum to zero). The intercept is estimated to be  $-4.79 \pm 0.15$ , yielding an overall probability to be cited of 0.0083. An effect size of 2 then amounts to raising this probability to 0.058, or a 7-fold increase (i.e. a so-called relative risk of 7). Although not completely symmetrical, an effect size of  $-2$  shows approximately a 7-fold decrease.

Figure 23: Estimates of effect sizes of all repositories on citations to datasets



Each repository is represented as an element on the y-axis, while the x-axis represent the effect of that repository on data usage.

Figure 24: Estimates of effect sizes of repositories on citations to datasets for non-null effect sizes



Each repository is represented as an element on the y-axis, while the x-axis represent the effect of that repository on data usage.

		Mean	3%	97%	N. datasets	% cited
tib.gfz	GFZ Data Services	2.76	2.29	3.24	406	17
ethz.wsl	EnviDat	2.75	2.17	3.35	136	21
dk.dtic	DTU Data	2.53	1.18	3.79	13	38

illinois.databank	Illinois Data Bank	2.20	1.62	2.73	255	10
sml.tcia	The Cancer Imaging Archive	1.99	0.47	3.51	4	75
unsw.repo	University of New South Wales	1.96	1.15	2.68	110	10
uiowa.prod	University of Iowa Libraries	1.88	0.72	2.91	55	9
tib.g-node	German Neuroinformatics Node	1.66	0.87	2.39	151	8
ardcx.usc	University of the Sunshine Coast	1.56	0.10	2.81	25	12
cdl.ucsd	UC San Diego	1.55	0.92	2.14	234	8
delft.ru	Radboud Data Repository	0.86	0.03	1.64	233	3
cern.zenodo	Zenodo	0.77	0.17	1.35	402	4
bl.cam	Apollo	-0.76	-1.40	-0.17	2 139	1
cnic.sciencedb	ScienceDB	-0.92	-1.68	-0.21	1 503	0
epa.epa	Environmental Protection Agency (EPA) Repository	-0.96	-1.72	-0.28	1 752	0
tib.suprabank	SupraBank	-1.72	-2.98	-0.62	355	0
ocean.ocean	Code Ocean	-1.81	-3.04	-0.71	748	0
figshare.vuw	Victoria University of Wellington	-1.83	-3.07	-0.75	866	0
bl.mendeley	Mendeley Data	-2.01	-2.50	-1.55	13 791	0
figshare.ars	figshare Academic Research System	-2.22	-3.42	-1.16	1 492	0
inist.cds	Strasbourg Astronomical Data Center	-3.10	-4.28	-1.92	19 502	0

There does not seem to be a clear pattern showing which repositories are positive and which ones are negative. There is a slightly negative correlation with the number of datasets in the repository, but the relationship is not very pronounced.

Datasets that are older show a clear tendency to have accumulated fewer citations (see image below). Although there is some variation across fields, most fields do not show a clear association with data citations (see image below.) One field that shows a higher number of data citations is, interestingly enough, bibliometric analysis. This may perhaps be due to the field being more aware of some of the developments around data sharing and data citations. Other fields that stand out include Microbial Identification, Melanoma, Landslide Hazards, and Membrane Gas Separation.

Figure 25: Effect of year on citations to datasets.

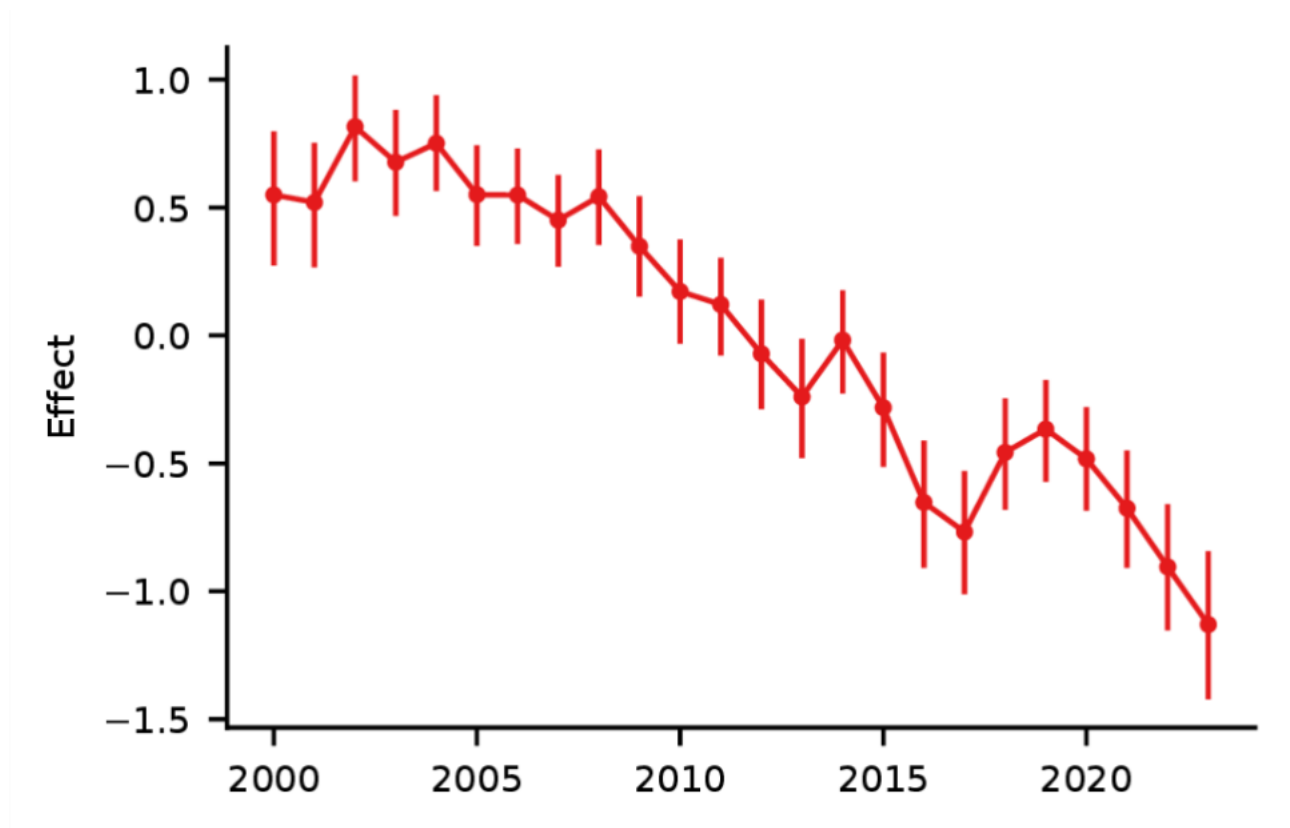
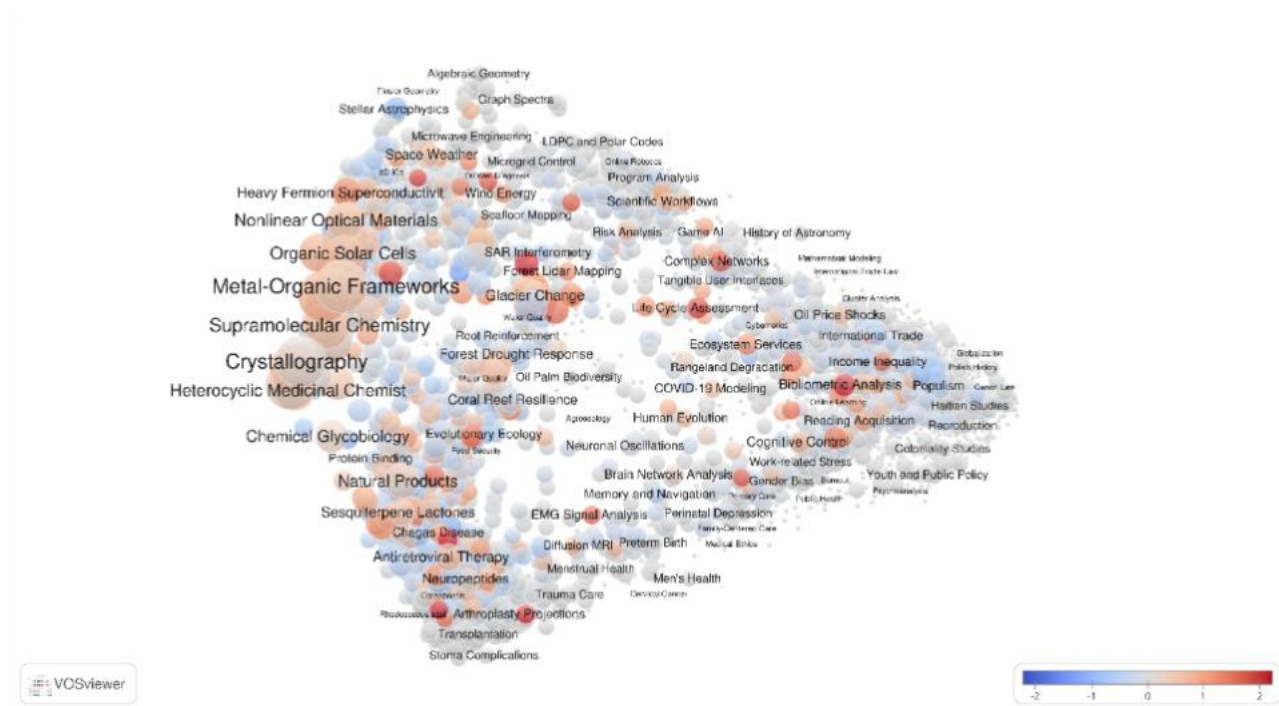


Figure 26: Effect of field on citations to datasets



Each field is plotted based on the citation relations to other fields, such that fields that are more related through citations are positioned closer to each other and fields that are less related through citations are positioned further apart. The axis have no special interpretation, only relative distances are informative. The colour shows the size of the effect, while the size of the field reflects the precision (i.e. the inverse of the standard deviation) of the estimate of the effect.

## 2.6.6. Data mentions analysis

### METHODOLOGY

We have sampled 200 publications from 2017, 2018 and 2019 from Scopus from categories in ASJC that fall under SSH (according to Scopus, check out ASJC categorizations on their website). We explicitly chose relatively recent years in order to increase the chances of data being reused. Of the 200 randomly sampled publications, only 187 publications were accessible<sup>1</sup>, with the remaining 13 being behind paywalls. Of these 187 publications, it turned out that 25 publications were not in English, and we excluded them from the remainder of our study, leaving us with 162 publications.

We manually extracted data mentions from full-text publications based on a fixed protocol. We compared our manual extraction to the algorithmic extraction using dataset and SciNoBo. For more detailed information, please refer to D3.4.

## RESULTS

We found reused data in 49 out of the 162 publications (30%). In total, we found 86 different sentences with any data reuse mentions, of 96 different datasets. There are 48 references to “repositories”, none of which refers to a known data repository as included in RE3Data<sup>4</sup>. Most mentions of repositories that we kept track of refer to bureaus of statistics, ministries and public institutes. For the 96 different datasets, there were 26 references (27%) that were associated with these extracted datasets. Of the 86 sentences, the majority (73) appeared in the main text, with only a few appearing in a caption (12) and only one in the acknowledgements.

**Table 1: Confusion matrix between manual and algorithmically extracted data mentions at the document level.**

	dataset						SciNoBo	
	Mentioned		Named		Implicit			
	No	Yes	No	Yes	No	Yes	No	Yes
Manual								
No	22	83	58	47	39	66	54	56
Yes	3	46	15	34	5	44	2	47

When comparing the manual data mentions to the extracted data mentions using dataset we can discern three types of datasets inferred by dataset: “mentioned”, “named” or “implicit”. At the document level, using the “mentioned” type in dataset has the highest recall, finding 46 of the 49 documents ([see table above](#)), amounting to a recall of 0.94 ([see table below](#)). The “named” type has a substantially lower recall, but shows the highest precision. However, the precision is relatively low for all types. Overall, the “implicit” has the highest [Equation]-score, balancing recall and precision reasonably well. SciNoBo shows an even higher recall than dataset, and also has the highest precision. Nonetheless, with a precision of 0.42, its performance also remains less than satisfactory.

**Table 2: Evaluation of performance at the document level.**

		Recall	Precision	$F_1$
dataset	Mentioned	0.94	0.36	0.52
	Named	0.69	0.42	0.52
	Implicit	0.90	0.40	0.55
SciNoBo		0.96	0.46	0.62

Moving down to the sentence level, we find that the precision decreases even further [see table below](#)). For dataset, we consider any type of data, whether “mentioned”, “named” or “implicit”. The recall at the sentence level is substantially lower than the recall at the document level. The precision is very low for both dataset and SciNoBo.

Note that the matching at the sentence level depends on the similarity calculations. After manual inspection of the potential matches (any sentences on the same page) and the similarity, a cut-off of 0.8 seemed most reasonable. Not requiring exact matches addresses some problems that appeared. For instance, some differences occurred due to references being excluded in the context in dataset. Some inexact matches missed parts of a sentence in the algorithmic data extraction due to apparent problems with sentence segmentation, for instance involving URLs. However, there are also some incorrect matches with a similarity higher than 0.8, notably sentences with a very similar structure.

**Table 3: Evaluation of performance at the sentence level.**

	Recall	Precision	$F_1$
dataset	0.44	0.022	0.042
SciNoBo	0.37	0.078	0.13

Moving further down to the individual mention level, we restricted the analysis to only matching sentences. That is, we only consider whether individual data mentions match if at least the same sentence is identified in both the manual and algorithmic labelling. In a manual evaluation, a cut-off similarity of 0.4 seemed most reasonable. Most algorithmically extracted data mentions below that cut-off simply refer to words such as “data”, “sample” or “variable” in dataset. In some cases, data collected by researchers (e.g. surveys or interviews) were reported as data mentions of reuse, but these should be considered not the reuse of data by others. In other cases, algorithmically identified data mentions refer in fact to descriptions of procedures and methods, for instance stating what variables are used in the study, or how they are calculated.

It is clear that most algorithmically extracted data mentions best match the manually extracted mentions of dataset instead of identifiers, references or repositories ([see table below](#) ). SciNoBo has both a higher recall and precision than dataset. The precision and the recall are still relatively low, in the best case reaching only levels of 0.6 to 0.7.

Table 4: Evaluation of performance at the mentions level

	dataset			scinobo		
	Recall	Precision	$F_1$	Recall	Precision	$F_1$
Dataset	0.46	0.42	0.44	0.61	0.60	0.60
Identifier	0	0	0	0.029	0.022	0.025
Reference	0.026	0.018	0.021	0	0	0
Repository	0.05	0.035	0.041	0.059	0.067	0.062
Any of the above	0.47	0.44	0.45	0.68	0.69	0.68

## 2.6.7. Qualitative analysis

Since our case study examines the impact of data availability on research uptake and usage, we wanted to focus our interview study on whether making data accessible influences its utilisation of open data by other scientists and how that varies between repositories. Specifically, we investigate data utilisation in the Netherlands, with preferably a focus on the social sciences and humanities (SSH). To recruit participants for our case study, we reached out to several academic networks. We explained the objectives of our project and provided an overview of our study's scope and aims.

With our interview plan, we aimed to explore data reuse practices among researchers, with a focus on the role of data repositories. It began with a brief introduction to the study's goals; specifically understanding how data repositories influenced the reuse of research data. Some of the questioned concerned current practices in data reuse, including how data is accessed and moderated, challenges regarding reusing data and qualities that make a data repository favourable. With these questions, we aimed to understand both the technical and social aspects of data reuse in the SSH.

So far, we have interviewed one participant - a sociologist and computational social scientist who is an associate professor at a Dutch university. The participant shared their experience

using repositories such as DANS-EASY<sup>13</sup>, LISS Panel<sup>14</sup>, OSF<sup>15</sup>, GitHub<sup>16</sup>, and CBS StatLine<sup>17</sup>. They expressed a clear preference for more curated platforms like DANS, citing its greater reliability and built-in moderation compared to open platforms like GitHub or OSF on which anyone can publish data without any moderation. Key challenges in data reuse included limited access to certain datasets, bureaucratic hurdles, and at times the need for extensive data cleaning and formatting. Overall, the participant strongly advocated for making open data sharing the default in publishing. A realistic suggestion was to have publishers enforce open data policies, while still allowing for legitimate exceptions such as embargoes. This interview was in line with our notion that the quality and trustworthiness of the data repository influenced how much it was used.

## 2.6.8. Interpretation of Results

Here we studied the effects that data repositories can have on subsequent data use by studying data citations and mentions. Our study used the Data Citation Corpus, which is mostly based on publications from the biomedical domain.

### DATA CITATION CORPUS AND DATA MENTIONS ANALYSES

Our results suggest that repositories seem to have some effect on citations. This is somewhat contrary to our expectations when we started this analysis. That is, we expected to find no influence of the repository on subsequent data use at all. Seeing that repositories do have some effect challenges our expectation. It seems to suggest that some data are more likely to be used when shared on a particular repository. Data from some more general repositories, such as Figshare and Mendeley seem less likely to be used, while data from some more specialised and institutional data repositories seem more likely to be used. However, this pattern does not hold in general, as data from some general repositories, such as Zenodo, are more likely to be reused, while data from some specialised repositories are less likely to be reused. Moreover, most repositories do not show any effect on reuse, even if some repositories do show some such effect.

---

<sup>13</sup> DANS is the Dutch national expertise centre and repository for research data, which used to be called EASY.

<sup>14</sup> LISS Panel is a longitudinal research project and database on Dutch society.

<sup>15</sup> OSF, or Open Science Framework, is a free online platform where researchers can share their projects and collaborate.

<sup>16</sup> GitHub is an online platform where developers can store their code, track changes, and collaborate with others.

<sup>17</sup> StatLine is the online database from Statistics Netherlands (CBS), providing a wide range of data on the Dutch economy and society.

It is not entirely clear what mechanism is responsible for causing differences in data reuse from different data repositories. One potential mechanism is that specialised repositories are more likely to be known by potential users, and hence more likely to be explored independently of the associated publication, and thus more likely to be reused. Another potential mechanism is that specialised repositories, especially discipline-specific ones, may offer greater metadata that is useful to explore potential datasets. Yet another mechanism could be related to some form of reputation or status of the repository, possibly related to quality control. What mechanism is behind this should be studied further.

When considering hosting a repository, for instance an institutional repository, it may be valuable to consider what its intended purpose is. If it is meant to stimulate additional visibility, for instance from certain institutions, it might be good to consider whether such effects are likely to be pertained or not. Possibly, data hosted in more general repositories may be less likely to be reused, such that it may be worth hosting a separate repository. On the other hand, it might be more worthwhile to share data in more specialised repositories, decreasing the necessity of hosting a separate repository for reasons of visibility.

Although our study suggests some association between citations and repository, it cannot be excluded that the identified association does not represent a causal effect. In particular, we have not controlled for the country here, which we earlier identified as one potential confounder, for instance through institutional repositories. Moreover, the size of certain datasets could also play a potential confounding role, but we are unable to control for it. An additional problem, and presumably a larger one, is that there are relatively few data citations and mentions, and so the estimates may suffer from some selection bias towards datasets that happen to be cited. Finally, the other associations, with field and year, should not be interpreted causally, as they were not the main focus of the analysis, and most likely do not represent causal effects.

## INTERVIEW STUDY

So far, we have interviewed one participant - a sociologist and computational social scientist who is an associate professor at a Dutch university. The participant shared their experience using repositories such as DANS-EASY<sup>4</sup>, LISS Panel<sup>5</sup>, OSF<sup>6</sup>, GitHub<sup>7</sup>, and CBS StatLine<sup>8</sup>. They expressed a clear preference for more curated platforms like DANS, citing its greater reliability and built-in moderation compared to open platforms like GitHub or OSF on which anyone can publish data without any moderation. Key challenges in data reuse included limited access to certain datasets, bureaucratic hurdles, and at times the need for extensive data cleaning and formatting. Overall, the participant strongly advocated for making open data sharing the default in publishing. A realistic suggestion was to have publishers enforce open data policies, while still allowing for legitimate exceptions such as embargoes. This interview was in line with our notion that the quality and trustworthiness of the data repository influenced how much it was used.

## 2.6.9. Conclusions

Open data has become increasingly important over the past decade. There has been increasing attention to the reuse of data. There have also been increasing efforts in order to try to algorithmically extract data mentions from publications, also in order to monitor developments in open data. These indicators are feeding into policy discussions about open data. Knowing the reliability and accuracy of these indicators is therefore important.

There is a great variation in data sharing and data usage across fields. In this paper we analysed data usage specifically in the social sciences and humanities, and studied how our manually extracted data mentions compared with the algorithmically extracted data mentions. We found that algorithmically extracted data mentions show a high recall (0.96 at best) of whether any data was reused in an article. This means that most publications with any data reuse will most likely be identified through algorithms. However, the precision of the algorithmic approach is much lower (0.46 at best), suggesting that more than half of the articles are incorrectly suggested by algorithms to show data reuse.

This finding suggests that the reportedly high levels of data use in the social sciences are most likely to be overestimates. For instance, the French Open Science Monitor reports that 60% of the publications from 2023 in the social sciences and 40% in the humanities show any data use<sup>5</sup>. A recent report on data mentions in the Dutch-funded research suggests that 83% of the publications from 2023 show some data use. Given our results here, these are likely to be overestimates of data reuse, and are roughly twice as high as in reality. Indeed, estimates of roughly 30-40% seem more realistic, and are also more in line with our manual extraction of data mentions.

Algorithmically derived indicators of data use should hence be interpreted with some restraint in the social sciences and humanities. Absolute interpretations, e.g. 83% of the publications show data use, are most likely not very accurate, and, as said, are presumably overestimates. Nonetheless, comparing such figures over time, such as a growth from 81% to 83% from 2022 to 2023 might provide some indication of the overall trend. Algorithmically derived indicator of data use may therefore be used to study the development over time. Comparing these numbers across disciplines might not be warranted. We now showed that the precision in the social sciences and humanities is not very high. Most likely, the precision will not be the same across domains, given also differences in data use and practices across domains. If precision indeed varies across domains, then comparing the resulting algorithmically derived indicators across domains is not accurate.

When going down to the individual detection of what datasets are mentioned, we note that results are not usable at this point. At the sentence level, the recall is already fairly low (about 0.40), but the precision is outright low (about 0.05). This means that only roughly 1 out of every

20 identified sentences with data reuse is correct. For those 1 out of 20 correctly identified sentences, about two-third of the data mentions are correctly inferred. Overall, roughly 3% of the algorithmically extracted data mentions can be considered correct. In addition, the algorithmically extracted data mentions miss more than half of the sentences that mention data reuse. In sentences that are correctly identified, it still misses about one-third of the data mentions. Only about a quarter of the data mentioned is correctly found algorithmically. This implies that studies cannot confidently rely on the extraction of individual data mentions in the social sciences and humanities.

Our qualitative study on the repository effect, while currently limited in its sample size, indicates that the specific repository used to deposit data may have an effect on its reuse. While openness is appreciated highly, repositories that have some metadata and deposit moderation proved more favourable for data reuse practices. This also underlines the preference for national data repositories as a central point for data sharing and re-use practices in SSH.

This contrasts with other domains, such as the biomedical domain, where data identifiers are more regularly used. For instance, the Global Data Citation Corpus relies on extracting identifiers from full-text, working on publications predominantly from the biomedical domains. In the social sciences and humanities, this approach will not work. In addition, much of the data mentioned seems to originate from governmental agencies, such as bureaus of statistics and ministries. If such datasets had permanent identifiers, perhaps it would allow tracing data uses in the social sciences and humanities more accurately. However, this requires efforts beyond academia itself, and involves government more broadly, relating to initiatives such as open government data.

## 3. Cross-Case Synthesis and Reflections

### 3.1. Framing the synthesis

The six case studies collectively provide grounded insights into how specific Open Science practices develop into academic, societal, and economic effects. Each case follows a pathway perspective, tracing the sequence from inputs and practices to outputs, outcomes, and eventual impacts, while considering the enablers, barriers, and feedback loops that shape this process. This synthesis brings together these findings to highlight consistent patterns across contexts, areas of divergence, and factors that appear to amplify or moderate impact.

The cases differ substantially in scope and empirical setting. They encompass:

- national infrastructures (Portuguese Repository Infrastructure RCAAP),
- thematic infrastructures (ELIXIR's Bioinformatics Resources),
- global emergencies (COVID-19 artefact reuse),
- field-specific dynamics (Effects of Data Repositories on Data Usage),
- cross-platform usage (French Open Access Infrastructure), and
- repository effects on data reuse (Effects of Data Repositories on Data Usage).

Taken together, they span a wide spectrum of disciplines, sectors, and institutional configurations. The diversity of contexts is valuable because it reveals both recurring impact pathways and context-dependent variations. Some practices, such as repository deposition or structured artifact reuse, display relatively consistent associations with downstream indicators, while others, such as journal-mediated openness, show more nuanced or mixed patterns depending on field and outcome.

By combining large-scale quantitative evidence with qualitative perspectives, the case studies also illustrate the importance of context-sensitive amplifiers: metadata quality, repository curation, collaboration incentives, and early dissemination all emerge as factors that shape the strength and direction of Open Science effects. The synthesis therefore advances the understanding that impacts cannot be attributed to openness alone, but are the result of interactions between practices, infrastructures, and broader systemic conditions.

Importantly, this synthesis should be interpreted within the work package's primary objective: operationalizing and testing impact indicators for Open Science evaluation. PathOS aimed to develop robust methodological approaches, apply big data, AI and mixed-methods designs, create new indicators and tools, and implement the strongest feasible causal identification strategies within observational constraints. The value lies not only in the specific findings but in

demonstrating how systematic impact assessment can be conducted across diverse OS contexts.

## 3.2. Common signals across cases

Across the six case studies, several patterns emerge that cut across differences in field, geography, and type of infrastructure. These signals suggest recurring pathways by which Open Science practices generate academic, economic, and societal effects, while also highlighting the contextual factors that shape their strength. However, the effects of openness are not uniformly positive. While most observed effects are positive, some appear neutral or context-dependent, reminding us that openness interacts with quality, visibility, and timing rather than operating as a uniform driver of impact.

**Repositories and infrastructures as enablers** Structured infrastructures appear repeatedly as decisive enablers of visibility, persistence, and consultation. Whether in national systems such as RCAAP, thematic infrastructures such as ELIXIR, or field-specific analyses of Green Open Access, the presence of repositories and platforms is associated with greater discoverability. However, the relationship with thematic endurance varies by access route, repository-based sharing (Green OA) showed positive effects on topic persistence, while journal-mediated openness (Published OA) showed negative effects. The French case further shows that Open Science platforms attract significant traffic from outside academia, underscoring their role in bridging scientific and societal domains.

**Reusability and practical uptake** The cases focused on COVID-19 and dataset reuse both demonstrate that impact depends not only on access but on the conditions that make research outputs practically reusable. Explicit referencing, traceability, and structured metadata emerge as crucial for enabling downstream use. The ELIXIR case similarly highlights how open resources, when aligned with FAIR principles, become integrated into innovation processes, including patenting activity. These observations suggest that openness translates into impact most effectively when combined with usability, curation, and quality assurance.

**Collaboration as a recurring pathway** Collaboration across sectors appears in several contexts as a pathway amplified by Open Science. Evidence from ELIXIR patents, RCAAP-supported publications, and Green Open Access in AI-and-Climate research indicates that open infrastructures can foster links between academia and industry. While the intensity of these collaborations differs across cases, their recurring presence points to a broader role for openness in enabling cross-sector partnerships that support both research progress and innovation.

**Contextual amplifiers** The case studies also underscore that Open Science effects are rarely uniform. Metadata quality, repository curation, early dissemination, and disciplinary norms emerge as amplifiers that strengthen or weaken impact pathways. For instance, curated repositories in the Netherlands case appear more conducive to data reuse than general-purpose platforms, while early dissemination during the COVID-19 crisis enhanced the benefits of artifact sharing. These examples illustrate that the impact of Open Science depends on the interaction between practices and their systemic environment.

Taken together, these signals indicate that openness is most effective when embedded within infrastructures that ensure visibility, traceability, and usability, and when accompanied by systemic supports that enhance collaboration and quality.

### 3.3. Amplifiers and moderators

The effects of Open Science practices identified across the case studies are rarely uniform. Instead, they are mediated by amplifiers and moderators that shape whether openness translates into sustained academic, societal, and economic impact. These contextual factors help explain why similar practices yield strong effects in some environments but limited or mixed results in others.

**Timing and early dissemination** emerge as central. In the COVID-19 case, early availability of datasets and software increased opportunities for reuse and downstream uptake, particularly in industry and technology transfer. Later outputs, although open, were less likely to secure similar influence, indicating that timeliness can magnify or diminish the benefits of sharing.

**Metadata quality and repository curation** act as further amplifiers. The Netherlands case showed that repositories with structured metadata and active moderation were more conducive to data reuse than broad, lightly curated platforms. In the RCAAP system, the emphasis on metadata validation and harmonisation enhanced findability and interoperability, positioning repositories as key enablers of visibility and discoverability.

**Sectoral and geographical reach** also functions as a moderator. The French case revealed that open platforms such as HAL and OpenEdition attract significant traffic beyond academia, including from public administration and industry, but with notable variations across countries and sectors. This indicates that the benefits of openness are not evenly distributed, but mediated by who is able to access, search, and use the available outputs.

**Collaboration** is another amplifier consistently observed across cases. The ELIXIR study demonstrated that open bioinformatics resources support patenting and innovation, with

stronger effects when industry and academic actors collaborate. Similarly, RCAAP outputs showed relatively greater impact on domestic collaborations between universities and companies. In AI-and-Climate research, Green Open Access was linked to more durable thematic persistence and stronger science–industry connections.

These observations highlight that openness alone is not a sufficient condition for impact. Rather, the combination of infrastructure design, disciplinary norms, quality assurance, and policy environment determines the magnitude and direction of Open Science effects. Amplifiers such as timing, metadata, and collaboration enhance the likelihood that open practices move beyond access to generate lasting academic visibility, societal relevance, and economic innovation.

## 3.4. Reflections on causality

The six case studies employed a variety of approaches to address the challenge of causal attribution in Open Science. Propensity-score matching was applied in the AI and Climate case to isolate the effects of Green and Published Open Access. Regression with interaction terms was used in the COVID-19 study to test how reuse effects varied by quality, visibility, and timing. The Netherlands case explored repository effects through a combination of data citation corpus analysis and qualitative interviews. The French case advanced an exploratory approach, leveraging server log data to develop a tool for investigating sectoral and geographical patterns of access. The ELIXIR case employed counterfactual reasoning to assess how the absence of bioinformatics resources might affect industry innovation and research progress. The RCAAP case combined quantitative contrasts between repository and non-repository publications with qualitative scenarios that asked stakeholders to consider a landscape without a national repository infrastructure.

Each design captures only part of the causal picture but contributes to building a more transparent evidence base. Across cases, the emphasis on clarity of assumptions and acknowledgement of residual uncertainties emerges as an important signal. Observational methods cannot remove all potential biases, but designs that combine clean treatment definitions, balance diagnostics, and clear temporal sequencing strengthen the credibility of findings.

A further insight is that Open Science practices often operate as amplifiers rather than substitutes. The benefits of openness tend to be strongest when research outputs are already characterised by high quality, strong visibility, or early dissemination. Conversely, openness alone is rarely sufficient to generate lasting influence if these other enabling conditions are absent. This pattern was visible in the COVID-19 case, where artifact reuse reinforced already-

visible and early contributions, and in the AI-and-Climate case, where repository self-archiving prolonged topic persistence particularly for themes with existing scholarly momentum.

The lesson across designs is that causal claims in Open Science must be framed carefully. Transparent recognition of methodological limits is as important as reporting positive results. When linked to clear causal pathways, the evidence becomes usable for funders, infrastructures, and policy makers, even when definitive proof is unattainable. Open Science effects are therefore best interpreted not as universal laws but as context-dependent mechanisms whose strength and direction depend on infrastructures, practices, and systemic environments. The methodological approaches developed and tested across these cases, including new indicators, big data analytics, mixed-methods designs, and causal identification strategies, represent advances in how OS impact can be systematically assessed, even when findings challenge conventional expectations about openness benefits.

### 3.5. Lessons for indicators and monitoring

The application of PathOS indicators across six diverse case studies demonstrates both the versatility of indicator-based monitoring and the challenges that arise when they are transferred into real-world contexts. Standard bibliometric and translational indicators, such as citation impact, field-weighted citation scores, industry co-authorship, and patent references, proved effective in capturing dimensions of academic and economic impact. Newer constructs, including topic persistence and cost-benefit estimation, expanded the capacity to assess longer-term and systemic effects. The French case further illustrated the potential of access-related indicators, such as the Open Access advantage derived from server log analysis, to capture societal uptake beyond academia.

At the same time, the limitations of indicators became visible. Automated detection of data mentions in the social sciences and humanities revealed high recall but low precision, meaning that broad trends can be tracked, but detailed or cross-domain comparisons risk misinterpretation. Similarly, indicators tied to clinical uptake in the COVID-19 case were sensitive to contextual conditions such as timing and research visibility, underscoring that indicator performance cannot be assumed to be uniform across fields or stages of research.

Several cross-cutting lessons emerge:

- **Alignment with policy questions:** Indicators are most informative when they are explicitly matched to a well-defined policy or evaluation question. Starting with the question, rather than the metric, makes gaps and limitations visible and prevents over-reliance on partial signals.

- **Dependence on data quality:** The effectiveness of indicators depends strongly on the completeness, accuracy, and curation of underlying datasets. Metadata quality, repository practices, and harmonisation procedures directly influence indicator robustness.
- **Need for triangulation:** Quantitative signals alone are insufficient. Qualitative insights, stakeholder perspectives, and scenario exercises are critical to interpret indicator results and identify mechanisms behind observed patterns.
- **Context sensitivity:** The same indicator can yield divergent insights depending on disciplinary norms, timing of dissemination, or geographical and sectoral environments. Indicators therefore require contextual interpretation rather than universal benchmarks.
- **Transparency and limits:** Communicating the assumptions, scope, and uncertainties attached to each indicator is essential for their credibility. Transparent recognition of what an indicator does not capture can be as important as reporting positive results.

Taken together, the case studies suggest that indicators can meaningfully inform policy design and evaluation when embedded in mixed-methods approaches, when grounded in high-quality data, and when their interpretive limits are acknowledged. They work best not as stand-alone measures, but as components of a broader evidence base that combines quantitative monitoring with qualitative understanding of how Open Science practices unfold in practice.

## 3.6. Implications for infrastructures and policy

The systematic application of impact indicators across six diverse case studies demonstrates that infrastructures play a decisive role in shaping how Open Science practices translate into academic, societal, and economic impact. Repositories, platforms, and open data resources emerge not only as technical compliance mechanisms but as amplifiers of persistence, visibility, and reuse. Their capacity to extend the longevity of research topics, to sustain discoverability, and to facilitate cross-sector uptake underscores the importance of sustained investment in their quality and reliability. Metadata enrichment, repository curation, and alignment with international standards are repeatedly highlighted as enablers of stronger outcomes.

In domains characterised by urgency or rapid change, such as health crises or climate research, timing becomes a central determinant of impact. The availability of data, code, and publications at an early stage appears to condition the extent to which openness translates into reuse and downstream influence. Infrastructure design and policy frameworks therefore need to prioritise mechanisms that support timely deposition and dissemination, without compromising standards of quality and traceability.

Another recurrent pattern across cases is the role of collaboration. Open infrastructures act as connectors between academia, industry, and other societal sectors, but these connections do not arise automatically. Where partnerships are present, evidence points to stronger innovation outcomes and wider uptake. Policies that incentivise collaboration, encourage shared standards, and reduce barriers for non-academic actors can help transform openness into concrete forms of knowledge transfer.

For policy makers and funders, the overarching implication is that investments in Open Science infrastructures should be understood as investments in enabling conditions, rather than as end goals in themselves. Their value lies in the way they support amplification effects: extending topic persistence, strengthening industrial engagement, broadening access across sectors, and ensuring that research outputs remain reusable over time. Balanced support for multiple routes of openness, combined with attention to metadata quality, timing, and cross-sector incentives, appears most conducive to realising the broader promise of Open Science.

## 3.7. Open questions and future directions

The cross-case synthesis makes visible several areas where important uncertainties remain. Some patterns are clear across multiple contexts, while others point to unresolved questions that call for further enquiry.

A first open question concerns the mechanisms underlying differences between routes of Open Access. In the AI and Climate case, repository-based self-archiving extended research topic vitality, whereas journal-mediated openness showed significant negative effects on long-term persistence. The reasons for this divergence are not yet fully understood. It remains to be explored whether the discoverability of repository deposits, the metadata practices that accompany them, or differences in how search engines and platforms index various versions play a role in shaping these outcomes.

Another area of uncertainty relates to the translation of research into clinical practice. In the COVID-19 case, documented reuse of datasets and software was positively linked to technological uptake and industrial collaboration. This raises questions about whether clinical adoption depends more on perceived authority and dissemination channels than on technical reusability. Understanding these dynamics is particularly relevant for fields where research feeds directly into decision-making under urgent conditions.

The repository study in the Netherlands highlighted further methodological challenges. Automated extraction of data mentions in the social sciences and humanities produced broad signals but lacked precision, raising concerns about how best to capture reuse in domains with diverse publication practices and limited identifier adoption. Addressing this gap will require

new methods that can combine computational detection with curated validation and may also depend on the wider uptake of persistent identifiers by data providers such as government agencies.

Future directions for research and policy include:

- Refining indicators to better capture context-sensitive forms of reuse, translation, and persistence.
- Extending cost-benefit analysis frameworks to additional infrastructures, allowing a clearer picture of returns to public investment.
- Deepening the integration of qualitative evidence, such as interviews, focus groups, and scenario testing, to complement quantitative monitoring.
- Exploring the role of amplifiers such as early dissemination, metadata quality, or cross-sector collaboration in shaping whether openness leads to measurable impact.

Across all cases, the lesson is that Open Science practices do not operate in isolation but within broader ecosystems of infrastructures, incentives, and disciplinary norms. Clarifying the conditions under which openness translates into lasting benefits remains an important task for both future research and policy development.

## 4. References

Besançon, L., Peiffer-Smadja, N., Segalas, C., Jiang, H., Masuzzo, P., Smout, C., ... & Leyrat, C. (2021). Open science saves lives: lessons from the COVID-19 pandemic. *BMC Medical Research Methodology*, 21(1), 117.

Catalano, G., Colnot, L., Vignetti, S., Correia, A., Príncipe, P., Lopes, P., & Seminaroti, E. (2025). RCAAP case study. Zenodo. <https://doi.org/10.5281/zenodo.15732145>

Chen, X., Bharti, N., & Marsteller, M. R. (2021). Use of Bibliometrics Data to Understand the Citation Advantages of Different Open Access Categories in Covid - 19 Related Studies. *Proceedings of the Association for Information Science and Technology*, 58(1), 410–414. <https://doi.org/10.1002/pr2.469>

Chițu, F., Mecu, A.-N., & Marin, G.-I. (2024). Exploring the Climate Change–AI nexus: A bibliometric and scientometric study. *Proceedings of the International Conference on Business Excellence*, 18, 1658–1670.

Clayson, P. E., Baldwin, S. A., & Larson, M. J. (2021). The open access advantage for studies of human electrophysiology: Impact on citations and Altmetrics. *International Journal of Psychophysiology*, 164, 103–111. <https://doi.org/10.1016/j.ijpsycho.2021.03.006>

De Filippo, D., & Mañana-Rodríguez, J. (2020). Open access initiatives in European universities: Analysis of their implementation and the visibility of publications in the YERUN network. *Scientometrics*, 125(3), 2667–2694. <https://doi.org/10.1007/s11192-020-03705-0>

Eger, T., Mertens, A., & Scheufen, M. (2021). Publication cultures and the citation impact of open access. *Managerial and Decision Economics*, 42(8), 1980–1998. <https://doi.org/10.1002/mde.3429>

Fraser, N., Brierley, L., Dey, G., Polka, J. K., Pálffy, M., Nanni, F., & Coates, J. A. (2021). The evolving role of preprints in the dissemination of COVID-19 research. *PLOS Biology*, 19(4), e3000959.

Harrison, P. W., Lopez, R., Rahman, N., Allen, S. G., Aslam, R., Buso, N., ... & Apweiler, R. (2021). The COVID-19 Data Portal: accelerating SARS-CoV-2 and COVID-19 research through rapid open access data sharing. *Nucleic Acids Research*, 49(W1), W619-W623.

Huang, C.-K., Neylon, C., Brody, T., Xia, J., & Ratan, N. (2024). Open access research outputs receive more diverse citations. *Scientometrics*, 129, 825–845.

Klebel, T., Cole, N. L., Tsipouri, L., Kormann, E., Karasz, I., Liarti, S., Stoy, L., Traag, V., Vignetti, S., & Ross-Hellauer, T. (2023). PathOS – D1.2 Scoping Review of Open Science Impact (Commissioned report). Zenodo. <https://doi.org/10.5281/zenodo.7883699>

Kobayashi, S., Falcón, L., Fraser, H., Braa, J., Amarakoon, P., Marcelo, A., & Paton, C. (2021). Using open source, open data, and civic technology to address the COVID-19 pandemic and infodemic. *Yearbook of Medical Informatics*, 30, 38-43.

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>

Piwo war, 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>

Piwo war, H., Priem, J., Larivière, V., Alperin, J. P., Matthias, L., Norlander, B., ... Haustein, S. (2018). The state of OA: A large-scale analysis of the prevalence and impact of open access articles. *PeerJ*, 6, e4375.

Sofi-Mahmudi, A., Raittio, E., & Uribe, S. E. (2023). Transparency of COVID-19-related research: a meta-research study. *PLOS ONE*, 18(7), e0288406.

Tse, E. G., Klug, D. M., & Todd, M. H. (2020). Open science approaches to COVID-19. *F1000Research*, 9, 1043.

Valenzuela-Escarcega, M.A., Ha, V.A., & Etzioni, O. (2015). Identifying Meaningful Citations. AAAI Workshop: Scholarly Big Data.

Wang, D., Song, C., & Barabási, A.-L. (2013). Quantifying long-term scientific impact. *Science*, 342(6154), 127–132.

Wang, L. L., Lo, K., Chandrasekhar, Y., Reas, R., Yang, J., Burdick, D., ... & Beltagy, I. (2020). CORD-19: the COVID-19 open research dataset. *arXiv preprint arXiv:2004.10706*.

Young, J. S., & Brandes, P. M. (2020). Green and gold open access citation and interdisciplinary advantage: A bibliometric study of two science journals. *The Journal of Academic Librarianship*, 46(2), 102105. <https://doi.org/10.1016/j.acalib.2019.102105>

## 5. Annexes

### 5.1. Data Mentions Study Interview Plan

#### 1. Introduction (5 minutes)

- Check whether participant received and signed the consent form.
- Obtain consent for recording, only for purpose of transcribing interview. Without consent, we will just take notes / transcribe manually.
- Briefly introduce yourself, your research, and the purpose of the interview.

The goal of the case study: To understand better how data repositories affect data reuse. We are interested in understanding how people come to reuse data for research, and if data repositories have any effect on whether data is reused or not. For example, data repositories may play a role in the visibility of certain datasets, or data repositories may have a certain reputation, leading people to trust datasets from such a repository.

We broadly define data as any evidence of phenomena for the purpose of research. This may range from spreadsheets or databases to text corpora or interview transcripts.

Data repositories are usually understood as repositories where researchers have deposited research data (openly). We here take a broader view, and also consider official government platforms, bureaus of statistics or data from NGOs or IGOs, since data in the social sciences often seem to come from these data sources.

- Outline the structure of the interview and reassure them about confidentiality and data usage.

#### 2. Background Information (5 minutes)

- “Can you describe your current role and area of research in a few sentences?”

- “How long have you been working in this field?”
- “Can you briefly describe what data you typically use in your research?”

### 3. Current Data Reuse Practices (20 minutes)

- “Can you describe one example of how you reused a dataset for research?”
- “Can you describe how you typically access data for your research?”
  - Offer the interviewee the possibility to demonstrate something.
- “How do you find datasets created by others?”
  - Through personal connections
  - Through journal publications
  - Through data repositories
  - ...
- “Can you name some data repositories that you are familiar with?”
- “Do you visit data repositories to find data for your research?”
- “How do you evaluate the quality or reliability of datasets shared by others?”
- “Are there some data repositories you consider more reliable or trustworthy?”

### 4. Challenges and Opportunities (15 minutes)

- “Have you encountered any difficulties when trying to find data to reuse in your research?”
- “Have you encountered any challenges when trying to reuse data for research?”
- “Are there particular formats, standards, or metadata that make it easier for you to reuse data for research?”
- “Did data repositories help you find or reuse data for research?”
- “What could make the process of finding or reusing data for research easier?”

### 5. Closing (5 minutes)

- “Is there anything else I haven’t asked about that you want to share?”
- Thank the participant for their time and insights.
- Ask if they have any additional thoughts or questions.

## 5.2. PathOS Case Studies Factsheet

This factsheet<sup>18</sup> provides a concise overview of the six PathOS case studies. Each case study models a distinct Open Science pathway, examining how specific practices (e.g., Open Access,

---

<sup>18</sup>[https://pathos-project.eu/images/Case\\_Studies/Unlocking%20the%20Insights%20from%20PathOS%20Case%20Studies%202.pdf](https://pathos-project.eu/images/Case_Studies/Unlocking%20the%20Insights%20from%20PathOS%20Case%20Studies%202.pdf)

Data Reuse, Repository Infrastructures, and Open Resources) produce academic, economic, and societal effects.

The factsheet summarises:

- The **objectives and scope** of each case study within the PathOS framework.
- The **methodological approaches** used to assess Open Science practices and their outcomes.
- The **main data sources and analytical tools** applied across cases.
- The **indicator framework** underpinning the PathOS evidence base, aligned with the PathOS Indicator Handbook.

It serves as a high-level reference for understanding how the case studies collectively contribute to mapping the pathways through which Open Science generates impact.