Open Science Impact Pathways

Deliverable 3.1

Case studies for evaluation of open science impact

Deliverable Number and Name	D3.1 Case studies for evaluation of Open Science impact
Due Date	October 31, 2023
Delivery Date	October 31, 2023
Work Package	WP3
Type	Report
Author	Nicki Lisa Cole, Simon Apartis, Erika Balsyte, Antonia Correia, Clare Garrard, Ioanna Grypari, Corinne Martin, Haris Papageorgiou, Pedro Principe, Despoina Sousoni, Petros Stavropoulos, Vincent Traag, Tommaso Venturini, Tim Willemse
Reviewers	Vincent Traag, Tommaso Venturini, Ioanna Grypari
Approved by	Ioanna Grypari
Dissemination Level	PU – Public
Version	1.0
Number of Pages	62

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
0.0	02/10/2023	First draft	NLC, CM, EB, DS, CG, TW, VT, PP, AC, PS, IG, HP, TV, SA
0.1	16/10/2023	Peer review	VT, TV, IG
0.2	23/10/2023	Revised draft	NLC, CM, EB, DS, CG, TW, VT, PP, AC, PS, IG, HP, TV, SA
0.9	25/10/2023	Final version after formatting and proofreading	NLC
1.0	31/10/2023	Final version after minor editing from project coordinator	NLC

Table 1: Document Revision History

Table of Contents

List of Tables

List of Figures

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

AI	Artificial Intelligence
APC	Article Processing Charge
CBA	Cost Benefit Analysis
CCSD	Centre pour la Communication Scientifique Directe
CNRS	Centre National de la Recherche Scientifique
CSOs	Civil Society Organisations
CWTS	Centre for Science and Technology Studies
DANS	Data Station Social Sciences and Humanities
EC	European Commission
EOSC	European Open Science Cloud
EU	European Union
FAIR	Findable, Accessible, Interoperable, Reusable
FCCN	Fundação para a Computação Científica Nacional
FCT	Fundação para a Ciência e Tecnologia
H2020	Horizon 2020
HAL	Hyper Article en Ligne
INIST	Institut National de l'Information Scientifique et Technique
INSEE	Institut National de la Statistique et des Etudes Economiques
ISP	Internet Service Provider
OA	Open Access
OD	Open Data
OM	Open materials
OS	Open Science
OSC	Open-Source Code
RCAAP	Repositórios Científicos de Acesso Aberto de Portugal
SARI	Serviço de Alojamento de Repositórios Institucionais
SME	Small and Medium-sized Enterprises (SMEs)
SSH	Social Sciences and Humanities
WOIS	Who is?

Executive Summary

PathOS is a Horizon Europe project aiming to gather concrete evidence of the impacts of Open Science (OS). It seeks to understand the progression from input to output, outcome, and eventual impact, taking into account both enabling factors and critical barriers. Recognizing and comprehending OS pathways is vital not only to estimate and measure the effects of a policy intervention but also to elucidate why and how these impacts arise.

PathOS centres its activities around six targeted case studies that will (a) drive the modelling of the pathways with all relevant elements of OS which can be measured both in terms of indicators for input and costs, and by making their connections, (b) support testing and operationalization of OS indicators (tools, data, flows) by providing access to data and experts who bring knowledge of the local environment, and (c) provide input to the Cost Benefit Analysis and to validate project results. The foci of the case studies include: 1. accelerating collaborations within academia & industry in Portugal; 2. research data and knowledge use outside of academia in France; 3. cross cutting effects due to Open Research data from a National Repository in the Netherlands; 4. Emerging topics fostered by Open Science with a focus on gender in AI and climate innovations; 5. Open Science practices during the COVID-19 pandemic within Horizon 2020 projects; and 6. Innovation from Open Research resources provided by ELIXIR.

The purpose of this deliverable is to report on the status to date of the case studies, about one-third of the way into the project's duration. This report includes a chapter dedicated to each case study which describes the development of the case studies in terms of research questions, causality narrative and impact pathway logic, discussion of enabling factors and barriers, research methods, and immediate next steps. As co-creation with expert stakeholders is a core practice of PathOS, the cases that have hosted focus groups to date (3 out of 6) report on how the insights gained have contributed to the development of the case study and supported the project's development of impact pathway models and impact indicators.

Together, these case studies focus on identifying impacts from four key Open Science aspects: Open Access publishing, Open/FAIR Data, Open-Source Code, and Open Materials. Cumulatively, the case studies cover government/policy mandates and government-sponsored portals/repositories, and the short-term impacts they aim to measure include academic citation advantage (from Open Access and Open Data), data reuse, and collaborations; the economic impacts of academic-industry collaborations and the industry use of Open Science resources; and the societal impacts of the use of OS resources by various societal actors (policymakers, media, civil society organisations, healthcare providers, etc.).

1. Introduction

1.1. Project overview

PathOS aims to identify and document Open Science Impact Pathways — the possible paths that connect the input of Open Science (OS) interventions to research output, outcome, and ultimately impact (academic, societal and economic). The project additional aims to identify the causal mechanisms that link these, as well as factors that enable or block the creation of OS impact.

Building on the RI-PATHS impact pathways model (Griniece et al., 2020), the PathOS model (Figure 1) conceives that the combination of resources and activities generates outputs, which then lead to short-term and long-term outcomes, and ultimately, to impacts.

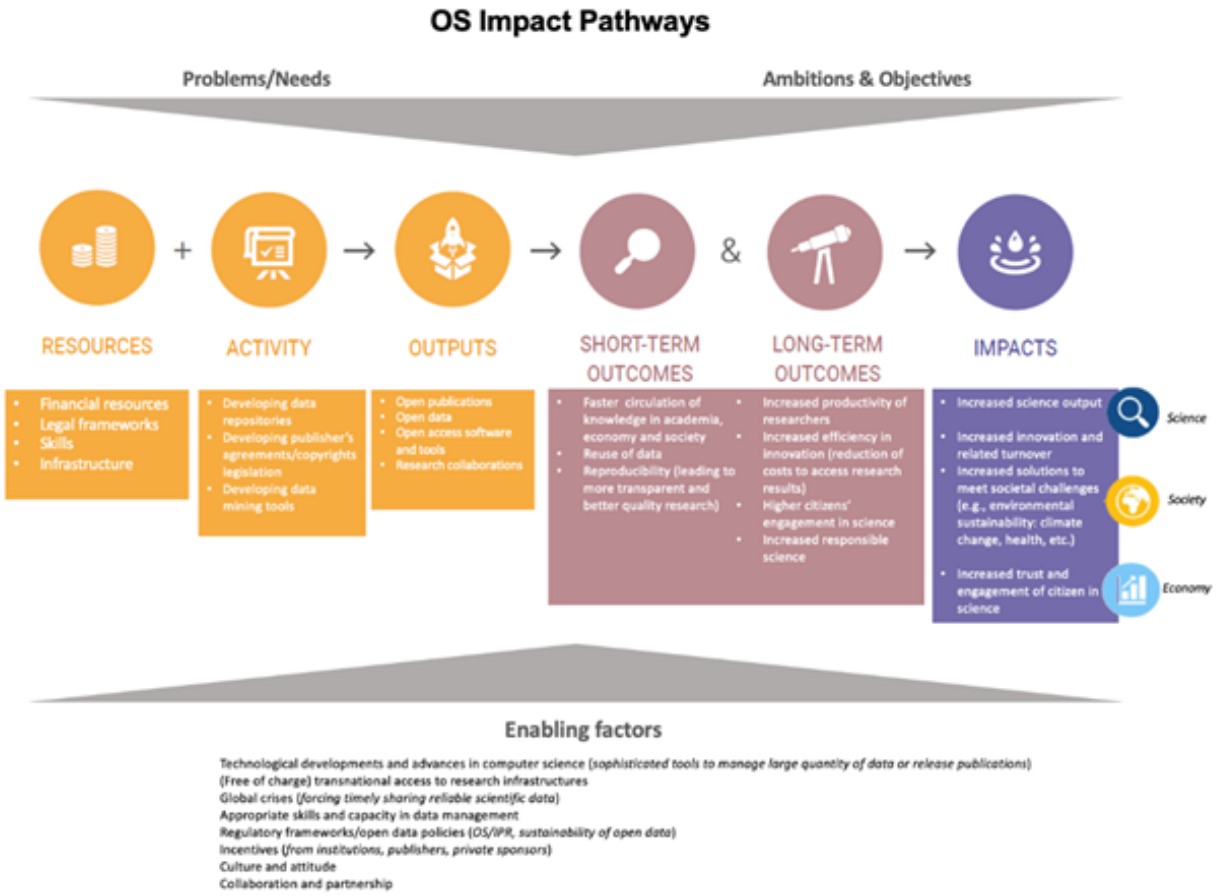

Figure 1: Draft PathOS OS impact pathways (Source: PathOS Description of Action)

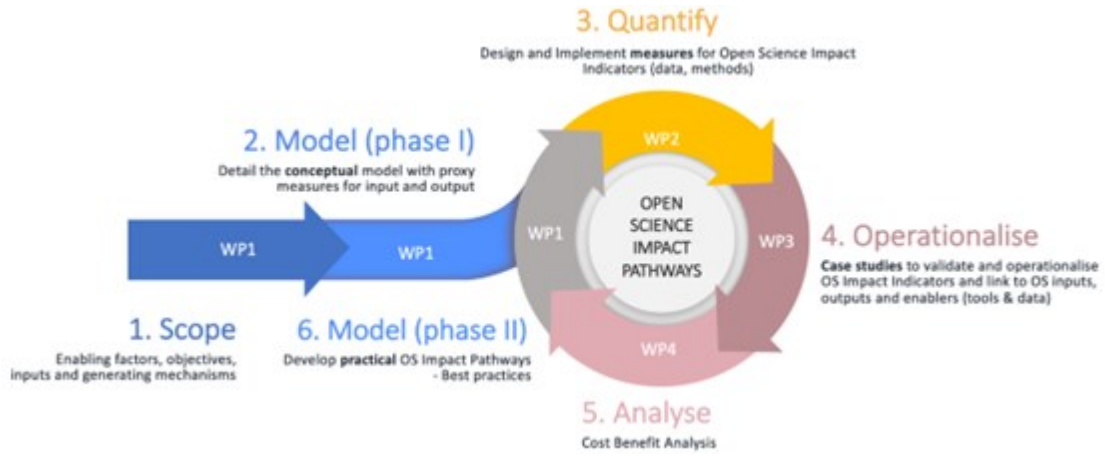

Figure 2: PathOS methodological steps

PathOS will identify Open Science Impact Pathways, their causal mechanisms and enabling and blocking factors through an iterative, co-creative process. This process, demonstrated in Figure 2, began (1) by scoping the existing evidence of academic, societal and economic impacts of Open Science (Klebel et al., 2023). Simultaneously, we (2) developed a conceptual model of Open Science Impact Pathways based on current approaches to science policy evaluation and Theory of Change (Dekker et al., 2023). Our (2) modelling, (3) quantification, (4) operationalization and (5) analysis of Open Science Impact Pathways are conducted in a phased approach and centre around seven case studies. Each case study focuses on a specific range of OS elements operationalized in a particular context and serves as the mechanisms through which PathOS develops and tests a set of tools to measure and analyse OS Impact and to link them (causally) to enabling factors using structural causal models (Pearl & Mackenzie, 2019) as a thinking tool.

1.2. Case studies approach and overview

The case studies have been developed to provide an end-to-end story, reflecting a variety of established OS practices at institutional and national levels and focusing on cross-cutting (e.g., scientific collaboration, gender, reproducibility) and downstream (e.g., acceleration of innovation, health and environmental) impacts.

Taking a co-creative approach with carefully selected expert stakeholders from across the R&I ecosystem, with the case studies we concretely capture all elements of OS that are relevant and can be measured (quantified both in terms of indicators for input and costs), test and operationalize OS impact indicators (tools, data, flows), and provide concrete input to the Cost Benefit Analysis process to help establish a framework to validate results.

The six case studies, detailed in subsequent chapters, include three national case studies for France, Portugal and Netherlands, all of which have an active and established OS ecosystem

consisting of strong policies, mature infrastructures, regulatory frameworks, and engaged research communities. The **French case study** focuses on the broader societal use and impacts of research data and knowledge facilitated by the national infrastructure for OS. The **Portuguese case study** focuses on the use of the national OS publication repository infrastructure and knowledge diffusion via academic-industry collaborations. The **Dutch case study** focuses on the academic impacts of research data availability, facilitated by a national data repository, and examines whether the national repository has a different effect than other types of repositories.

Two of our case studies are European-wide in their scope and study the effects of Horizon 2020 OS interventions on scientific research during the **global COVID-19 pandemic**, on **climate science**, and **on gender-related issues in AI**.

Our sixth case study is domain-specific and focuses on the use of OS resources by the **bioinformatics industry**, facilitated by the ELIXIR OS platform and resources.

1.3. Co-creation with expert stakeholders

In PathOS we take a co-creative approach with expert, context-relevant stakeholders to inform the development of each case study. These stakeholders include institutional leaders and research managers, industry, infrastructure providers, meta-science and scientometric experts, research funders, policy analysts, and domain-relevant researchers.

These stakeholders are engaged through a series of case-specific focus groups with 4-8 people designed to draw on their experience and expertise in developing the foci and methods of the cases, the development of indicators, and the modelling of impact pathways. The stakeholder groups created for each case will follow the results of the project and meet on a regular basis (3-4 times). The first series of focus groups, occurring from spring through autumn 2023, are organized as a needs assessment exercise to develop the case-specific conceptual pathway model and provide feedback and insights on how it fits their context. The aim of this exercise is to identify expert-informed, targeted pathways and to facilitate the modelling process.

A second series of focus groups, conducted from winter through spring 2023-24, will focus on the development of impact indicators and identify themes and areas that might be missing in the case-specific impact pathway model. A third series will be conducted in autumn 2024 to facilitate a cost-benefit analysis (CBA) process for selected case studies, and a final series will be conducted for all cases during spring 2025 to assess the results of the operationalization of the OS impact indicators, with discussion of the causal effects of the measurements and any potential biases as well as enabling factors and barriers.

Key insights from each focus group are centralized and disseminated across the project to inform modelling, indicator development, and cost benefit analysis.

1.4. Purpose and overview of this deliverable

The purpose of this deliverable is to describe each case study in detail. Therefore, each case is detailed in the subsequent chapters. Each of these chapters introduces the case and the questions it pursues, details the current state of the impact pathway logic for the case and the causality narrative that supports it, specifies impact targets for the study, provides an overview of the research methods for the case study, and explains the next steps in the development of it.

A synthesis chapter follows these, which focuses on overlapping issues, concerns, and big picture insights offered by the work of each case study so far. This deliverable concludes with a chapter that focuses on the next steps for the case studies.

2. Case 1: Accelerating collaborations within academia and industry

2.1. Introduction

The Accelerating Collaborations within Academia and Industry case study explores the usage of the Portuguese publication repository infrastructure for Open Access (RCAAP)[1], to understand whether the availability of open access publications increases both the visibility of higher education institutions and publications and their usage by knowledge-intensive industries and Small and Medium-sized Enterprises (SMEs), fostering collaborations.

Collaborations between academia and industry and the visibility of the research publications deposited in repositories from the RCAAP infrastructure will be analysed by 1) studying the usage of Open Access publications, and 2) doing a citation and network analysis to study collaborations in specific domains, regions and sectors.

The uptake of Open Access (OA) publications available through RCAAP portal by Portuguese SMEs and industry will be studied, thus contributing to a better understanding of the impact of OA in non-academic contexts. Additionally, we aim to study the level of compliance with FCT's OA policy and test the generalised assumption about the potentially positive impact of OA mandates on innovation activities.

The geographical coverage of this case study is Portugal, and the domain coverage includes all disciplines, although some disciplines may be more likely to show collaborations with industry and SMEs. The defined time range is from 2015 to 2020.

2.1.1. Background/state of the art

RCAAP has been developed by FCCN, the Portuguese Foundation for National Scientific Computing, with the technical and scientific collaboration from Minho University, to collect, aggregate and index OA scientific publications from the institutional repositories of national higher education institutions and national OA journals, containing thousands of scientific and academic documents. RCAAP benefits from both the Portuguese legislation on academic degrees — as the legislation mandates the publication of PhD and master thesis in an institutional repository belonging to the RCAAP network[2] — and the Foundation for Science and Technology's (FCT) national policy for Open Access.

[1] https://www.rcaap.pt

[2] Decreet DL115/2013, August 7th mandates the legal deposit of master dissertations and PhD thesis and Portaria 285/2015, September 15th, regulates the terms under which this mandatory deposit occurs.

Besides the availability of the infrastructure, RCAAP also provides continuous support to repositories belonging to its network, by updating the underlying software (Dspace), providing a helpdesk, giving training and fostering a community of repository managers over the years.

RCAAP also provides a hosting service that many repositories benefit from. The Institutional Repository Hosting Service (SARI) is intended to be used by any institution in the scientific and higher education system to store its repository with its own individualised corporate identity. In addition to customising the image of the repository, each institution can also define and implement the configurations and parameterisations it considers appropriate to its organisational structure and its policies for self-archiving publications and managing the repository. The Institutional Repository Hosting Service is provided on a Software as a Service basis, i.e., it is based on RCAAP infrastructures (hardware, hosting, connectivity, base systems, applications, perimeter security, backup service, monitoring and alarming) which are managed and operated by the project team.

Additional reinforcement to the usage of the repositories is given by the Portuguese main funder, as compliance with the national mandate[3] makes it compulsory for the publications resulting from FCT funding to be deposited in a repository belonging to the RCAAP infrastructure.

Presently, RCAAP shows a Directory hosting different types of providers, from institutional repositories (academic) to repositories from Hospitals, Laboratories and Institutes, the Common Repository, a repository for "orphan" researchers (i.e., researchers without institutional affiliation); a repository for long tail data; institutional repositories from Brazil and a portal for Brazilian journals (OASISbr)[4]. As this case study focuses on collaborations between the academic sector and industry, only institutional repositories will be considered in the analysis.

2.1.2. Research question(s)

The most important research questions for this case study are the following:

1) Does the availability of RCAAP infrastructure have an impact on the submission of publications, and consequently on open access uptake?

We are interested in finding evidence whether the existence of the repository infrastructure RCAAP influences the submission of full text in Portugal, by making research available and ensuring that research outputs are preserved and accessible for all relevant stakeholders.

2) Is there an effect of the OA policies of Portugal's main funder FCT on the submission of metadata and full text by researchers?

[3] https://www.fct.pt/documentos/PoliticaAcessoAberto_Publicacoes.pdf

[4] https://oasisbr.ibict.br/

We aim to understand if FCT's OA policy influenced the submission of metadata and full text by researchers, by studying the level of compliance and its evolution over time, to test the generalized assumption about the potentially positive impact of OA mandates on innovation activities.

3) Does the national repository infrastructure and OA repositories foster the reuse of publications by companies?

We will try to identify to which degree the availability of OA publications led to collaboration, which are the companies using publications from the RCAAP network, and if there has been an increase in usage of publications from RCAAP repositories by companies.

4) Is there evidence of the usage of publications from repositories by industry, e.g., citations in patents?

We aim to find out if OA publications from repositories belonging to RCAAP network are cited in patent literature.

2.2. Causality / Impact pathway logic

The context of this case study consists of Portuguese legislation, mandating the deposit of master dissertations and PhD theses in institutional repositories, and the global rise on Open Access policies and mandates, more specifically FCT's OA mandate, dating May 14th, 2014, making the self-archive of publications financed by this funder in a RCAAP-indexed repository compulsory.

RCAAP operation benefits from financing from FCT and the operation is run both by FCT and UMinho, providing skilled staff and developing several activities: the maintenance of the RCAAP platform; regular upgrades in the software (Dspace); metadata curation and the insertion of several APIs (e.g. a projects database); a hosting service; a helpdesk team; a regularly updated training platform (eLearning); and support to a community of users, mainly repository managers.

The expected short-term outcomes are an increased discoverability of publications, leading to more views and downloads of publications, increased citations, and higher probability of collaborations between authors from different institutions but also from companies. This collaboration could assume different shapes, from the creation of start-ups and spin-offs to citations of OA publications in patents and industrial innovations. The latter would be medium-term outcomes.

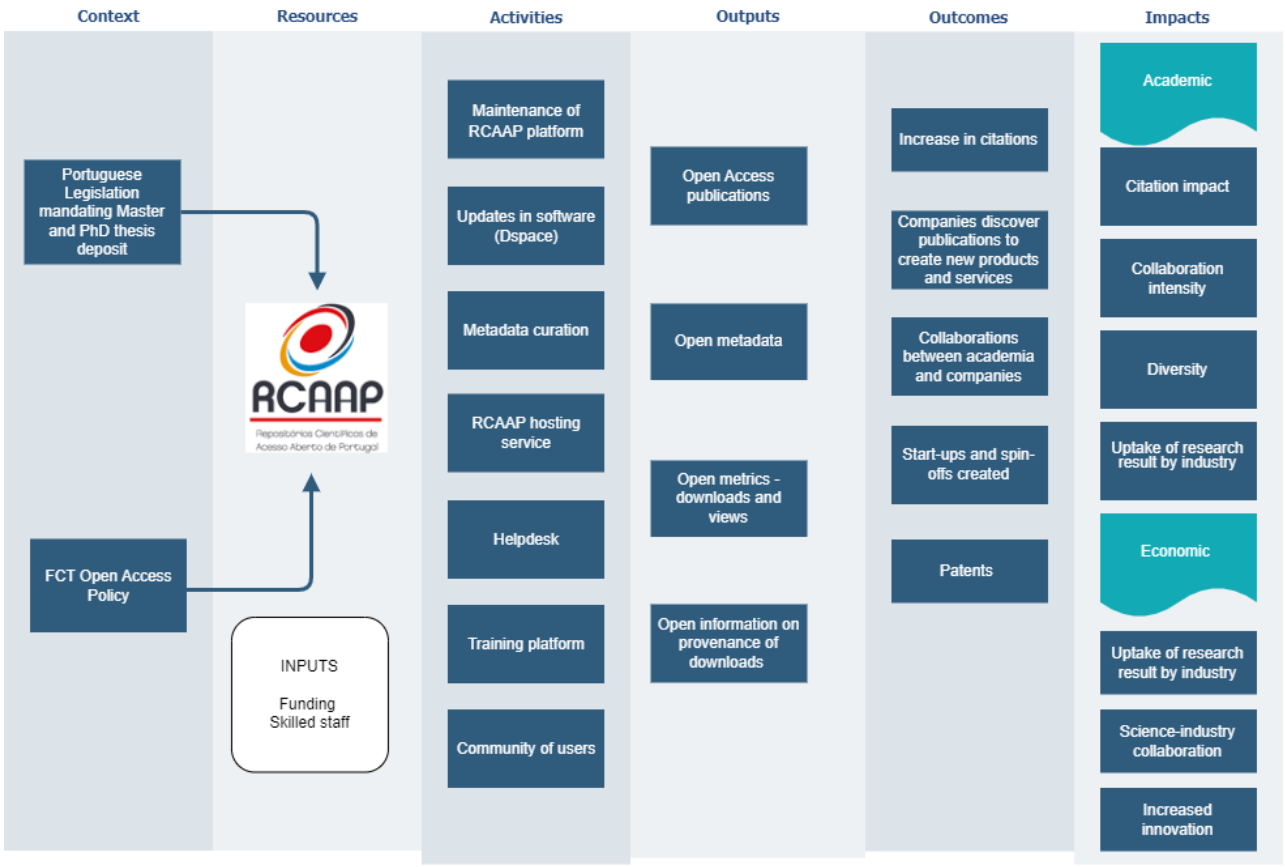

Figure 3: RCAAP impact pathway logic

The main research question is if the availability of OA repositories has an impact on the submission of metadata and full-text, and subsequently in OA uptake, leading to an impact in the visibility, collaboration diversity and intensity, increase of added value, and uptake of research results by industry and patents.

In trying to establish causality, the following relevant factors were identified:

- Availability of the full-text of publications — repositories many times have only the metadata available, and if not monitored, embargo periods on full-text may be longer than expected.
- Metadata quality — the implementation of personal identifiers and in-kind APIs in repositories is very recent, so institutional affiliations and collaborations may be harder to discover.
- Implementation of FCT OA policy — OA policy appeared in 2014, but still in 2017 the monitoring of FCT OA policy was not yet fully implemented and FCT allowed a transitory phase. Only in more recent years more effective monitoring was put in place and this probably influences the level of compliance.
- Fields of science — although institutional repositories from RCAAP network are multidisciplinary, publication patterns are different in different fields of science, and this will probably influence collaborations.

- Time — OA has been gaining relevance over the time, leading to more publications being deposited in repositories. This will probably lead to more collaboration.
- Different drivers — industry's main concern is problem-solving, so citations will need to be complemented with other types of documents and practices analysis.

2.3. Impact targets

This case study targets academic short-term impact, namely the citations reuse of RCAAP OA publications by companies and possible resulting collaborations between industry and academia.

At a longer term there are other potential impacts that OA publications may have, such as faster advancements in research, an increased speed in knowledge sharing, an increased efficiency and robustness of research results at an academic level, and at an economic level, increased innovation and economic growth. In this case study, we limit ourselves to the relatively short-term academic and economic impact of publications use by other researchers.

2.4. Methods

The main methods consist of focus groups, text and network analysis and citation analysis.

A set of focus groups with a panel of experts in the fields of OS, innovation and policymaking is planned. The first focus group on this case study took place March 28th, 2023, with the aim to introduce the project, the case study itself, and collect feedback on the envisioned methodology and pathway for the study. Eight experts from academia and industry debated if the national repository infrastructure and its repositories foster the reuse of publications and collaborations between academia and industry, what would be the research areas or disciplines more likely to show collaborations, the best way to measure them and the impact of the main funders' OA policies. Their feedback was most useful to realign the methodology and sources of information.

Publications will be retrieved from the OpenAIRE Research Graph for the years 2015-2020, relating Portuguese repositories. The OpenAIRE Research Graph[5] is an open resource that aggregates a collection of research data properties (metadata, links) available within the OpenAIRE Open Science infrastructure for funders, organizations, researchers, research communities and publishers to interlink information by using a semantic graph database approach.

[5] https://graph.openaire.eu/

Text mining will be performed on the publications for references to SMEs/ companies, using a list of companies from Orbis database. Publications will also be mined for references to FCT funding, using an FCT database of funded projects and grouped in the different fields of science.

On the other hand, Portuguese patents will be mined for citations of publications from RCAAP repositories, by using information from a patents database.

The metrics — citations, views and downloads — will be analysed using OpenAIRE usage statistics, and citations via OpenCitations.

Participants from the first focus group suggested studying social networks could be beneficial to this case study, as industry's main driver is not publishing articles, so this approach will also be integrated in the methods.

2.5. Next steps

The next steps will be the retrieval and analysis of a sample of publications from OpenAIRE. This first analysis will allow to refine the methodology and give a more faithful projection of the activities to be performed and timings. After this, a full analysis of the publications will be done.

The impact of the OS mandate of FCT is being considered for the cost benefit analysis to be performed within the PathOS project, and the analysis of the dataset of publications, as well as an inquiry on running costs with the RCAAP operation.

3. Case 2: Research data and knowledge use outside of academia

3.1. Introduction

The goal of this case study is to investigate the use of the three main French Open Science platforms — Open Edition[6], HAL (Hyper Article en Ligne)[7] and Recherche Data Gouv[8] — focusing, in particular, on the way in which their websites are visited by users outside the academia. We focused on these three platforms because they constitute the backbone of OS in France, and their study will offer us a window into the way in which Open Science trickles through society.

OpenEdition is a publication portal in the humanities and social sciences. It was founded in 1999 by the Centre National de la Recherche Scientifique (CNRS) and other research institutions (l'École des hautes études en sciences sociales, l'Université d'Aix-Marseille et l'Université d'Avignon). It mostly serves the humanities and social sciences and publishes books, articles, preprints and working papers. The main publishing languages are French (70%), English (11%) and German (11%). Most of its visitors come from France (35M), the Maghreb (Northwest Africa) (8M), the United States (5M), Italy (4M) and Mexico (4M). In 2021, it recorded a total of 110M visits and had a total of 1M uploaded documents since its creation in 1999.

HAL is an online platform developed in 2001 by the Center for Direct Scientific Communication (CCSD) of the CNRS, intended for the submission and dissemination of articles by researchers, published or not, and theses, emanating from educational establishments and French or foreign research institutes, public or private laboratories. It has now been functionally attached to two successive versions of the French national plan for Open Science (the 2018-2021[9] plan and 2021-2024[10] one) and covers all disciplines as well as other types of open resources: articles (60%), communications (15%), theses (7%), preprints (5%), books (4%), pictures (4%), source codes, etc. The main publishing languages are English (60%) and French (29%). In 2021, a total of 1M documents were published in it since its creation in 2001 and they were downloaded more 80M times.

Recherche Data Gouv is a more recent platform launched by French ministry of higher education in 2022 as a part of the national plan for OS and of the French data, algorithm and source code policy Roadmap[11]. It aims at joining the network of the European Open Science

[6] https://www.openedition.org/

[7] https://www.openedition.org/

[8] https://recherche.data.gouv.fr/fr

[9] https://www.ouvrirlascience.fr/plan-national-pour-la-science-ouverte/

[10] https://www.ouvrirlascience.fr/deuxieme-plan-national-pour-la-science-ouverte-2021-2024/

[11] https://www.ecologie.gouv.fr/feuille-route-donnee-des-algorithmes-et-des-codes-sources

Clound (EOSC)[12]. The platform focuses on data (observational, experimental, raw and processed, surveys, text corpora) and source code. In 2023, it covered mostly agricultural studies (35%), earth and environmental sciences (24%), medicine, health and life sciences (18%) and computer science (6%).

3.2. Background/state of the art

Research on the use of these platforms is scarce, especially concerning their potential impact on society, academia and industry. There are a few unpublished works that we use as a starting point for our case study:

1. Joël Gombin, Pierre-Carl Langlais. *Usages alpha. Étude préliminaire à orientation méthodologique*. Final report to ISTEX/ANR funded project Usages Alpha (ANR-10-IDEX-0004-02)[13].
2. Romain Deveaud, *Modalités d'accès au savoir ouvert sur les plateformes d'OpenEdition*. Final report to ISTEX/ANR funded project Usages Alpha (ANR-10-IDEX-0004-02)[14].

These two reports document a web analytics application developed during the project "Usages alpha" which is called Umberto[15]. Drawing on a dataset of Open Edition's connection logs from 2018-2019, Umberto helps identify and investigate readers of Open Edition's documents coming from outside academia. The two reports highlight that many OS resources hosted by Open Edition are indeed accessed by this type of reader, who is not the expected audience of the platform. It more generally helped to identify societal and economical actors who are using open science platforms.

3.3. Research question(s)

In this case study, we will examine the use of OS platforms by non-academic users and organisation focussing on three major questions related to this use. Together the answers to these questions should give us a comprehensive picture of OS dissemination in France as well as allow us to collaborate with the most important actors in this context.

1. *Who are the private and public organizations most frequently accessing French OS platforms?*

Extending the methodology of the earlier prototypes described in 3.2 and matching the information contained in the visit logs of the three platforms with public information about the

[12] https://eosc-portal.eu/

[13] https://nuage.cis.cnrs.fr/s/cM5gSfiYRadsimw

[14] https://nuage.cis.cnrs.fr/s/RJ2g3ygye72rfEE

[15] https://analytics.huma-num.fr/OpenEditionLab/umberto_oe/

IP addresses of institutions and companies, we will be able to analyse which organizations use these platforms the most and which content is of more interest to them.

2. *When are these platforms used the most and why?*

Through time-series analyses of different types of actions on the platform (publication of new items, page views, downloads, etc.), we will be able to detect peaks, regular oscillations, and long-term trends in the production and use of OS in France. Combined with an in-depth investigation of the most relevant patterns, this will help us gain insights on the impact of the Open Science produced in France and the motivation for its use, because it will allow us to understand which are the most important situation and context in which OS becomes visible outside the academic world.

3. *Which websites refer to these platforms and how are they connected?*

Using information about the websites that have referred users to the OS platforms, we will investigate which online venues are the most important in allowing the dissemination of OS, serving as channels to distribute OS beyond the three platforms were OS is published. We will then proceed to map the connections among them, by means of both semi-automatic crawling (tracing the networks of hyperlinks that connects them) and ethnographic investigation (by a close reading of the most important or interesting dissemination events in the period of the investigation). This will help us draw the landscape of French OS. We also expect to find, among the referrers of the platforms we investigate, websites that are not based in France. Their analysis will help highlighting the connections of French OS with the rest of Europe and the world.

3.4. Impact targets

Drawing on the work carried out in the second work package of this project, we can imagine using the visit logs of the three platforms described above to assess the uptake and impact of the OS resources they offer on different societal issues. As described in the PathOS indicator handbook, this requires going through five different steps, which are described in more detail in PathOS indicator manual of Open Science that we retrace here and apply to this case study.

1. *Defining the relevant societal issues and their desired solution.*

While there is no obvious agreement about which societal issues OS should contribute and how, a variety of initiatives exists in France to measure a variety of indicators of societal development. Following, once more, the PathOS indicator handbook, we suggest relying on the UN Sustainable Development Goals. While these goals and their measures can and have been criticized (Briant Carant, 2017; Fehling et al., 2013; Kopnina, 2016), they do offer a standardized set of societal ambitions that are monitored by a plurality of actors and can be compared over long periods of time and across different countries and regions.

2. *Assessing their improvement or deterioration of the chosen societal issues.*

For this step, we will use the statistics curated by the INSEE[16]. Even though there is always a risk in relying on official figures, because of course they present the official viewpoint of the State or organisation issuing them, the INSEE is generally considered an independent and trustworthy organisation. Its statistics offer a standardized quantification of all seventeen SDGs starting from the early 2010s (the exact year varies according to the indicator). For each SDG, different indicators are offered, and sub-national data can be found for some of them.

3. *Assessing the existence of Open Science resources available on different societal issues.*

Here our task is to evaluate the quantity and, if possible, the quality of the different types of *OS* resources that address the issues under investigation — in this case study scholarly publications and scientific datasets. It is also important to be able to assess the ratio of Open versus non-open resources, so that the availability of OS resources can be separated from the availability of scientific resources in general.

For preparing this indicator we inspect the three platforms discussed above and quantify the number of resources that they contain relative to the number of resources (publications or datasets) that are specifically dedicated to each of the 17 SDGs — we also observe the variation of this number year after year. To decide if an OS resource is relevant for a goal, we can consider whether its title or abstract contains keywords related to that goal after having defined a dictionary of SDG keywords[17]. Particular attention should be dedicated here to the HAL portal as it contains both open and non-open resources, offering a baseline to calculate the ratio of OS over non-OS resources addressing each societal issue.

4. *Assessing the uptake of Open Science by the social actors active on different issues.*

Since OS can influence a given issue only if taken up by the social actors engaged on that issue, it is necessary to assess to what extent different OS resources are mobilised in the actions and discourses concerning the issues in question. This starts by identifying the most important collective actors (public institutions, CSOs and private companies) active on each SDG and locating the website (or social media account) that each of them employs to put forward its position statements and action reports. This list of digital venues then needs to be searched, crawled and scraped to identify references to scientific resources of different types. This should allow to calculate the ratio of references to OS over non-OS scientific resources in the statements and documents of each actor active on each societal issue. The ratio for different issues can then be compared among them as well as to the general ratio for each website.

[16] the French national statistics institute https://www.insee.fr/fr/statistiques/2654964

[17]https://www.leidenmadtrics.nl/articles/consensus-and-dissensus-in-mappings-of-science-for-sustainable-development-goals-sdgs)

5. *Disentangling the effect of Open Science from the many other dynamics that may influence the evolution of social issues.*

The idea here is to associate the evolution of each SDG through time and the mobilization (or lack of mobilization) of OS resources in the discourses around it. Proving the existence of a real association between the two remains challenging. To do so we propose two research directions:

a. A comparative approach, across different issues and across different periods, aiming to establish that a higher level of Open Science mobilization is regularly associated with positive evolution of a given societal issue.
b. A qualitative exploration of the nature of the Open Science mobilized, the identity of the actors mobilizing them, the role played by OS in their strategies, the reception/reaction of the other actors and, finally, the precise dynamic of evolution of the issue at stake. This qualitative exploration is meant to reveal the causality paths that lead to Open Science impact.

3.5. Causality/Impact pathway logic

As said by Case Study 3 (see 4.2 below), "in any observational study, causality is typically difficult to establish".

In this case, we will try to study the uptake of OS resources available in OA on OS platforms, that is to say the short-term uptake of those resources in societal areas such as citizen science, education, policy making, legal sector, press and public debate, medical practice, civil society, NGO's and artistic sector, as well as in several economic sectors such as banking, aviation, energy, cars, manufacturing, public firms, agribusiness, insurance, health, real estate and transport.

We posit that OS platforms providing OA to science could have an effect on the diversification of the readership of science with more societal and economical actors reading science than in fee-based platforms, the diversification of co-producers of science and the emergence of specific topics, especially those related to the SDGs. We would also like to see how these effects varies depending on the languages of the resources (English prevailing on HAL and French on OpenEdition), their type (HAL & OpenEdition being mostly articles, RechercheDataGouv being mostly data) as well as according to the disciplines and concepts that the resources are about.

At strict platform level, it is nearly impossible to draw conclusions on long-term effects of OS platforms such as the effect on democracy, trust in science, science literacy, on gender or ethnic inequalities. The effects of scientific information on beliefs and representations, skills, the social and political organization of groups, and their productive activity is a classical question of audience reception sociology[18] that can only be addressed through in-depth qualitative

[18] https://en.wikipedia.org/wiki/Audience_reception

research (as mentioned in 3.4.5.b) and goes far beyond a quantitative log analysis of OS platforms.

However, as mentioned in section 3.4.5.a, we can identify the emergence of concepts related to a specific Sustainable Development Goal (SDG) in open science resources and compare it to the indicators which quantify that particular SDG, though without the ability to establish a correlation, let alone a causation. Online surveys and semi-structured interview might also give us a few complementary indications on how uptake translates into longer-term impact.

3.6. Methods

In order to classify and characterize users of the three Open Science Platforms of our study, we will first collect and analyze the connection logs with *logstash*[19] and another tool developed by CNRS team INIST (Institut de l'Information Scientifique et Technique) called Ezpaarse[20], which cleans the raw log files and enriches them with various information, including the metadata of consulted resources.

The logs contain the IP addresses of each connection to a resource. Using semi-automated scrapping of WHOIS[21], we will constitute a repository of ranges of IP addresses and which organization they have been attributed to, helping us better characterize the users of the resources. Another piece of information contained in the connection log is the referrer, i.e., the webpage that points the user to the open science resource. If this resource is pointed at by a webpage of a specific societal group or economic actor (i.e., the webpage of a bank or of a ministry), it will help us identify the user more precisely (I.e., at least as "someone being interested in the webpages of a certain societal / economical group").

The logs also contain metadata of the consulted resources and a unique resource identifier thanks to whom we can link the resource to a SolR[22] database containing extensive information about all open science resources available on the platforms. Then, thanks to scientometrics, we can map those resources, that is to say cluster them using several variables: language, concepts, resource type (code, article, data…), discipline, as recently done by T. Venturini on AI[23] in his scientific landscape analysis.

One of our fundamental working tools are *datasprints*. They aim at discussing and solving technical and methodological issues and review intermediary results by gathering the various

[19] https://www.elastic.co/fr/logstash/

[20] https://www.ezpaarse.org/

[21] https://fr.wikipedia.org/wiki/Whois

[22] https://solr.apache.org/

[23] http://www.tommasoventurini.it/ai2s2/

stakeholders (head of platforms, data analysts, sociologists, system administrators) of our research around a table.

During those datasprints we make intensive use of data visualization and exploration tool Kibana[24] that helps us visualize significant relationships between referrers, IP addresses and resource metadata. First datasprint took place on June 7th, 2023, and was focused on Open Edition.

Online questionnaires and semi-structured interviews will be used to gather information to further qualify users when IP address and referrer do not give relevant information (i.e., Referrer is Google, or IP address is the one of a commercial ISP (Internet Service Provider).

3.7. Next steps

The first focus group was held in October 2023 and participants stressed the necessity to do qualitative analysis to complete our quantitative log analysis, in order to be able to move from short-term uptake to longer-term impact assessment. This will be integrated into our upcoming milestones to the extent possible.

- December 2023: Our second *datasprint* will have a primary focus on HAL.

We'll transition from working with one month of logs for prototype analyses to a full year for both platforms, and will complete the IP range repository.

- March 2024: We'll host our second focus group.

We will design an online survey to be administered on the first two platforms and ensure its feasibility with platforms directors and system administrators. This survey should help fill in information gaps that couldn't be obtained from log analysis.

- May 2024: Our third datasprint will be dedicated to RechercheDataGouv.

We'll study the feasibility of counterfactual analyses run on both pirate platforms and for-pay platforms to try and isolate specific causal effects of open science platforms.

We will run comparative analyses based on differences between our three platforms (data vs articles, English vs French).

- September 2024: We'll hold our third focus group.

Following feedback from the focus group, we will design semi-structured interviews to further investigate points that couldn't be clarified by the online survey.

- November 2024: Our fourth datasprint will focus on all three platforms, and we'll deliver a final and comparative analysis.

[24] https://www.elastic.co/fr/kibana/

4. Case 3: Cross cutting effects due to Open Research data from a National Repository

4.1. Introduction

This case study investigates the effect of data availability in a particular repository on data use. Do scientists make more use of publications for which data has been made available in a specific repository rather than another repository? More specifically, the case study will review if there is a difference whether the data is being made available through a national data repository or through other international (e.g., Zenodo) or disciplinary repositories (e.g., Inter-university Consortium for Political and Social Research), with a particular interest on the EASY repository.

These questions will be investigated by studying the EASY database, which is maintained by DANS in the Netherlands. This is a data repository that contains datasets from the social sciences and humanities (SSH) with a focus on the Netherlands. The case study will specifically investigate data sharing practices in SSH and how the use of an Open database like EASY might affect data uptake in general and compared with other data sharing practices.

DANS is the Netherlands' institute for permanent access to digital research resources (Doorn, 2020). It encourages researchers to make their research output adhere to the Findable, Accessible, Interoperable and Reusable (FAIR) principles. One of the services DANS provide is the EASY digital archive which has been steadily growing over the years reaching about 120,000 datasets in 2019. These datasets relate to all kind of fields within SSH, but archaeology represents a large majority with 80% of the datasets being attributed to this field. Data uptake from EASY has been on the rise in proportion to the increase of datasets added to the repository. This uptake is quite skewed, where some datasets in the database reach up to 7,000 downloads and the lower half percentile of datasets barely gets any downloads. We will be studying the uptake from the perspective of research publications that reused these datasets.

The main audience for this case study consist of researchers, and open data repositories. Researchers could get insights out of this case study on how open data repositories currently function and are being used by other researchers. A common understanding on this topic could improve scientific collaboration, data sharing and research output in general. The case study could also provide insight into the data reference practices that are currently in use in SSH fields. Insight in these dynamics might provide opportunity to make data sharing more findable and effective. Lastly, insights in the use of data repositories by researchers could be a source of recommendations for improving data structuring and open data repository use.

4.2. Causality narrative/impact pathway logic

In any observational study, causality is typically difficult to establish. Here, we try to identify what we believe to be the main causal factors, so that we can develop a strategy that would allow us to identify the causal effect. In particular, we are interested in the causal effect of where a dataset is shared (i.e., in which repository) on the usage of the dataset by others. In a sense, this is similar to an earlier study that tried to identify the causal effect of where a study was published on the later citations of that study (Traag, 2021). This is the model of the factors that we identified to be relevant in this context:

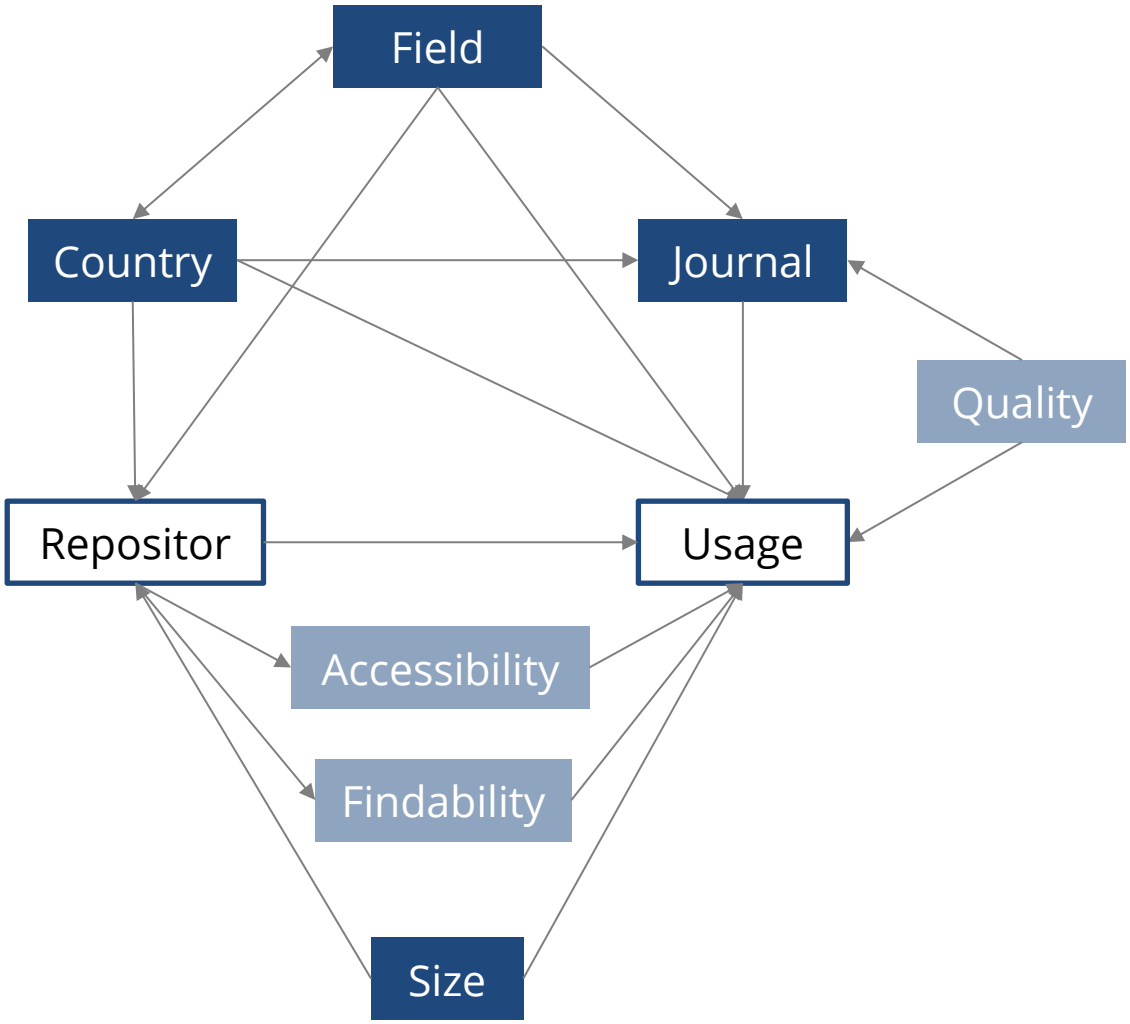

Figure 4: EASY impact pathway logic

We identified the following factors to be relevant in this context.

- **Accessibility.** Datasets that are not made accessible 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.
- **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** (of accompanying publication). Datasets that are part of a publication in a high-impact journal may be more frequently used than datasets that are part of a publication in a lower-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.

We need to be careful about controlling for all these factors directly, as some factors should *not* be controlled for. For instance, we may expect accessibility and findability to be affected by the repository where they are shared, so that sharing on one repository may make the dataset much more likely be found by prospective users, for instance, because the repository is indexed by some search engines or bibliometric data sources. In that case, we should *not* control for these two factors, otherwise we would be missing some causal effects, because the effect of sharing data on a particular repository might be in part mediated by the accessibility and findability. These factors are not observed by us in this study and might even be unobservable and are greyed out Figure 8. Since we do *not* control for these factors, this is fine.

An important unobservable factor is that of "quality", understood as a broad multidimensional concept that covers both the research behind the construction of the particular datasets, and the quality of the dataset itself. We might expect quality to affect where a dataset, or the publication that introduces the dataset, is published; that is, in which journal. At the same time, quality might affect the usage: if a data set is of higher quality, it is more likely to be used later. The journal itself might also have an effect on the later usage of the data. Luckily, we can close that non-causal pathway by controlling for the journal (which we observe). However, note that journal is also acting as a collider (i.e., it is a common effect): the field affects in which journal research gets published, and conditioning on the journal would hence open a non-causal pathway. We therefore need to condition additionally also on the field. Similarly, we also need to condition on the country. These factors are coloured dark blue in the picture of the model.

Finally, there is one factor that, in our assumed model of causal effects, only acts as a confounder, and that is the size of the dataset. Some datasets might be larger and therefore more likely to be shared on a particular repository, and this might also affect its usage. We hence simply need to control for size.

4.3. Impact targets

This case study targets academic short-term impact, namely the reuse of data by other researchers. This is quite a direct impact that sharing data on a data repository might have. There are other, longer term, impacts that data sharing may have, such as making research more efficient (because of the possibility of re-using data) and increasing the robustness of research results (because open data allows direct replication). However, in this case study, we limit ourselves to the relatively short-term academic impact of data use by other researchers. Although this is a relatively narrow impact target focus of the case study, this impact is already quite challenging to define and measure, as we will see.

4.4. Methods

Interviews, surveys, focus group discussions:

A set of focus groups with a panel of experts in the field of open science, data reuse and open data repositories are planned. Besides providing valuable input for the case study this panel also advises what to focus on and what methodologies to consider in the investigative part of the case study.

Text data analysis

Unfortunately, the use of data by other researchers is often not clearly indicated by researchers. For instance, researchers are used to citing prior literature, but citing datasets is not yet as common. Especially in the social sciences and humanities, datasets are being referred to in quite diverse ways (Gregory et al., 2023). This means that data citations (Robinson-Garcia et al., 2017) are unlikely to provide a sufficiently clear picture. Most likely, a text-based approach where machine learning and natural language processing would be used to extract references to datasets would be most promising[25], although this is a relatively challenging task.

[25] For instance, https://github.com/kermitt2/datastet or https://github.com/DataSeer/dataseer-ml. Other alternatives might also be developed in the context of PathOS.

4.5. Next steps

As indicated, we want to use text analysis for extracting data mentions from SSH literature. This essentially requires us to go through a couple of steps. First of all, we need to identify full-text publications that allows us to extract data mentions. Although there is already a database of full-text publications available at CWTS, we might want to restrict our analysis to open access publications, so that the analysis can be easily reproduced. To develop the methodology that allows us to extract data mentions itself, we need to first test existing methodologies. There are some existing implementations, such as datastet or dataseer-ml. Other alternatives might also be developed in the context of PathOS. There is also an ongoing effort by DataCite and other collaborators to build an Open Global Data Citation Corpus. Even if that data source is not yet developed fully, some of the methodologies that are being developed might be used.

One critical aspect of doing this is to test and validate the extraction of data mentions. This means that we will have to manually identify data mentions for a test dataset that could later then be tested against an automated process. In addition, this will provide some information about the referencing practices in SSH and its heterogeneity.

5. Case 4: Open Science practices during the COVID-19 pandemic

5.1. Introduction

The COVID-19 pandemic posed an unparalleled challenge, necessitating swift scientific action and effective communication of research findings. Open Science (OS), with its principles of transparency and accessibility, emerged as a pivotal approach in facilitating this accelerated research landscape.

This case study examines **the use of Open Datasets**, a key instrument of OS, during the pandemic's research phase. These datasets, freely available to everyone, represent the principles of reuse, modification, and unrestricted sharing. Our focus narrows down to COVID-19-related publications and research projects that actively used these datasets. Our goal is to discern the specific impact of using open datasets on scientific advancements.

Preliminary assessments indicate that using these accessible datasets significantly propelled research. Being openly available, these datasets facilitated swift integration, application, and iterative improvement grounded in collective feedback. The open framework also furnished a golden opportunity for self-correction in research. Given that numerous COVID-19 publications underwent retractions or adjustments, the community's capacity to swiftly access, verify, and correct erroneous findings became critical. Hence, the recurrent cycle of **uptake, reuse, correction, and subsequent reuse became a reinforcing loop**, amplifying the strength and accuracy of scientific conclusions, even though the urgency of the situation occasionally led to outputs without rigorous vetting.

The objective of this case study is to gauge the influence of using open datasets on the pace and reliability of COVID-19 research findings, spotlighting the repetitive feedback loop of data uptake, reuse, correction, and reuse during this crisis.

The stakeholders influenced by our case study comprise:

Funders: A perspective into the tangible outcomes of their financial allocations, especially those channelled to projects that employed open datasets, can guide effective resource allocation and OA mandates.

Scientists: Our analysis offers insights into the merits and hurdles of OS practices and the utilization of open datasets, encouraging more calculated research strategies and partnerships.

Policy Makers: As policy choices hinge on punctual and precise research, grasping the significance of using open datasets can shape data-sharing protocols and guidelines, especially in urgent scenarios.

General Public: A realization of the concrete advantages of research clarity, particularly in endeavours funded by taxpayers, accentuates the value of OS in confronting worldwide challenges.

In sum, this case study strives to untangle the intricate dynamics between using open datasets and the rapid scientific advancements during the COVID-19 pandemic, spotlighting the promise of Open Science in fostering collaborative and agile global solutions.

5.2. Impact pathway logic/Causality Narrative

5.2.1. Causality Narrative: Open Data's Influence During the COVID-19 Pandemic

The overarching aim of our study is to discern the causal impact of using open datasets on the pace and quality of scientific research during the pandemic. As we delve into this investigation, it's imperative to acknowledge and control for several key factors:

- **Pandemic Urgency:** The pandemic instigated an unprecedented surge in research endeavours, with many researchers racing against time. This naturally occurring acceleration could cloud our observations, making it vital to adjust for this intensified urgency.
- **Dataset Quality:** The inherent quality of datasets, open or otherwise, can significantly shape research outcomes. As we focus on the utilization of open datasets, we must account for potential variations in their quality, ensuring we compare like-for-like datasets in terms of comprehensiveness and accuracy.

To establish causality, we will employ a combination of methodologies. A crucial component will involve benchmarking the rate and robustness of research outputs that utilized open datasets against comparable outputs from a time devoid of such intensive open dataset utilization. By doing so, we aim to isolate the influence of open datasets from other confounding variables, offering a clearer perspective on their genuine impact on scientific research during this critical period.

5.2.2. Impact Pathway Logic

As we delve into the role the use of open datasets played in scientific breakthroughs during the COVID-19 pandemic, it is important to also understand the mechanism behind our findings as well as the relationship between the expedited pace of research and its broader societal and economic implications.

Mechanisms at Play: Assuming our examination shows that open datasets indeed hastened research, the next logical step would be to discern the processes that facilitated this

acceleration. At the heart of this is the iterative nature of open science. Researchers globally had access to these datasets, which could have propelled a collaborative spirit and collective improvement. By examining metrics such as the number of citations, cross-references, and the breadth of interdisciplinary applications, we can gauge the breadth and depth of collaboration and knowledge dissemination. Furthermore, looking at the timeline of publications might give insights into the recursive feedback loop - the rate at which errors were corrected, innovations made, and knowledge built upon.

Impact Analysis: With an understanding of the mechanisms in place, we then seek to **correlate** the accelerated scientific progress with tangible societal and economic impacts during the pandemic. This involves tracking the ripple effect of swift research — from influencing public health guidelines to potentially impacting economic policies that are rooted in scientific understandings. By comparing the timelines of scientific findings with the implementation of related measures we aim to paint a clearer picture of the effects of rapid, open research during critical times.

Figure 9 below summarizes the impact pathway logic of this case study.

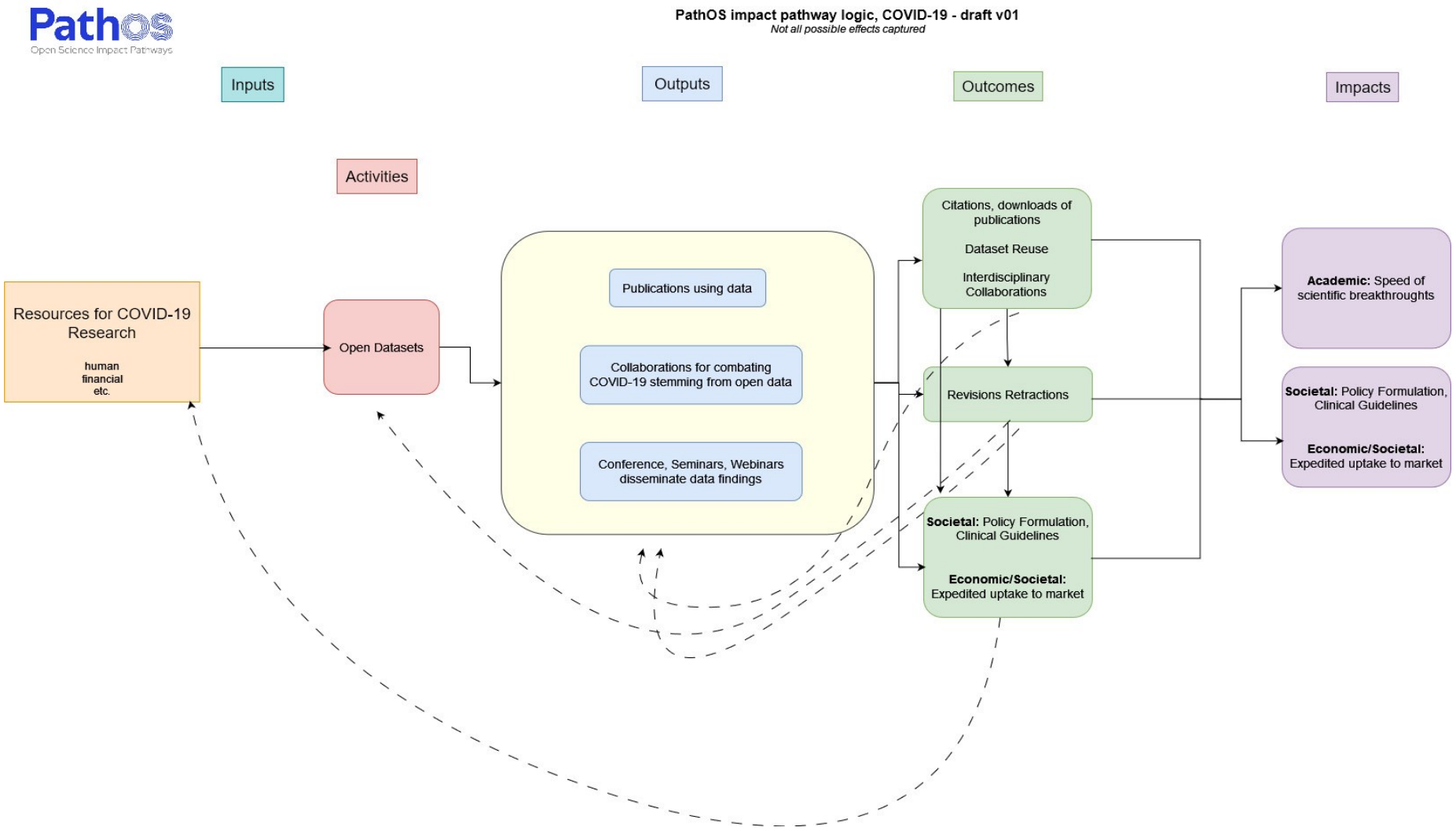

Figure 5: Open Data impact pathway logic

5.3. Impact targets

Within our exploration of the effects of open dataset use during the COVID-19 pandemic, specific **Impact Targets** serve as our guideposts to understand and measure the consequential impacts on scientific progress and broader societal implications.

Scientific Progress — Causality: A few key metrics help to delineate the direct influence of open datasets. The time between the release of an open dataset and subsequent related publications offers a tangible measure of the pace of scientific advancements. Concurrently, the number of retractions or corrections serves as a barometer for the quality and reliability of this accelerated research. Furthermore, the rise in cross-disciplinary collaborations points towards the versatility and encompassing nature of the data in use, bridging various scientific domains. An uptick in citations for these studies highlights their foundational role in the overarching research narrative of the pandemic.

Impact Analysis — Correlation: Beyond the immediate realm of research, the implications of these studies permeate into the economic and societal fabrics. On the economic front, the uptake of research-driven innovations signals the transition of academic knowledge into actionable solutions. Societally, the incorporation of these research findings and innovations into policy documents and clinical guidelines underscores their pivotal role in shaping public health responses and strategies.

Category	Description
Scientific Progress — Causality	Speed of scientific breakthroughs (time between dataset release and related publications)
	Number of retractions/corrections
	Cross-disciplinary collaborations
	Increased citations
	Characteristics/Mechanisms at Play:
	• Dataset reuse
Impact Analysis — Correlation	**Economic**: Uptake or research innovations
	Societal: Uptake of research/innovations in policy documents/clinical guidelines

Table 2: Causality and impact analysis of Open Data on COVID-19 research

5.4. Methods

This section outlines our methodological framework, presenting the specific categories of analysis, our objectives within each, and the tools and data sources we will be leveraging, all of which will be described as part of the PathOS Handbook of Indicators[26]. Our methodological approach is designed to mainly trace the trajectory of scientific progress influenced by the use of open datasets but also to gauge broader economic and societal implications. This comprehensive analysis will provide a holistic perspective on the ways open datasets shaped the pandemic's research landscape.

Analysis Category	Objective/Description	Tools & Data
Group Selection & Control Testing	Identify publications in the domain of interest and partition them into designated groups (see causality narrative) and run tests verifying controls.	Comparative statistics & SciNoBo[27] Toolkit *OpenAIRE Graph[28] & COVID-19 Databases*
Dataset Use Recognition	Classify research based on the type and source of datasets used, and whether they are created or reused in the scientific work.	Dataset & Software NLP extraction toolkit *OpenAIRE Graph, COVID-19 Databases*
Scientific Progress Indicators	Measure key indicators such as research speed, retractions, cross-disciplinary collaborations and so on, as presented in the previous table.	SciNoBo Toolkit, Citation tracking and analysis. *OpenAIRE Graph, COVID-19 Databases*
Impact Analysis	**Economic:** Track the integration of research innovations into the economy and their broader economic implications.	Innovation extraction system, Citation analysis *OpenAIRE Graph, Company Websites*
	Societal: Track the influence of research findings in shaping clinical guidelines, and public health strategies, linking them to specific SDG targets.	Citation analysis, SDG Classifier *Pubmed*

Table 3: Open Data methodological approach

[26] https://handbook.pathos-project.eu/

[27] https://doi.org/10.3389/frma.2023.1149834

[28] https://graph.openaire.eu/

5.5. Next steps

Before starting the data work for this case study, we will convene its first focus group where we will present the study's objectives and the rationale underlying our chosen methodology. Our aim is to collect feedback on any overlooked dynamics and to gauge the alignment of our study with their domain-specific knowledge. Such views will be instrumental in refining our impact pathway logic and finalising the causality narrative. We will delve into the impact targets, determining their priority and homing in on specific datasets and correlations within various elements of the pathway model.

6. Case 5: Emerging Topics Fostered by Open Science: Gender in AI and Climate Innovations

6.1. Introduction

This case study sets out to understand the role of Open Science in the emergence of new research topics, especially within the domains of AI and climate change. By championing transparency and inclusivity, Open Science potentially paves the way for a diversified research landscape, inviting perspectives and research interests that traditionally might have been side-lined. In AI, for instance, there is an escalating conversation about whether AI tools themselves are biased, specifically regarding gender. Similarly, in climate change discourse, new technologies are making their presence felt.

As societal challenges evolve and diversify, there is a pressing need for scientific research to be both *adaptable* and *directly responsive*. Open Science offers more than just wider access to research; it suggests a shift in how research topics are identified and pursued. By promoting *inclusivity* and *transparency*, Open Science could facilitate the **exploration of a broader range of research questions**, many of which align more closely with contemporary societal needs. By examining how new research areas have surfaced, we aim to understand the potential influence of Open Science on the landscape of scientific inquiry.

Within the framework of Open Science, various instruments and methodologies can act as potential catalysts for the introduction of new research topics. This includes transparent protocols, collaborative research platforms, open peer review, and more. However, for the scope of this case study, our analysis will concentrate on **Open Access publications and the different routes to Open Access available**, such as Green OA (OA in a repository) vs. Gold, Hybrid or Bronze OA (published OA)[29]. These elements have been selected due to their foundational role in disseminating knowledge, facilitating research, and enabling collaboration, thereby potentially influencing the direction of scientific investigations.

Our focus on **gender bias in AI** and **climate change** directly ties to key *United Nations Sustainable Development Goals (SDGs)*. Investigating gender biases in AI aligns with SDG 5, emphasizing Gender Equality, ensuring technological advancements do not perpetuate disparities. Simultaneously, delving into emerging topics in climate change reinforces the commitments of SDG 13, championing proactive Climate Action, and related SDGs. Through

[29] Gold OA: OA in a fully OA journal, Hybrid OA: OA in a hybrid journal with a license, Bronze OA: OA in a hybrid journal without a specific license. A hybrid journal is one that includes both OA and non-OA articles.

these subjects, we aim to show how Open Science can meaningfully connect research to pressing global challenges.

Objective: This case study seeks to determine the extent to which Open Access to publications and the different routes to OA have influenced the emergence of specific research areas. We will be focusing on:

1. the emerging topic of Gender and AI, within the AI research domain, which revolves around the potential gender biases in AI tools, and
2. emerging topics and technologies in climate change research.

The following **stakeholder groups** can be influenced by the findings of this case study.

Policymakers & Research Funders: Decisions regarding research funding, intellectual property, and educational mandates can be informed by the connection (or lack thereof) between Open Science practices and the evolution of research topics, guiding future strategies and mandates in order to align scientific efforts with societal needs.

Scientific Publishers and Journals: Their publishing models might be influenced by the relationship between Open Access and research topic emergence, ensuring they remain central to scientific discourse.

While the immediate implications of this study's findings will be most pertinent to the identified stakeholder groups, it is worth noting that academic institutions, industries, and the general public also have a stake in the outcomes. These broader groups will be particularly interested in the actual topics that emerge, and their potential applications. However, their connection to the mechanisms of Open Science as a driving force might be more indirect and rooted in the broader landscape of research and its societal impacts.

6.2. Impact pathway logic/Causality narrative

6.2.1. Harnessing the Horizon 2020 (H2020) Mandate for Causality Analysis

To understand the causal effect of OA on the emergence of new research topics, we employ a strategy using various comparison groups and the Horizon 2020 OA mandate[30]. The mandate

1. requires deposit of articles in a repository (i.e., Green OA) and
2. encourages publishing in Open Access (Gold, Hybrid, or Bronze OA) via the remuneration of article processing charges (APC), given specific eligibility conditions.

[30] Article 29.2 of theH2020 Programme AGA – Annotated Model Grant Agreement, https://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/amga/h2020-amga_en.pdf

Defining the Comparison Groups: To dissect the effect of OA on topic emergence, we classify "*high quality" peer-reviewed* publications *from 2014 to 2020* (see Confounding Factors below) into three distinct groups.

"HIGH QUALITY" PEER-REVIEWED PUBS IN 2014 - 2020	NON-OA	GREEN OA ONLY	GREEN AND GOLD/HYBRID/BRONZE OA
Non-H2020 funded	Group A (control)		
H2020 funded	NA[31]	Group B	Group C

Table 4: Funder and OA comparison groups

Group A encompasses a mix of OA and non-OA publications, reflecting the broader publishing spectrum in the AI and Climate domains. Some publications in this group may naturally be OA, influenced by journal preferences or external mandates. This varied collection offers a baseline, capturing the usual evolution of emerging topics without any specific directive's influence, on average.

In contrast, when juxtaposing this baseline (Group A) with the distinct OA approaches of **Group B and C** — both influenced by the H2020 mandate — we aim to pinpoint any distinct OA effects on the emergence of topics. Specifically, a comparison between Group A and Group B will shed light on the role of repository-driven Green OA. A marked variation in topics suggests the importance of unrestricted repository access in driving research innovation. Meanwhile, the contrast between Group A and Group C highlights the combined impact of Green OA and direct OA routes, such as Gold, Hybrid, or Bronze. Noticeable differences between these groups underscore the effects of varied OA strategies in shaping research trajectories.

Attributing Topic Emergence to Specific Groups: Topics in research do not emerge from isolated publications but are rather a result of collective academic discourse. To attribute the emergence of a topic to a particular group, we will identify a pattern where a critical mass of publications from that group converges on a particular topic over time. If, for instance, a topic like "Gender and AI" becomes prominent mainly in Group C, it suggests that the immediacy of Gold/Hybrid/Bronze Open Access had a pivotal role in catalysing that topic's emergence within the AI research landscape.

Controlling for Confounding Factors: In our analysis, it's crucial to ensure that the observed effects genuinely arise from the influence of Open Access (OA) and not from other external factors or inherent differences in the groups. We account for **temporal effects** by selecting publications within the H2020 timeframe (2014-2020), ensuring external events or advancements influence all groups uniformly. We control for the **quality of publications** to ensure that any observed effect is not merely a byproduct of higher research quality. For non-

[31] H2020 publications may belong to this set due to IPR reasons, but we exclude them from the analysis.

H2020 publications (Group A), quality is controlled by selecting papers with comparable *citation and usage statistics* as those in H2020-funded groups. Additionally, a *peer-review* sampling ensures consistent quality perception across the groups. Lastly, by focusing on specific domains (AI or Climate), we account for **domain-specific** trends, publication habits, and inherent differences in the pace of research advancements.

Enabling Factors in Our Analysis: The research landscape is shaped by numerous enabling factors that could influence the emergence of new topics. These include technological advancements, trends in collaborative and interdisciplinary research, the global reach of OA, nuances in research funding, and heightened public engagement with OA materials. However, in constructing our comparison groups, we have selected high-quality, peer-reviewed publications. This approach ensures a certain level of uniformity in exposure to these enabling factors. Group A, even though broader in scope, consists of papers that meet these stringent criteria, minimizing the variability that could arise from factors like regional disparities in research resources or infrastructural differences.

6.2.2. Impact Pathway Logic

This study's primary objective centres on establishing a clear causal link between OA the emergence of new research topics, with a particular focus on Gender & AI and new climate change-related topics. To enrich our understanding of OA's influence, we explore two further dimensions:

Mechanisms at Play: We will analyse the **characteristics of the emerging topics that can be attributed to OA**. By examining features such as multi-disciplinarity, we will be able to better understand the underlying mechanisms that OA might be promoting. For example, a surge in multi-disciplinary topics in the realm of AI might suggest that OA is facilitating broader academic participation (e.g., from the Social Sciences), leading to areas like gender & AI.

Correlations with Impacts: Beyond identifying the emerging topics due to OA, we will also assess their potential broader societal impacts. For example, we can examine whether new climate change topics produce industrial innovation (industry uptake) or whether AI & gender is a topic that can also be traced all the way to policy documents, suggesting the beginnings of societal impact. While we will not be establishing direct causality in this section, the correlations can shed light on the societal relevance and footprint of the topics created in an OA context.

The Impact pathway logic is summarized in Figure 10 and tries to capture the above aspects, with the arrow in bold being the direct causal link that we aim to establish. The impact pathway starts with resources invested in Horizon 2020 projects and the infrastructure supporting them. The central intervention here is the OA mandate, which leads to an increase in OA publications. These open-access articles gain more visibility and attract multidisciplinary collaborations.

As more people access and collaborate on these publications, new research topics emerge. These new topics can have wide-ranging impacts. In academia, they influence ongoing research

directions. In the economy, they can guide the adoption of novel technologies in industries, and in the society, they can influence policies and public perspectives.

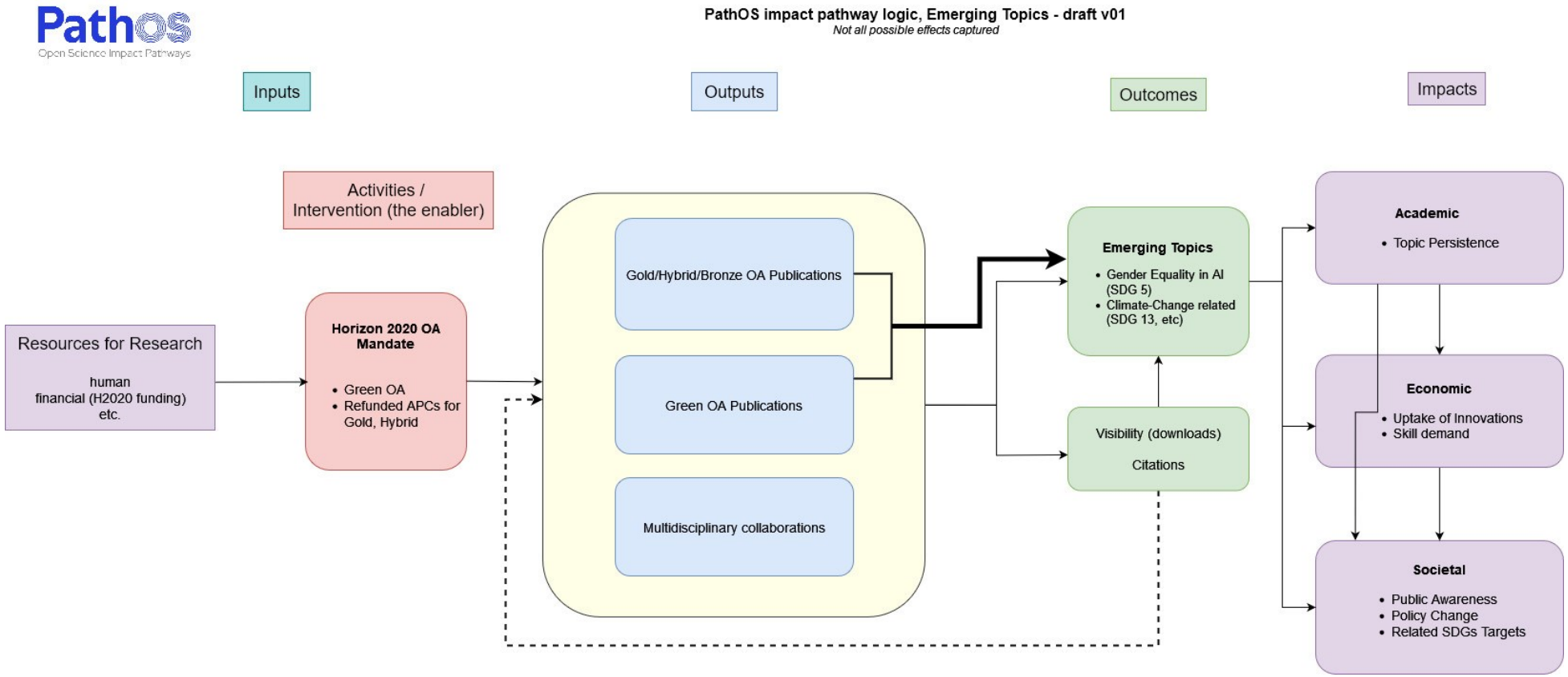

Figure 6: H2020 OA mandate impact pathway logic

6.3. Impact targets

In our analysis, we prioritize a rigorous and systematic approach by identifying proxy indicators for each category, ensuring that our conclusions are grounded in robust data-driven insights. At the forefront of our findings is the emergence of new research topics directly attributed to OA. Beyond this causal relationship, we delve into the characteristics and mechanisms at play that these emerging topics exhibit, such as their multidisciplinary nature, the amplified participation from non-academic entities, how the relate to specific SDGs and their enhanced visibility in the academic sphere.

We then extend our analysis to correlate these findings with broader impacts, segmented into academic, economic, and societal domains. This entails evaluating the longevity and prominence of these new topics within academia, the creation of new innovations and demand for specialized skills in the economic sector, and the resonance and adoption of these research topics in policymaking and SDG-related outcomes.

Category	Description
Emerging Topics — Causality	Emerging topics attributed to OA
	Characteristics/Mechanisms at Play:
	• Multi-disciplinarity
	• Relation to SDGs
	• Increased non-academic participation
	• Enhanced visibility and citations
	• Representation of women as authors
Impact Analysis — Correlation	**Academic**: Persistence and prevalence of new topics
	Economic: Emergence of new innovations and skill demand
	Societal: Uptake of research/innovations in policy documents

Table 5: Causality and impact analysis of OA

6.4. Methods

The case study aims to extract the causality of OA in emerging topics, understand those topics' defining characteristics, and examine the larger academic, economic, and societal impact they foster. The methodology, articulated across distinct analytical facets, combines various tools and processes to build an end-to-end storyline. The table below summarises our approach.

Analysis Category	Objective/Description	Tools & Data
Group Selection & Control Testing	Partition publications into designated groups (see causality narrative) and run tests verifying controls.	Comparative statistics *OpenAIRE Graph*[32] & *CORDIS*[33]
Topic Recognition	Classify publications to domains, categories and topics (Fields of Science (FoS) levels 1 through 5), identify AI and Climate publications within Groups A, B and C.	SciNoBo toolkit[34] *OpenAIRE Graph*
Emerging Topics	Identification of emerging topics and their attribution to Groups A, B and C, and thus to to OA.	SciNoBo Toolkit *OpenAIRE Graph*
Multi-disciplinarity	Survey the intersection of various research disciplines underpinning the emerging topics.	Discipline analysis *OpenAIRE Graph*
SDG Relevance	Determine the relation of emerging topics to the SDGs	SDG classifier *OpenAIRE Graph*
Visibility Analysis	Ascertain the emerging topics' academic prominence and acceptance.	Citation analysis *OpenAIRE Graph*
Women's Representation	Survey the presence of women as (first) authors in publications of emerging topics.	Gender detection system *OpenAIRE Graph*
Impact Analysis	**Academic:** topic analysis on sustainability, dispersion and ensurance of emerging topics attributed to OA.	Citation tracking and analysis. *OpenAIRE Graph*
	Economic: track emerging topics in industry and skill demand	Skill detection system, SciNoBo Toolkit

[32] https://graph.openaire.eu/

[33] https://cordis.europa.eu/

[34] https://doi.org/10.3389/frma.2023.1149834

		EURAXESS[35], Company Websites
	Societal: track emerging topics in policy documents	SciNobo Toolkit *Dataset TBD*

Table 6: OA methodological approach

6.5. Next steps

Before diving into data collection and analysis, we will first convene our first case study focus group. This meeting brings together experts from climate change, Open Science and gender, as well as data scientists and economists.

In the focus group, we will present the study's objectives and the logic behind our approach. We are looking for feedback on any dynamics we might have missed and to understand how our study aligns with their field expertise. This will help us refine our impact pathway logic and solidify the causality narrative. We will discuss the impact targets, prioritising them and focusing on datasets and connections across different elements of the pathway model. This expert-in-the-loop approach will be key in making sure the case study is both relevant and convincing.

[35] https://euraxess.ec.europa.eu/

7. Case 6: Impact of open bioinformatics resources on industry

7.1. Introduction

ELIXIR unites Europe's leading life science organisations in managing and safeguarding the increasing volume of data generated by publicly funded life sciences research. It coordinates, integrates and sustains bioinformatics resources (databases, software and tools, workflows, standards, ontologies, cloud computing, and training) across the ELIXIR Member countries[36]. This ensures that life scientists in academia and industry, within and beyond Europe, can access resources that are vital for their research.

Open Science is a founding principle of the ELIXIR research infrastructure; hence all its resources are open by design. ELIXIR's funders[37] (including various funding schemes of the EU, Member countries of ELIXIR within and beyond the EU, various foundations and trusts, etc.) typically support this principle, and often require that Open Science is practiced. ELIXIR was a contributor to the seminal publication describing the "FAIR" (Findability, Accessibility, Interoperability, and Reusability) principles (Wilkinson et al., 2016), which support the Open Science principles.

Previous ELIXIR-led work has shown that Small and Medium-sized Enterprises (SMEs) use public-funded open/FAIR resources like those provided by ELIXIR as business models (Garcia et al., 2018), and for creating innovative added-value products and services that they sell to larger industry clients (Lauer et al., 2021).

The results of these studies suggest that socioeconomic and societal benefits are generated by ELIXIR resources. In PathOS, we aim to investigate these pathways (and underlying logic) in finer detail, including looking at causality aspects. The 'bioinformatics case study' will hence build on and expand the previous ELIXIR-led work to unravel the whole impact pathway and shed light on:

- the effect of ELIXIR's open resources on fostering innovation in the industry sector, and
- how this then translates to identified/quantified socioeconomic and societal benefits (impacts).

Put more simply, this case study aims to examine the use/uptake of ELIXIR resources (which are open) by industry and what socioeconomic/societal value is generated by this. For clarity, the bioinformatics case study does not aim to look at the uptake of Open Science practices by

[36] https://elixir-europe.org/about-us/who-we-are/nodes

[37] https://elixir-europe.org/about-us/how-funded

the industry sector. The industry sector can freely use ELIXIR's databases/tools/standards/etc., (since they are open and follow FAIR principles) but develop/maintain their own internal databases and other tools that may or may not abide by Open Science principles.

One particularity of ELIXIR, compared to most research infrastructures, is that the users of its resources can, and are, located anywhere on the planet. Users only require an internet connection to access ELIXIR's bioinformatics resources: they do not need to register, nor apply, nor pay, to use ELIXIR's resources[38] — this is atypical in the world of research infrastructures. Whilst this takes Open Science to a new level, this also makes it more challenging to know who those users are and how the resources assist their work, and hence which outcomes and impacts are generated. Evidence of such outcomes and impacts, notably of socioeconomic/societal nature, are crucial to ELIXIR's long-term sustainability as a public-funded research infrastructure that is not permitted to generate revenue from the use of its resources.

7.2. Impact pathway logic

The Open innovation ecosystem plays a pivotal role in addressing the grand challenges in life sciences, with publicly available data resources at the core of the ecosystem. Intergovernmental, publicly funded infrastructures like ELIXIR contribute substantially to this ecosystem by offering a wide array of resources and services that are freely accessible to users in both academia and industry. These resources are crucial for advancing research and driving economic growth, particularly considering the expanding market size for life sciences and the growing demand for computational tools and people to manage, analyse, and interpret the ever-increasing volume of data in life sciences. The contributions of the publicly funded infrastructures are not only substantial 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.

Discussions early in the project helped refine the **scope** of the bioinformatics case study to focus on four ELIXIR resource '**types'** (i.e., categories, Table 7) as opposed to considering '**named'** resources. This high-level categorisation was deemed helpful in streamlining and simplifying ELIXIR's extensive portfolio of bioinformatics resources (more than 400 in total) into something more manageable for taking the case study forward as part of Work Package 3, in which ELIXIR is considered more general. In contrast, during the collaboration with Work Package 4 (Cost-Benefit Analysis), named resources (such as UniProt[39], a knowledge base) will be looked at in detail.

[38] Some exceptions exist, such as for sensitive human data, but this is beyond the scope of the case study.
[39] https://www.uniprot.org/

	DESCRIPTION
Deposition database	Deposition databases are used by life scientists, within and beyond ELIXIR, to 'deposit' the data they have created as part of their research work. Examples of such data include nucleotide sequences (deposited in the European Nucleotide Archive) or protein sequences (deposited in the Protein Data Bank). Other life scientists are then able to re-use the deposited data as input into their own research. See examples at https://elixir-europe.org/platforms/data/elixir-deposition-databases
Knowledge base	Knowledge bases are dynamic bodies of scientific knowledge, which life scientists refer to as part of their research work. Examples include Uniprot, a comprehensive resource of protein sequence and functional information, and Ensembl, a centralised resource on vertebrate and model organisms
Interoperability resource	Interoperability resources support the delivery of FAIR principles, for instance through establishing connections between data and other resources or acquiring and exposing metadata. Examples include the Ontology Lookup Service, a single access point to the latest biomedical ontology versions, and FAIRsharing, a registry of data standards and repositories
Software and tools	A number of ELIXIR resources help life scientists find, register and benchmark software and tools. Examples include bio.tools, a registry of software and tools, and WorkflowHub, a registry for life science workflows

Table 7: High-level categorisation os some of ELIXIR's >400 bioinformatics resources

Figure 11 shows an impact pathway logic for an 'ELIXIR deposition database', one of the four resource types considered for the case study. A version can be downloaded for detailed viewing.[40] Input and feedback from PathOS colleagues led to this version of the pathway logic, and it is anticipated that similar ones will be built for the other resource types (Table 7).

[40] https://drive.google.com/file/d/1jVhfi6Ax3std82jznFjoeTODO99jzH7z/view?usp=drive_link

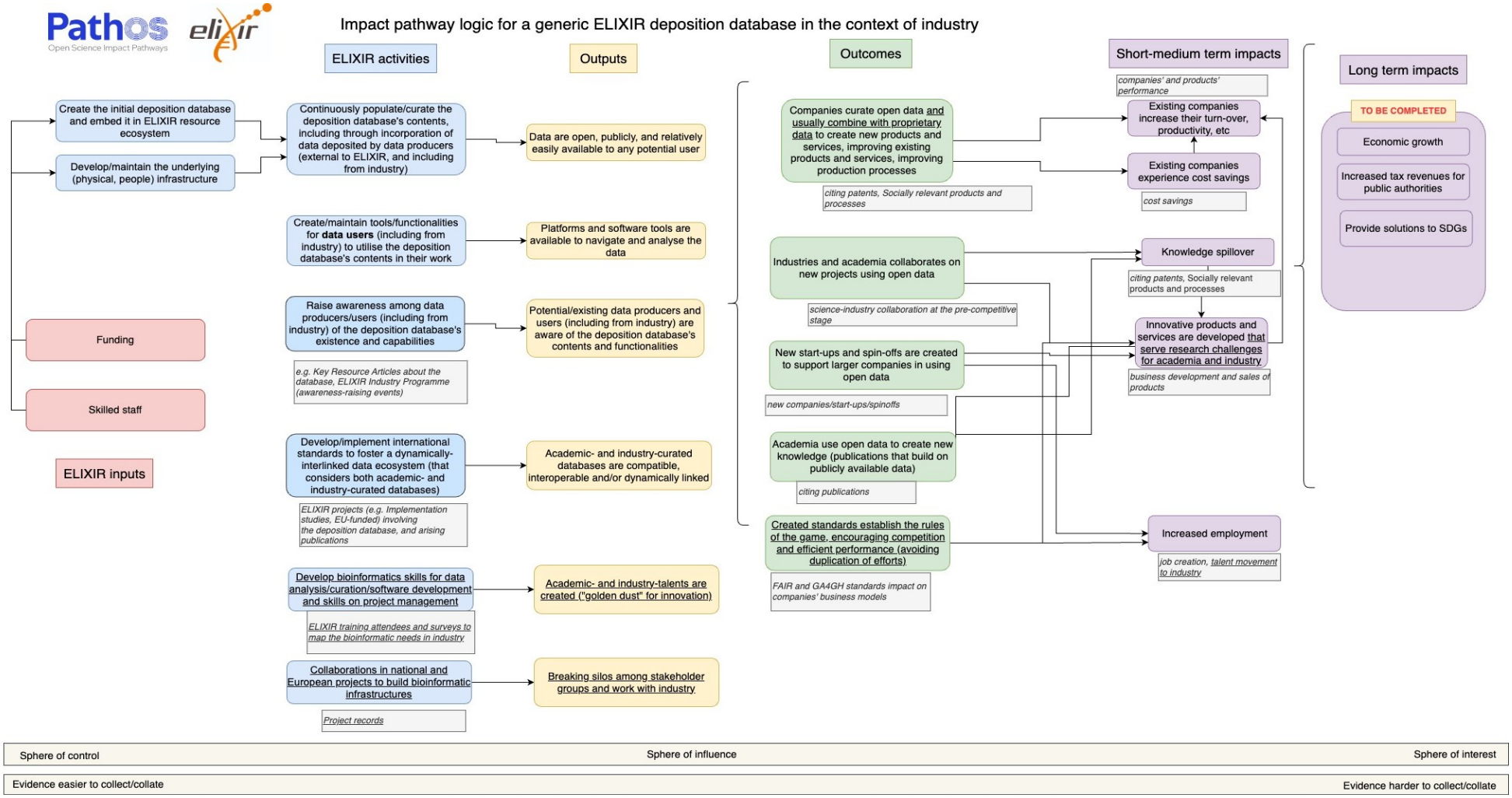

Figure 7: Impact pathway logic for an 'ELIXIR deposition database', one of the four resource types considered for the case study.

7.3. Impact targets

The impact targets, i.e., indicator themes, of the bioinformatic case study encompass science, economy, and society (Table 8). Due to the industry focus of the case study, the economy features more prominently, and this is where more effort will be expanded during the remainder of PathOS.

	SHORT-TERM	MEDIUM-TERM	LONG-TERM
Science	Researcher skills development/upgrade for academia and industry.	Public-private collaborations via scientific projects.	Sustainability of bioinformatics resources and equity in access for all. Establishments of standards in data sharing.
Economy	Usage of ELIXIR resources by industry. Precompetitive projects on similar challenges.	Usage of research results by industry. Innovation output (e.g., new products and processes).	Jobs creation in industry, and movement of talents between academia and industry. Creation of start-ups/spin-offs. Increase of turnover, in productivity. Increase in tax revenues.
Society			Socially relevant products and processes (e.g., bioinformatics applications of societal benefit in health, food security, the environment).

Table 8: Indicator themes (selection) for the bioinformatics case study

7.4. Causality narrative

The impact pathway logic presented in Figure 11 illustrates how ELIXIR's inputs and activities are currently understood to lead to outputs, outcomes and impacts, in the context of a generic ELIXIR deposition database that industry would use. The logic has not yet been tested and PathOS will work to provide evidence in support of causality links in this pathway logic.

Inputs/Activities

- ELIXIR, with its skilled staff and funding, plays a crucial role in the bioinformatics ecosystem. ELIXIR staff are responsible for creating, developing, and maintaining the deposition databases and associated analytical tools. These resources are not isolated; instead, they are part of a broader international ecosystem of interlinked bioinformatics

resources and open science standards. This ecosystem includes a variety of resources such as software and tools, workflows, standards, ontologies, cloud computing, and training.

- To ensure these resources remain state-of-the-art and relevant to user needs, ELIXIR undertakes awareness-raising activities. These activities promote the resources and foster a community of experts from academia and industry. This community collaborates to discuss visionary ideas, identify bottlenecks, and devise solutions to major challenges in the data-driven life science sector.
- In addition to these technical aspects, publicly funded infrastructures like ELIXIR invest in the professional development of their members, equipping them with transferable skills such as project management and awareness of different stakeholder groups' approach to life sciences, including the needs and practices of the private sector, and international perspectives.

Outputs/Outcomes

- As a result of the operation of the infrastructure and the awareness-raising activities, users (from industry and academia) can openly explore, extract and analyse existing data which are vital for their work. In 2021, ELIXIR conducted a survey involving industry users (Lauer et al., 2021), which highlighted the importance of open resources for bioinformatics companies. The study indicated that these resources are typically combined with proprietary data, providing companies with the ability to enhance their existing products and services or develop new ones, underscoring the potential for growth and innovation in industry.
    - The first Focus Group event noted that resources (such as deposition databases) are used in complex combinations (so as to construct more elaborate products), hence the importance of keeping in mind the broader ecosystem around the resource being investigated. They acknowledged the difficulty of disentangling various effects when public and proprietary data are combined.
    - The first Focus Group event also highlighted the usefulness of working in partnerships (resource providers and industry) at the precompetitive stage, so as to better foster the reuse and repurposing of data.
- Another effect to highlight relates to skills development, notably the creation of talent, seen as the "gold dust" for innovation. Open research data are indeed notoriously complex to use in the sense that they require specialised skills, which are usually obtained in academic contexts, and very prized by industry (resulting in talent movement between the two sectors).
    - The first Focus Group event highlighted this effect, which had not sufficiently been considered in earlier versions of the pathway logic.
- Industry outreach activities across ELIXIR enable industry engagement and research collaborations, notably 'via projects' or 'at precompetitive stage'. Over the years, ELIXIR

has created a record of activities, and now showcases best practices in industry engagement for public infrastructures.

- The value (to industry) of these outreach activities was highlighted during the first Focus Group event, for instance given the complexity of using resources primarily developed for academic research.

Short/medium-term and Longer-term impacts

- Short/medium-term expected impacts include increases in productivity and turn-over, knowledge spillovers and increased employment.
- In the longer-term, the above is expected to contribute to economic growth, increased tax revenues, as well as the provision of solutions of societal benefit (e.g., bioinformatics applications of societal benefit in health, food security, the environment), notably in the context of the UN Sustainable Development Goals[41].

The first Focus Group event noted that it will be challenging for PathOS to unravel and disentangle complex impact pathways to which many, beyond ELIXIR, contribute. Furthermore, participants of the Focus Group noted that innovation takes time, meaning that the temporal distance between ELIXIR's activities, such as the operation of a deposition database, and its effects through use by industry, are likely to be large, and hence more difficult to track.

7.5. Methods

The bioinformatics case study aims to identify and verify the pathway logic relating to the effect of ELIXIR's open resources on (1) fostering innovation in the industry sector, as well as (2) how this translates to identified/quantified socioeconomic and societal impacts. A range of methods is, and will be, employed to collect relevant data, information and knowledge around the set of impact targets (i.e., indicator themes) presented in Table 8 above. Methods include:

- **Review of existing literature** (building on prior ELIXIR-led work such as Garcia et al., 2018, and Lauer et al., 2021), for instance to collate concrete examples of how industry uses ELIXIR resources in their work, what value they see in doing so, and what innovation is enabled.
    - The **Focus Group events** are expected to help collect additional such examples and testimonials of how industry uses ELIXIR's resources, and indeed this already happened during the first Focus Group event (July 2023). The information collected was then used to refine and arrive at the impact pathway logic presented in Figure 11. The first Focus Group had 10 external guests (a number

[41] A number of UN Sustainable Development Goals are relevant to ELIXIR's activities, e.g. Goals #2 Zero hunger, #3 Good health and wellbeing, #4 Quality education, #5 Gender equality, #6 Clean water and sanitation, #7 Affordable and clean energy, #9 Industry, innovation and infrastructure, #13 Climate action, #14 Life below water, #15 Life on land.

having current or prior experience in the industry sector; economists; resource providers; industry officers in research infrastructures).

- Events organised by **ELIXIR's Industry Programme**[42] will also be capitalised on to collect data, information and knowledge, for instance on science-industry collaboration (via projects, or at the pre-competitive stage). This will be done through examining presentations delivered by speakers of both sectors, cross-sectoral panel discussions, as well as through interviews and informal discussions in the margins.
    - This aligns with feedback from participants at the first Focus Group event, who encouraged case study leads to continue speaking to industry, to really understand how they use and repurpose ELIXIR resources, including at what part of the value chain.
- ELIXIR's Industry Programme also has access to a **network of industry officers** located in many of its country Nodes. Insights into various topics of relevance to the case study (e.g., industry's use of ELIXIR's resources, science-industry collaborations, establishments of standards in data sharing) will be collected via dedicated discussions as part of the regular meetings of this network.
- High-level, qualitative, insights into bioinformatics applications of societal benefit (in health, food security, the environment) that are enabled through the use of ELIXIR's resources have been obtained through **text-mining** of ELIXIR resources names in lens.org, a patent and scholarly literature search facility. This has revealed that ELIXIR resources are widely mentioned by name in patent applications[43], an indication of their usefulness to bioinformaticians of all sectors (academic, industry), and across the globe.
    - This work will undergo further refinement to provide a more qualitative analysis of industry-affiliated patents and publications by extracting information about the invention, tracking the product's progress in the market (if applicable), diving deeper in the mention itself (to understand how important it was to the patent or product based on it), and examining the size and business development of the company. This analysis will provide a deeper understanding of the socioeconomic impact stemming from innovations in life sciences that are reliant on open infrastructure.
    - This aligns with feedback received as part of the first Focus Group event, during which participants encouraged the case study leads to use a mix of qualitative and quantitative approaches, with the latter providing signals that are best investigated using the former.
- Alumni networks and online databases (such as LinkedIn) will be used to gather data on talent movement between academia and industry.

[42] Innovation and SME Forums https://elixir-europe.org/industry/forums ; Bioinformatics Industry Forums https://elixir-europe.org/industry/suppliers-forum

[43] ELIXIR's contribution to innovation https://elixir-europe.org/about-us/impact/patents

- Working with economists from Work Package 4, as part of the Cost-Benefit Analysis (CBA), **interviews, surveys and questionnaires** will be employed to gather the necessary data. CBA is indeed a data-intensive methodology which will require, for the case study, to quantify and, as far as possible, monetise effects attributable to a selected ELIXIR resource. Several of the indicator themes are expected to be derived from this cross-Work Package collaboration, e.g., jobs creation in industry, creation of start-ups/spin-offs, increase of turnover, increase in productivity, and increase in tax revenues.

7.6. Next steps

Next steps for the bioinformatics case study include:

- Take stock of the work accomplished, and results achieved, to date to inform the structure of the second Focus Group:
  - For instance, both new and recurrent attendees may appreciate a summary of the outcomes from the first event and how the insights provided[44] were taken on-board by the case study team.
  - The impact pathway logic, along with preliminary results on indicator themes (presented in Table 8) could be presented for feedback.
  - A set of questions needs to be drafted so as to fill gaps in the case study.
- In parallel, continue the work required to assemble data, information and knowledge relating to the indicator themes.
  - For a number of these (especially the long-term economic and societal ones), this will be done through a collaboration (with colleagues in Work Package 4) on the Cost-Benefit Analysis of a selected ELIXIR resource.
  - Where relevant, which is especially the case of the simpler indicator themes that ELIXIR can monitor themselves (even after the end of the funded phase), results will be added to ELIXIR's Impact Dashboard[45]. This will represent an important online legacy for the case study, as well as being used in ELIXIR's efforts towards the long-term sustainability of its research infrastructure and the 'free at the point of use' resources that it offers.

[44] https://pathos-project.eu/a-year-in-review-the-pathos-focus-groups-key-takeaways

[45] https://elixir-europe.org/about-us/impact

8. Synthesis

This chapter aims to synthesise the foci, aims, impact targets and pathways to impact studied by the six case studies described in prior chapters.

Four key OS aspects are in focus across the case studies: Open Access, Open Data, Open-Source Code (OSC), and Open Materials (OM) (see Table 9). The interventions (or resources) studied that foster sharing OS resources are primarily government-sponsored portals/repositories, except for the European Commission's OA and OD mandates, which went into effect with the Horizon 2020 framework programme. From these, our case studies aim to quantify academic, economic and societal short-term impacts. These include the academic impacts of citation advantage (from OA and OD), data reuse, and collaborations; the economic impacts of academic-industry collaborations and the industry use of OS resources; and the societal impacts of the use of OS resources by various societal actors (policymakers, media, CSOs, healthcare providers, etc.), and the fostering of gender equality within the research realm.

OS aspect	Intervention	Academic impact	Economic impact	Societal impact
OA	RCAAP OA portal; Open Edition, HAL, RDG, EC mandate	Citation advantage; collaboration	Academic-industry collaborations; industry use of OA	Societal use of OA (public institutions, CSOs, media); gender inclusive science
OD	RDG, ELIXIR, EASY, EC mandate	Data reuse, citation advantage; collaboration	Industry use of OD	Societal use of OD (public institutions, CSOs, media); gender inclusive science
OSC	HAL, RDG, ELIXIR		Industry use of OSC	
OM	ELIXIR		Industry use of OM	

Table 9: Cumulative areas of focus across the case studies

The medium and longer-term impacts studied across our case studies include:

Academic

- Advancements in interdisciplinary research
- Increased speed of knowledge development
- New emerging topics and areas of study

Economic

- Industry and economic growth
- New industry sector development
- Industrial innovation
- Faster product development
- Skill demand shift

Societal

- Societal benefits from product and service development
- Job creation
- Greater public awareness of, and trust in, science
- Greater progress toward SDGs
- Advancements in gender equality

Across the case studies, the following impact-enabling factors have been identified:

- Legislation mandating OA and/or OD publication and/or the submission of research outputs in open repositories
- Science policy that mandates and/or strongly encourages OA and/or OD publication and/or the implementation research output submission in open repositories
- Policy and legislation that funds and creates OS platforms
- OS platforms
- Staff and resources to maintain, curate resources, and promote engagement with OS platforms
- Training for using OS platforms

Much of what is described above was generated through research and theorizing on the part of our case study teams, but some parts were identified through case study focus groups conducted through 2023. Importantly, the focus groups, conducted with relevant expert stakeholders selected based on the content and geography of each case, provided important feedback on research methods.

Specific to Case 1, which focuses on academic-industry collaborations in Portugal, focus group participants suggested that studying social networks rather than citation patterns might reveal more about collaborations, and that patent analysis could prove useful.

For Case 3, which focuses on SSH data reuse in the Netherlands, focus group participants suggested that a variety of methods might be necessary to measure data reuse because citation practices vary widely both across and within research disciplines, and data reuse might not always be acknowledged through formal citations.

Finally, for Case 6, which focuses on bio-medical industry use of OS resources, focus group participants recommended that qualitative methods could prove more useful than quantitative methods in this case, given that companies tend to mix proprietary and OS resources in research and product development. Because of this, quantitative measures of impact could be

difficult to achieve. Focus group participants suggested interviews with industry stakeholders to gauge industry use of OS resources and their impacts, and suggested focusing on human factors, like staffing crossovers between industry and OS resource providers. Further, focus group participants that resource use differs between small and large companies, and therefore separating them for analytic purposes is important.

The contributions from the three focus groups that have been hosted to date have shown our case teams that measuring any kind of impact is challenging and complex, but they have also provided critical insights into steps the teams can take to meet this challenge.

9. Next steps

Currently, all case studies are proceeding with the development of causality narratives and impact pathway models. Cases 1, 3, 5 and 6 are doing so with the lessons learned through the first round of focus groups, while cases 2 and 4 are scheduling their first-round focus groups for autumn 2023. Following this, the second round of case study focus groups will take place in the winter and spring of 2024. At that stage, focus group discussions will focus on the development of impact indicators for each case. Case teams will then build on focus group contributions to further develop impact indicators prior to operationalising them. Finally, in spring of 2025, a third round of focus groups will be held to assess the results of the operationalisation of the OS impact indicators, provide insights on the causal/bias effects of the selected measurements, and for connecting them to the enabling factors.

Throughout the duration of the project, each case study will feed results and insights into other project tasks and outputs, including the OS Indicator Handbook, Cost-Benefit Analysis for selected case studies, and the creation of data and tools for the long-term evaluation of OS, and a final report on the operationalization of impact indicators within the case studies.

10. References

Briant Carant, J. (2017). Unheard voices: A critical discourse analysis of the Millennium Development Goals' evolution into the Sustainable Development Goals. *Third World Quarterly*, *38*(1), 16–41. https://doi.org/10.1080/01436597.2016.1166944

Dekker, R., Karasz, I., & Stoy, L. (2023). *PathOS - D1.1 Open Science Intervention Logic*. Zenodo. https://zenodo.org/record/7801286

Doorn, P. (2020). Archiving and Managing Research Data—Data services to the domains of the humanities and social sciences and beyond: DANS in the Netherlands. *Der Archivar*, *73*(1), 44–50.

Fehling, M., Nelson, B. D., & Venkatapuram, S. (2013). Limitations of the Millennium Development Goals: A literature review. *Global Public Health*, *8*(10), 1109–1122. https://doi.org/10.1080/17441692.2013.845676

Garcia, P. R., Smith, A., & Blomberg, N. (2018). Public data resources as a business model for SMEs. The Role of Public Bioinformatics Infrastructure in supporting innovation in the life sciences. *F1000Research*, *7*(590), Article 590. https://doi.org/10.7490/f1000research.1115445.1

Gregory, K., Ninkov, A. B., Ripp, C., Roblin, E., Peters, I., & Haustein, S. (2023). *Tracing data: A survey investigating disciplinary differences in data citation*. Zenodo. https://doi.org/10.5281/zenodo.7865097

Griniece, E., Angelis, J., Reid, A., Vignetti, S., Catalano, J., Helman, A., Barberis Rami, M., & Kroll, H. (2020). *Guidebook for Socio-Economic Impact Assessment of Research Infrastructures*. https://zenodo.org/record/3950043

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*. Zenodo. https://zenodo.org/record/7883699

Kopnina, H. (2016). The victims of unsustainability: A challenge to sustainable development goals. *International Journal of Sustainable Development & World Ecology*, *23*(2), 113–121. https://doi.org/10.1080/13504509.2015.1111269

Lauer, D. K. B., Smith, A., Blomberg, D. N., Sitjà, X. P., Talbot-Cooper, C., Rothe, P. H., & Conde, D. S. (2021). Open data: A driving force for innovation in the life sciences. *F1000Research*, *10*(828), Article 828. https://doi.org/10.7490/f1000research.1118745.1

Pearl, J., & Mackenzie, D. (2019). *The Book of Why*. https://www.penguin.co.uk/books/289825/the-book-of-why-by-judea-pearl-and-dana-mackenzie/9780141982410

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

Traag, V. A. (2021). Inferring the causal effect of journals on citations. *Quantitative Science Studies*, 1–9. https://doi.org/10.1162/qss_a_00128

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