Enhancing Trust, Integrity, and Efficiency in Research through Next-Level Reproducibility Impact Pathways

Deliverable D4.3 – Pilot implementation reflection report including assessment of efficacy & recommendations for future developments

28/11/2025

Lead Beneficiary: AmsterdamUMC (VUmc)
Author/s: Barbara Leitner, Joeri Tijdink, Friederike Elisabeth Kohrs, Alexandra Bannach-Brown, Sven Arend Ulpts, Fakhri Momeni, Eleni Adamidi, Thomas Klebel, Eva Kormann, Adrian Marangoni, Jesper W. Schneider, Allyson L. Lister, Elli Papadopoulou, Haris Papageorgiou, Petros Stavropoulos, Stefania Amodeo, Thanasis Vergoulis, Susanna-Assunta Sansone, Tony Ross-Hellauer

Reviewer/s: Aarhus & Know Center

Funded by
the European Union

D4.3 Pilot implementation reflection report including assessment of efficacy & recommendations for future developments

Prepared under contract from the European Commission
Grant agreement No. 101094817

EU Horizon Europe Research and Innovation action

Project acronym:	**TIER2**
Project full title:	**Enhancing Trust, Integrity, and Efficiency in Research through Next-Level Reproducibility Impact Pathways**
Start of the project:	January 2023
Duration:	36 months
Project coordinator:	Dr. Tony Ross-Hellauer

Deliverable title:	Pilot implementation reflection report including assessment of efficacy & recommendations for future developments
Deliverable n°:	D4.3
Version n°:	1.2
Nature of the deliverable:	Report
Dissemination level:	Public

WP responsible:	WP4
Lead beneficiary:	AmsterdamUMC

TIER2 Project, Grant agreement No. 101094817

Due date of deliverable:	Month n°35
Actual submission date:	Month n°35

Deliverable status:

Version	Status	Date	Author(s)
1.0	Draft	07.11.2025	Barbara Leitner, Joeri Tijdink, Friederike Elisabeth Kohrs, Alexandra Bannach-Brown, Sven Arend Ulpts, Jesper W. Schneider, Fakhri Momeni, Eleni Adamidi, Elli Papadopoulou, Haris Papageorgiou, Petros Stavropoulos, Stefania Amodeo, Thanasis Vergoulis, Thomas Klebel, Eva Kormann, Adrian Marangoni, Allyson L. Lister, Susana-Assunta Sansone, Tony Ross-Hellauer AmsterdamUMC, Charite, Aarhus,

			GESIS, OpenAIRE, Athena Institute, KNOW, UOXF
1.1	Review	14.11.2025	Sven Ulpts, Thomas Klebel Aarhus, KNOW
1.2	Final	28.11.2025	Barbara Leitner, Joeri Tijdink, Friederike Elisabeth Kohrs, Alexandra Bannach-Brown, Sven Arend Ulpts, Jesper W. Schneider, Fakhri Momeni, Eleni Adamidi, Elli Papadopoulou, Haris Papageorgiou, Petros Stavropoulos, Stefania Amodeo, Thanasis Vergoulis, Thomas Klebel, Eva Kormann, Adrian Marangoni, Allyson L. Lister, Susana-Assunta Sansone, Tony Ross-Hellauer AmsterdamUMC, Charite, Aarhus, GESIS, OpenAIRE, Athena Institute, KNOW, UOXF

Table of contents

Executive Summary

Deliverable 4.3 presents the evaluation procedure of the Pilots conducted in TIER2, and the process of design, testing, and evaluating the eight Pilots developed to address the needs of key stakeholder groups within the project's scope. The tools were created to enhance collaboration, knowledge sharing, and practical implementation of project objectives, with stakeholder groups selected based on their strategic relevance, operational impact, and capacity to generate meaningful feedback for tool refinement.

Eight Pilot activities were conducted to test the tools across diverse contexts and user environments. Each Pilot focused on specific application areas, allowing for targeted evaluation of usability, effectiveness, and scalability. The Pilots covered a range of themes—from data integration and policy support to community engagement and technical capacity building—reflecting the project's holistic approach to stakeholder engagement and real-world validation.

Comprehensive evaluations were performed for each Pilot, assessing both the individual outcomes and the synergies between them. This analysis demonstrated that cross-pilot learning and collaboration significantly enhanced the overall impact of the tools, creating a more coherent ecosystem of solutions. The findings highlight the importance of adaptability, user-centred design, and continuous feedback mechanisms to ensure sustained relevance and utility.

Looking ahead, Deliverable 4.3 outlines the future direction for tool improvements, further developments, recommendations and scalability. Recommendations emphasize the need for ongoing technical support, structured governance for tool ownership, and strategic partnerships to ensure long-term sustainability. The deliverable concludes by underscoring the potential for these tools to be integrated into broader frameworks and policy processes, thereby amplifying their contribution to innovation, collaboration, and impact within the project's domain. These recommendations will also lead to several overarching recommendations on the project level (and beyond).

List of Abbreviations

CA – Citance Analysis
CSS – Computational Social Science
DAS – Data Availability Statement
DMP – Data Management Plan
EM – Exact Match
EU – European Union
FAIR – Findable, Accessible, Interoperable, Reusable
FWCI – Field-Weighted Citation Impact
FWRI – Field-Weighted Reusability Index
KPI – Key Performance Indicator
KPMs – Knowledge Production Modes
LM – Lenient Match
maDMP – Machine Actionable Data Management Plan
NWO – Netherlands Organization for Scientific Research
OSF – Open Science Framework
PID – Persistent Identifies
RAA – Research Artefact Analysis
RCCI – Reproducibility Composite Confidence Index
RCI – Reproducibility Confidence Indicator
RDA – Research Data Alliance
RCT – Randomised Controlled Trial
RFOs – Research Funding Organizations
RI – Reusability Index
RMP – Reproducibility Management Plan
RPOs – Research Performing Organizations
RPP – Reproducibility Promotion Plan for Funders
RPs – Reproducible Practices
RN – Reproducibility Network
SE – Science Europe
TESK – Task Execution Engine
TOP – Transparency and Openness Promotion Guidelines
WG – Working Group

1. Introduction

Description and evaluation of Pilots fostering Reproducibility Practices (RPs)

This Deliverable reports on the design, implementation, and evaluation of eight Pilots aimed at fostering reproducibility tools and practices across the European research landscape. Developed within the TIER2 project, these Pilots address challenges in making research transparent, verifiable, and reusable across diverse epistemic traditions, institutional environments, and disciplinary cultures. The objective of this deliverable is to present the rationale and aims of the Pilots, document their development and implementation process, and assess their effectiveness and transferability to broader research contexts.

1.1. Aim and objective of this deliverable

The overarching aim of the Pilots was to create and test practical solutions that enhance reproducibility in research practice. Specifically, the Pilots aimed to:

1. Translate conceptual understandings of reproducibility into concrete tools, workflows, and policy approaches;
2. Co-create solutions with researchers, funders, and publishers to ensure contextual relevance and usability;
3. Evaluate the feasibility, acceptability, and potential for adoption of these solutions in real-world settings;
4. Identify potential barriers, enablers, and insights to inform sustainable implementation beyond the project's duration that will eventually result in concrete recommendations.

1.2. Overview of the Pilots

The eight Pilots collectively span key steps in the research lifecycle, different epistemic contexts/scientific domains, and key actors responsible for shaping research culture. They include conceptual decision tools (Pilot 1), research planning instruments (Pilots 2 and 5), technical infrastructures for computational reproducibility (Pilots 3 and 4), a monitoring dashboard (Pilot 6), behavioural interventions in scholarly communication (Pilot 7), and editorial policy practices (a handbook; Pilot 8). Together, the Pilots illustrate how reproducibility can be strengthened through aligned action at multiple levels: researchers, infrastructures, funders, and publishers.

- **Pilots 1–4** focused on supporting researchers and research teams by developing frameworks, platforms, and checklists to embed reproducibility into the design, execution, and documentation of research.
- **Pilots 5–6** addressed funders and institutions by providing policy guidance and monitoring tools that enable systematic support for reproducible practices.
- **Pilots 7–8** targeted journals and publishers, examining how editorial workflows can encourage data sharing and the transparent reporting of digital research objects.

1.3. History of the Pilot development and cocreation process

The Pilots followed a co-creation approach, involving stakeholders in iterative rounds of design, testing, and refinement. Co-creation activities included structured workshops, interviews, surveys, usability testing, and stakeholder engagement. This process aimed to ensure that each Pilot responded to the needs and constraints of real research environments, while allowing adaptations to diverse disciplinary and institutional settings. Ethical approvals were obtained where required, and all co-creation processes adhered to principles of inclusivity and reflexivity.

The Pilots evolved over time. While some experienced difficulties in reaching their original planned scope (e.g., the development of a decision-support prototype in Pilot 1), others shifted direction to accommodate stakeholder feedback or feasibility considerations (e.g. Pilot 6, the monitoring dashboard that can also be used by institutions and eventually publishers). This iterative process reflects the project's commitment to practical relevance and real-world applicability.

1.4. Process of evaluation

Evaluation across the Pilots combined qualitative and quantitative measures, tailored to each Pilot's design and user community. AmsterdamUMC led the process through monitoring the 8 Pilots and encouraging them to collaborate in their efforts. Methods of evaluation included interviews, surveys, controlled experiments, adoption tracking, workflow testing, and implementation case studies. The purpose of evaluation was not only to assess effectiveness, but also to identify barriers and enablers for conditions necessary for successful uptake and long-term sustainability of these (co)created tools.

Across Pilots, three cross-cutting themes emerged as central to fostering reproducible research:

- Embedding reproducibility into existing workflows rather than adding new administrative burdens;
- Aligning reproducibility practices with cultural and institutional incentives;
- Ensuring community engagement and cocreation of policy and implementation to sustain change.

1.5. Concluding Remarks

Together, we believe that the eight Pilots demonstrate how reproducibility can be operationalized across diverse research contexts, from individual laboratory workflows to funders policies and journal practices. Co-creation and iterative evaluation were put into place to assure that the solutions are feasible, adaptable, and sensitive to epistemic diversity, disciplinary norms, and resource constraints.

The insights gained from these Pilots provide a foundation for broader adoption of reproducibility practices, informing future initiatives, policy frameworks, and research infrastructure development. By documenting successes, challenges, and lessons learned, Deliverable 4.3 contributes not only to the immediate goals of the TIER2 project, but also to the long-term vision of a European

research ecosystem in which reproducible, transparent, and trustworthy science becomes standard practice. We aim to include the recommendations that follow from the Pilots in Deliverable 3.2 (TIER2 synthesis and recommendations).

2. Pilot 1 - Decision Aid: Relevance and Feasibility of Reproducibility

Author: Jesper Wiborg Schneider, Sven Ulpts

2.1. Introduction

The main purpose of TIER2 Task 3.1 was to establish a conceptual framework for reproducibility acknowledging epistemic diversity and different research settings. The work comprised of two components: First, an examination of definitions and understandings of reproducibility across research areas (Ulpts & Schneider, 2024), and second, an examination of how the appropriateness of reproducibility depends on epistemic diversity and specific research settings (Ulpts & Schneider, 2023). The findings are summarised in Deliverable D3.1.

Several attempts have been made to frame reproducibility by classifying types of research according to approaches, methods, or fields (e.g. Leonelli, 2018; Penders et al., 2019; Tuval-Mashiach, 2021). While informative, these framings remain limited because they focus narrowly on methodological design, while neglecting the epistemological dimension of knowledge production. This dimension is essential for understanding whether reproducibility is relevant to a given kind of research. Since epistemic traditions can vary within and across methods or fields, neither of these offer a suitable unit of analysis.

We therefore proposed knowledge production modes (KPMs) as an alternative analytical framework. KPMs capture both the epistemic and social aspects of knowledge production. They are local in the sense that they are organized around a particular subject matter, an epistemic orientation, and preferred methodologies within a specific research situation, but they can also scale up to form parts of research specialties.

Our framework enables assessment of the appropriateness of different forms of reproducibility for diverse research situations. Appropriateness has two components: **relevance** and **feasibility**. Relevance is assessed primarily on epistemic factors: the aims and ways of knowing that guide the research (its epistemology), the criteria and practices that establish quality and trustworthiness, and the research goals. Importantly, goals can also be non-epistemic and override epistemic considerations, for example when commercial or proprietary interests motivate or constrain the work. Feasibility, in contrast, depends on practical aspects: the nature and complexity of the subject of investigation (e.g. whether it is stable or dynamic, interacting with its environment or relatively independent), the degree of uncertainty involved, and the resources required. Uncertainty has two dimensions. Theoretical uncertainty refers to how well the subject matter is understood and how far such understanding can guide investigation. Methodological uncertainty concerns how well the methods and procedures themselves are understood, used, and justified.

The framework therefore is a more comprehensive proposal for an analytical tool which can address pertinent epistemic and social questions in relation to the appropriateness of

reproducibility, taking epistemic diversity and research contexts into consideration. In that sense, it became TIER2's conceptual framework for reproducibility across different contexts.

From this conceptual work came the idea to try to operationalise it in the form of a guided decision tool that could help indicate whether reproducibility was relevant and, if so, to what extent it was feasible in the given context. None of this was pre-planned and written into the TIER2 application. The idea emerged at the end of the work in T3.1 and as such the resources available for development were limited.

2.2. Development of prototype

To begin with, we did not commit to a specific stakeholder but sought a development that primarily reflected the idea behind the framework.

Our first step was to transform the KPM framework into a prototype schema. This schema functioned as a decision tree: starting from a general question, users were directed to the next relevant level of the tool depending on their answers. An overview of the schema is provided in Figure 2.2.1.

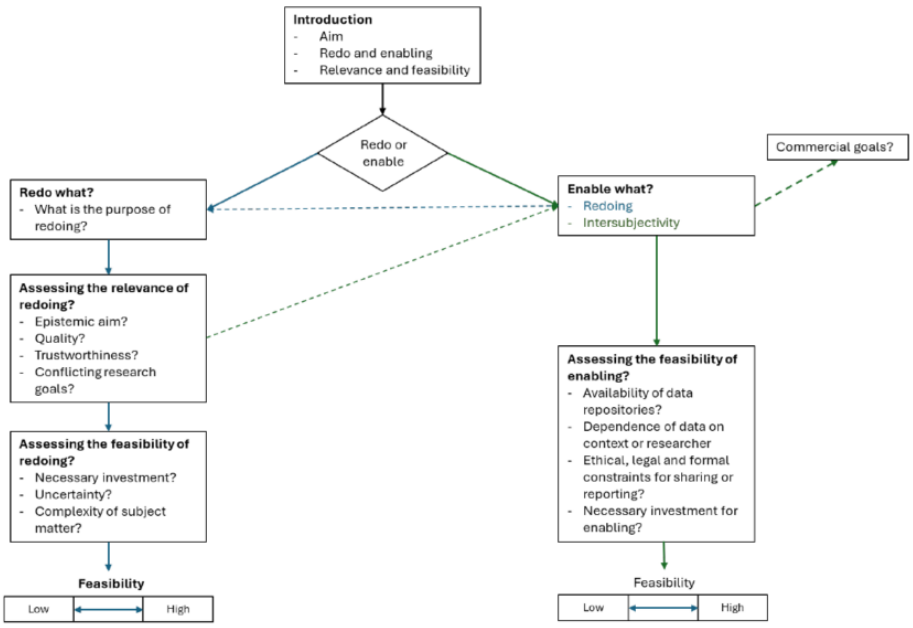

Figure 2.2.1. *A simplified schematic version of the prototype decision tool.*

To address the conceptual confusion documented in Ulpts and Schneider (2024), we used two terms within the prototype: **redoing**, which covers reproducibility, replicability, repeatability, and similar terms; and **enabling**, which essentially translates to transparency. Remember, the idea was that the aid should support epistemic diversity and thus different aims and epistemic functions.

The main point is that a user is introduced to the aid and then asked whether the aim is to redo or to enable something, either future redoing, or future intersubjectivity with the research at hand. Depending on the answer, the user follows different paths, eventually establishing whether redoing is relevant, and, if so, to what extent it is feasible. Likewise, if the aim is enabling, the user is given a set of questions that establish the degree to which enabling is feasible.

In short, the prototype consisted of two components: Relevance and feasibility. The basic idea in the relevance component was to get rid of qualifiers, such as direct replication, and instead get the user to map the intended epistemic function (purpose) to the actual parts of the research which should be varied or kept the same (the actual practices). For this, the component and an epistemological part that mapped functions to practices are presented, but also a part that queried the actual goal of the research. Eventually, relevance was established, and the user would either continue to feasibility, or if not relevant, to enabling or termination.

The feasibility component consisted of three parts: one that examined the complexity of the subject matter; one that assessed the resources available; and one that surveyed the presumed methodological and theoretical uncertainties associated with the research. The answers were graded, and a final feasibility assessment was provided.

The prototype therefore included both epistemic and social dimensions of knowledge production and, in principle, supported epistemic diversity. Numerous different aims and functions of redoing could be explicitly expressed; other aims, such as enabling or transparency, could likewise be made explicit; and concerns about actual feasibility could also be explicitly addressed

2.3. Operationalisation and cognitive testing

Initially, the individual elements of the prototype were cognitively tested by an epistemically diverse set of researchers employed at the Department of Political Science at Aarhus University, Denmark. What they tested were the individual components, as the functionality was operationalized later. We conducted a number of reruns, simplifying the questions to reduce the cognitive load on respondents. When deemed suitable, the schematic prototype was fully operationalized

It was operationalized using PHP (Hypertext Preprocessor), an open-source scripting language especially suited for web development. It runs on the server side, generating dynamic content that is sent to the user's browser. The server website is available at https://cfa-research.au.dk/tier2/index.php. Please note that although this website is currently incomplete, it is still used for development and testing, and its content may therefore change. Several versions have been developed and tested, focusing on making a functional app intended for piloting among stakeholders. However, this never materialised as we will discuss below.

2.4. Discussion

After a challenging phase implementing functionality in the prototype, we conducted another round of cognitive testing. This revealed that the decision tool had several issues. Most importantly,

optimal use presupposed substantial knowledge of methodology and epistemology, as well as detailed familiarity with the specific research under consideration. In other words, the complexity was high—indeed too high—given that the intended stakeholders included funders, i.e., external users tasked with evaluating a piece of research (e.g., a grant application). The testing also indicated issues with the redoing and enabling components, particularly the relevance component. While the component worked as intended, it appeared not to be well aligned with the unit of analysis, namely 'a specific piece of research'.

It turns out that the KPM framework's level of analysis plays a non-trivial role for the relevance component and may be geared more toward a level at which KPM is seen as a loosely defined research community rather than a 'a specific piece of research'. The aim of acknowledging and operationalizing epistemic diversity inevitably let to a degree of complexity that required an unfeasible level of expertise from the intended users. Therefore, while we think that it is still an insightful and useful analytical aid, it might be a too fine-grained unit of analysis to operationalize into a workable tool. Eventually, the conclusion from these tests was that the prototype needed to be simplified if it were to be used with stakeholders. Eventually, the idea was to convert the relevance component into a kind of policy brief that succinctly described the epistemic challenges KPMs may face and, consequently, that redoing (i.e., reproducibility) is not necessarily relevant for all KPMs—and that this should be acknowledged. To reduce complexity, it was proposed to narrow the tool to feasibility only, and only feasibility in relation to potential future reproducibility. The idea was that grant applicants would complete this reduced part of the tool as part of their application, thereby producing a self-assessed feasibility evaluation that funders could subsequently use in their appraisal of proposals.

This revision of the prototype was never fully developed and therefore never reached the stage of a pilot-testing. Resources did not allow us to take this further.

We still believe the idea was worth pursuing, precisely because the analytical KPM framework strongly encourages asking questions to 'research', to understand what the actual nature and properties of the research at hand are as well as how they relate to the relevance and feasibility of reproducibility. Conversely, we must acknowledge that developing such an online tool has been highly challenging, not least due to limited resources and our limited programming expertise, but also conceptually. We also recognize that the complexity challenges the pertinence of such a tool for the intended audience. Even so, we consider it worth the attempt. Please note that we have consistently treated this as an exploratory exercise, something coming out of WP 3.1 worth pursuing and did not initially plan for piloting and evaluation—the course of action depended on how development unfolded. In that sense, the tool differs markedly from the other Pilots.

We will report on the tool and make our prototype and files available, and we will prepare a brief report outlining the main points stakeholders should keep in mind when it comes to relevance and feasibility of reproducibility and how to address it. This will be published on the TIER2 website.

3. Pilot 2 - Reproducibility Management Plan (RMP)

Authors: Elli Papadopoulou, Maria Kontopidi

3.1. Introduction

Reproducibility in research is increasingly seen as requiring attention in the research enterprise. Tools like Data Management Plans (DMPs), which public research funders have widely adopted as mandatory deliverables, have significantly contributed towards establishing best practices that can ultimately lead to reproducible results. Science Europe has attempted to harmonize DMP templates across countries and domains through practical guides for international alignment of research data management. However, traditional DMPs focus primarily on data handling and occasionally delve into the software management or equivalent complementary research processes, leaving significant gaps in addressing comprehensive reproducibility needs.

Recent efforts have broadened the scope of DMPs to encompass software management[1] [2] machine learning algorithms, and other research outputs (Grossmann et al., 2024; Gebru et al., 2018). Despite these advances, services that support researchers in writing and actively managing their plans remain limited, and the publishing of planning outputs in scholarly communication channels falls short. Studies acknowledge DMPs as one tool for reproducibility, yet they often highlight limitations in the information covered by DMPs, particularly regarding data sharing practices, and do not address the collective planning of reproducibility activities throughout the project lifecycle.

The literature reveals a critical gap: while various tools support specific aspects of reproducibility (pre-registrations, electronic notebooks, research object management), no comprehensive service exists for planning and managing reproducibility activities across the full research lifecycle. This gap is particularly problematic because reproducibility is often treated as an aftermath exercise through reproducibility studies rather than as a proactive principle for getting started, organising and connecting research activities, people, tools, and information. Within this Pilot, we developed both a conceptual framework (RMP practice) and technical implementation (ARGOS tool) to address these gaps at multiple levels. The RMP practice provides the content model, i.e. a taxonomy of questions organised into thematic families covering the research lifecycle. ARGOS provides the technical infrastructure, offering configurable templates, persistent identifier (PID) integration, and machine-actionable exports that make RMPs practical, shareable, and monitorable (Adamidi et al., 2025). Our stakeholder selection reflected the interconnected nature of research. Researchers need practical tools from project inception. Funders require mechanisms to monitor compliance with their policies. Research institutions need standardised approaches to support their communities. Reproducibility communities practice reproducibility at different settings and occasions. By engaging all these groups, we ensured RMPs address both reproducibility needs and established processes across the research ecosystem. We chose this dual approach, i.e practice and tool, because reproducibility planning requires both conceptual

[1] https://www.software.ac.uk/guide/writing-and-using-software-management-plan
[2] https://elixir-europe.org/sites/default/files/documents/software-management-plan.pdf

clarity and practical implementation. By extending the familiar DMP practice rather than creating something entirely new, we leveraged existing knowledge while addressing broader reproducibility needs. The machine-actionable nature of our technical efforts enables other research information systems to consume and/or enhance RMP data, supporting automated checks and actionable guidance throughout the whole process. Our Pilot addresses a key research question**:** "How can reproducibility be systematically planned and managed across the research lifecycle in a way that is both comprehensive and practical?"

To assist our investigations, we broke down our research question as follows:

- RQ1: What reproducibility elements and activities span the research output management lifecycle that should be captured at the planning stage?
- RQ2: What questions should an RMP ask to address reproducibility needs and how do these differ across epistemic contexts?
- RQ3: How can RMPs be made machine-actionable and interoperable with existing research infrastructure and to what extent can standards encode reproducibility-relevant elements?
- RQ4: How can DMP platforms be upgraded to ensure reproducibility is embedded and followed in publicly funded project?

We explicitly considered confounding factors including prior DMP familiarity, institutional context, disciplinary norms, technical expertise, and resource availability. Barrier and enabler evaluation addressed capability (knowledge gaps, technical skills), opportunity (time, institutional support, tools), motivation (perceived value, career incentives, requirements), technical challenges (usability, reliability), and policy factors (mandates, guidelines).

3.2. Methodology

Our stakeholder engagement used three complementary recruitment strategies. First, we leveraged existing TIER2 stakeholder networks established in WP2, including the collaboration with sister projects and reproducibility networks (RNs), ensuring continuity with earlier project activities. Second, we engaged directly with participants at relevant conferences and TIER2-organised workshops capturing diverse perspectives. Third, we utilized monthly ARGOS community calls to delve into discussions on reproducibility planning. Our inclusion criteria prioritized active involvement in research (data) management planning, execution, or administration, combined with experience in data management, reproducibility or related practices. We deliberately sought diversity across career stages from PhD students to senior researchers, institutional types including universities and funding agencies, and geographic locations primarily within Europe but extending to international participants. This approach yielded 89 unique external participants engaged across various activities, representing 12 European countries plus three non-European nations, with balanced representation of researchers (58%), funders (19%), data stewards and librarians (15%), and others (8%).The TU Graz granted ethical approval on June 7, 2023, according to the standard consent form and processes that they established as the research does not involve medical subjects. The study posed minimal risk, involving only surveys, workshops, and feedback sessions. All participants received assurance of data confidentiality and were informed of withdrawal rights as well as recognition to our

deliverables. We employed multiple co-creation tools to ensure outputs reflected genuine community needs. Three focus groups engaged scientists, research investigators, and reproducibility professionals to identify elements visible in traditional DMPs, practices performed during research, and questions needed at the planning stage. Two policy workshops with funders and administrators collected perceptions of RMPs as policy-supporting tools and explored adoption barriers and enablers.

Monthly ARGOS community calls provided regular touchpoints for usability feedback, averaging 15 participants per session. Bilateral consultations with TIER2 consortium experts ensured alignment with state-of-the-art reproducibility research and facilitated integration with complementary pilot activities, particularly the Decision Aid (Pilot 1) and Monitoring Dashboard (Pilot 6).

Our development methodology comprised two parallel workstreams addressing conceptual and technical dimensions. The theoretical workstream began with literature synthesis analyzing existing practices like DMPs, Software Management Plans, and domain protocols to identify reproducibility elements while aligning with TIER2's WP3 conceptual framework. In our co-creation activities, we focused primarily on upgrading the Science Europe (SE) DMP template with questions' modifications and extensions. Input involved both generic and discipline-specific information, however it was not enough to be able to provide a domain protocol.

The technical workstream focused on ARGOS implementation and the adoption of the RDA DMP Common Standard. We extended the maDMP to accommodate reproducibility elements, defining an RMP profile with new entities and properties mapped to PID ecosystems. Technical efforts involved, among other things, configuring the FAIRsharing API for automated content enrichment, implementing qualified references linking datasets, publications, software, methods, workflows, and contributors, and creating export pipelines producing RMP extensions. Platform integration work onboarded RMP templates to ARGOS, implemented user interfaces, and established connectors to OpenAIRE Graph, Validator, and MONITOR.

We conducted structured data collection through multiple mechanisms. User requirements assessment gathered feedback on SE DMP template effectiveness and completeness. Usability testing surveys captured user experience through questions about ease of use, navigation clarity, and perceived value. Peer review surveys in collaboration with OSTrails, collected expert feedback on technical specifications and DMP Common Standard extensions (Manola et al., 2025).

Workshop activities included active participation in conference sessions, and three co-creation workshops for RMP content development combining presentations, collaborative activities using tools like Miro boards, and structured feedback collection.

We employed qualitative analysis to encode open-ended responses and focus group transcripts, identifying recurring themes through both deductive (theory-driven) and inductive (data-driven) approaches. Content analysis organized workshop outputs into actionable categories, while narrative synthesis integrated findings across different data sources.

Our evaluation examined five key dimensions. First, conceptual validity assessed whether the RMP concept resonated with stakeholders and whether questions addressed genuine

reproducibility needs. Second, practical utility examined whether researchers could complete RMPs with reasonable cognitive load. Third, technical feasibility tested whether ARGOS could support RMP workflows and produce truly machine-actionable exports. Fourth, adoption potential explored whether funders would require RMPs and institutions would support their use. Fifth, interoperability validated whether proposed extensions integrated well with the DMP Common Standard.

Formative evaluation during development employed several approaches. Iterative feedback loops presented drafts at workshops and collected immediate reactions using collaborative annotation tools. Summative evaluation post-implementation measured outcomes through user satisfaction surveys distributed after full RMP completion. Evaluation occurred at monthly intervals through community calls, with formal interviews with CHIST-ERA supporting refinement of templates and UI/UX navigation.

Bi-monthly pilots' meetings discussed integration points and potential integration timelines with Pilots 1 and 6. The Decision Aid tool (Pilot 1) was intended to provide contextual guidance on reproducibility relevance and feasibility for different research types, which we intended to embed within RMP templates (however, due to lack of resources, the Aarhus team could not finish the development of this tool). Similarly, we aimed at enabling monitoring dashboards (Pilot 6) to consume RMP data to support indicators that track reproducibility across projects and programs. Though we were able to provide RMP data to the monitoring dashboard, integration with Pilot 1 was not feasible at the end due to lack of resources to provide an API for the Decision Aid Pilot tool to be consumed by ARGOS.

This is an early-stage evidence base and not intended to yield "best" or broadly generalizable results. The sample is small and skewed toward motivated early adopters; timing benefited from a favourable policy climate; the institutional context reflects European open-science maturity; projects had access to TIER2 support; and funder Pilots likely boosted uptake. These factors constrain external validity. Despite these limits, the results establish a solid baseline and a credible path to shift reproducibility from an afterthought to a proactive, planned practice embedded in everyday research workflows. CHIST-ERA's move to a Single Plan uniting DMPs and SMPs underscores this shift and delivers the first real-world demonstration of RMPs adopted by funders via the ARGOS platform.

3.3. Results

This section presents our findings organized by research question, followed by key performance indicators. We provide both quantitative metrics and qualitative insights from our evaluation activities.

Research Questions addressed

RQ1: Reproducibility Elements Across the Research Lifecycle

Through engagement with 89 stakeholders, we identified reproducibility elements spanning the research lifecycle that warrant capture at the planning stage. The elements we identified achieved high consensus, with core questions rated relevant (≥3/5) by over 80% of participants across all workshops. These universally applicable elements include research design documentation,

materials and methods specification, data and evidence description, analysis procedures, and verification approaches.

We successfully extended the Science Europe DMP template across seven major reproducibility domains: project-level reproducibility objectives and strategies; detailed specifications for data collection, processing, and analysis methods; software and computational environment documentation; materials and protocols specification; quality control and validation procedures; sharing and preservation plans for all research outputs; and verification and testing approaches. While participants contributed both generic and discipline-specific information, this input proved insufficient to generate complete domain-specific protocols. Nevertheless, the framework demonstrated strong applicability across epistemic contexts by focusing on functions rather than prescribing specific methods.

RQ2: RMP Questions and Epistemic Context Adaptability

Our co-creation process resulted in a comprehensive question set addressing reproducibility across different epistemic contexts. The questions address seven core dimensions: reproducibility objectives for the research; data collection methods; analysis tools and methods; computational environment documentation; quality control measures; output sharing approaches; and verification procedures. Through monthly ARGOS community calls averaging 15 participants per session, we gathered iterative feedback that validated question effectiveness and completeness across disciplines.

Cross-epistemic challenges emerged clearly through this validation process. Terminology differences proved the primary obstacle—"data," "reproducibility," and "methods" mean different things across disciplines. Additional challenges included tensions between detailed procedural specifications and interpretive frameworks, varying definitions of verification across domains, and different sharing constraints based on disciplinary norms and data sensitivity. We addressed these challenges through contextual glossaries, flexible question interpretation, balanced required and optional sections, and a focus on sharing plans rather than mandated sharing. The configuration model, which allows funders and institutions to version templates and add contextual guidance, proved effective for organic adaptation to specific epistemic contexts.

RQ3: Machine-Actionability and Interoperability

We successfully demonstrated that RMPs can be made machine-actionable and interoperable with existing research infrastructure through extension of the RDA DMP Common Standard. Technical implementation achievements include:

- Extension of the maDMP to accommodate reproducibility elements, defining an RMP profile with new entities and properties mapped to PID ecosystems;
- Configuration of the FAIRsharing API for automated content enrichment; Implementation of qualified references linking datasets, publications, software, methods, workflows, and contributors through persistent identifiers;
- Creation of export pipelines producing machine-actionable RMP extensions; and
- Establishment of connectors to OpenAIRE Graph, Validator, and MONITOR.

The ARGOS platform successfully supported RMP workflows including template customization, progressive disclosure interfaces, automated content enrichment, and export generation in multiple formats. We extended the machine-actionable DMP (maDMP) standard with an RMP profile featuring new entities and properties mapped to PID ecosystems, and peer review surveys conducted in collaboration with OSTrails confirmed these extensions integrated well with the DMP Common Standard.

Integration validation demonstrated practical interoperability across the research infrastructure ecosystem. We successfully established connectors to OpenAIRE Graph, Validator, and MONITOR; configured the FAIRsharing API for automated content enrichment; and implemented qualified references linking datasets, publications, software, methods, workflows, and contributors through persistent identifiers. Funders who piloted the system reported substantial practical value, with an estimated 40% reduction in time spent aggregating reproducibility information compared to ad hoc methods. This demonstrates that machine-actionable monitoring at scale is both technically feasible and practically valuable.

RQ4: Upgrading DMP Platforms for Embedded Reproducibility

CHIST-ERA's adoption of RMPs via the ARGOS platform provides the first real-world demonstration of funders requiring reproducibility management plans. Nine completed CHIST-ERA project RMPs demonstrate the feasibility of embedded reproducibility planning in funded projects. CHIST-ERA's strategic move to a Single Plan uniting DMPs and SMPs delivers concrete evidence that DMP platforms can be successfully upgraded to support comprehensive reproducibility planning.

Platform upgrades implemented include Onboarding of RMP templates to ARGOS with customizable configurations; Implementation of user interfaces supporting progressive disclosure and guided workflows; Integration with complementary systems (OpenAIRE Graph, Validator, MONITOR); Support for automated content enrichment through APIs; and Generation of machine-actionable exports for downstream consumption.

Template customization enables funders to adapt to specific requirements while maintaining interoperability. Structured formats facilitate consistent review, and integration minimizes administrative overhead. Nine completed CHIST-ERA project RMPs demonstrate feasibility of embedded reproducibility planning in funded projects. Implementation challenges identified include initial setup requiring technical capacity; assessment rubric development needs; staff training requirements; and balancing thoroughness with researcher burden. However, the successful Pilot demonstrates these challenges are surmountable with appropriate support.

Key performance indicator results

KPI 1 - Innovation: We successfully delivered a novel RMP concept not previously available, confirmed through literature review, formal definition in preparation for peer-reviewed publication, and recognition by the Advisory Board members.

KPI 2 - Inclusivity: We achieved external stakeholder engagement with 89 unique participants across 12 European and three non-European countries. Stakeholder composition balanced

researchers (58%), funders (19%), data stewards/librarians (15%), and others (8%), using multiple engagement methods. This broad input enhanced relevance and legitimacy while building a community of practice supporting sustained adoption.

KPI 3 - Reproducibility completeness: Although the current guidance is not yet sufficient to address all domains contexts, reflected in the variability of quality, the results indicate a continued need for guidance improvement while also demonstrating that structured planning can successfully capture the necessary information.

KPI 4 - Validity: Eleven out of 17 funders reported reproducibility as an important indicator for DMP evaluation; however, the process is encouraged for adoption rather than mandated as in the case of data management.

KPI 5 - Adoption rate: We met our target with 9 CHIST-ERA completed project RMPs.

Furthermore, the RMP Pilot highlighted the following opportunities and challenges:

Table 3.3.1: *Table presenting the opportunities and challenges which emerged during the RMP Pilot.*

Opportunities	Challenges
Stronger funder interest than anticipated: 11 out of 17 funders reported reproducibility as important despite no initial mandate expectations; Quality community-contributed content exceeding expectations from co-creation activities; Successful technical integration with multiple systems demonstrating robust interoperability.	Greater quality variability in completed RMPs than expected, indicating need for enhanced guidance; Stronger timing sensitivity within project lifecycle—retrospective RMP creation for ongoing projects proved significantly more difficult than prospective planning; Critical importance of institutional support infrastructure—absence of dedicated personnel created major barriers.

Evaluation Results

Conceptual validity: The RMP concept resonated strongly with stakeholders, confirmed through Advisory Board recognition and sustained engagement across 89 participants. Questions addressed genuine reproducibility needs, validated through focus groups and iterative refinement.

Practical utility: Researchers could complete RMPs with reasonable cognitive load when provided with appropriate support structures. However, time investment, initial learning curves, and technical knowledge requirements remain challenges. Quality variability in completed RMPs indicates need for continued guidance improvement.

Technical feasibility: ARGOS successfully supported RMP workflows and produced machine-actionable exports, achieving 99.2% uptime. Successful integration with OpenAIRE Graph, Validator, and MONITOR demonstrated technical viability.

Adoption potential: Nine completed CHIST-ERA RMPs and eleven out of 17 funders reporting reproducibility as important evaluation indicator demonstrate promising adoption potential. Current encouragement rather than mandates suggests gradual adoption pathway.

Interoperability: Proposed extensions integrated successfully with DMP Common Standard, validated through peer review surveys and successful technical integration with multiple systems.

Importantly, our study limitations constrain generalizability. The modest sample of 39 completed RMPs limits statistical power and edge case assessment. Self-selection bias means participants were likely more motivated than average, potentially yielding optimistic metrics. Geographic concentration in Europe may not represent global needs.

Regarding research planning and preregistration, we demonstrate reproducibility planning as distinct from but compatible with preregistration, show researchers value structured planning prompts, provide evidence that planning improves practice implementation, and address broader scope beyond research design. For open science and transparency, we demonstrate machine-actionable metadata enables monitoring, show planning tools can promote open and best practices, and provide evidence that structured planning increases adoption.

3.4. Discussion

The RMP Pilot successfully transitioned reproducibility planning from concept to functional reality. We established RMPs as a novel approach extending Data Management Plans to comprehensively address reproducibility across the research lifecycle. Through extensive co-creation with 89 external stakeholders, we developed new questions and guidance based on the SE template, and deployed ARGOS as a machine-actionable platform achieving 99.2% uptime.
Broader policy implications include recognition that reproducibility planning requires more resources than DMPs, and acknowledgment that cultural change toward routine reproducibility planning will take time and sustained effort. The Pilot achieved all five key performance indicators, with particularly strong results for innovation and inclusivity, and promising though variable results for reproducibility completeness. Adoption metrics showed promising early uptake.
The Pilot fundamentally reframes reproducibility from verification after completion to planning from inception. By breaking abstract concepts into concrete, manageable questions, RMPs make reproducibility accessible to researchers without specialized training. Machine-actionability enables systematic monitoring rather than anecdotal assessment, allowing funders to track patterns and measure policy effectiveness quantitatively. Integration with existing DMP practices reduces adoption barriers by leveraging familiar workflows rather than creating standalone requirements.

For reproducibility in research practice and policy, the Pilot has cultural implications beyond technical functionality. Explicit planning signals that reproducibility is valued and expected, normalizing documentation as part of research practice. Transparency increases as plans become

public artifacts, and accountability improves through auditable commitments. This shifts research culture from "can we reproduce this?" to "how will we enable reproduction?"
Broader policy implications include recognition that reproducibility planning requires more resources than DMPs, and acknowledgment that cultural change toward routine reproducibility planning will take time and sustained effort.

The Pilot addresses several literature gaps: planning-stage focus versus assessment emphasis, cross-domain framework with domain adaptations versus domain-specific guidance, implementation evidence versus prescriptive theory, machine-actionable structure versus unstructured text, and systemic enablement versus individual practices focus.

- For reproducibility frameworks, we operationalize theoretical concepts from Leonelli (2018), Plesser (2018), and Nosek et al. (2015) into concrete answerable questions, provide evidence of feasibly documentable reproducibility information, and show reproducibility can be planned prospectively rather than only assessed retrospectively. We align with context-dependent reproducibility recognition while adding practical implementation layers to theoretical frameworks.

- Regarding research planning and preregistration, we demonstrate reproducibility planning as distinct from but compatible with preregistration, show researchers value structured planning prompts, provide evidence that planning improves practice implementation, and address broader scope beyond research design. For open science and transparency, we demonstrate machine-actionable metadata enables monitoring, show planning tools can promote open and best practices, and provide evidence that structured planning increases adoption.

- Extending FAIR principles, we demonstrate applicability to research processes beyond data, show researchers can create FAIR metadata about reproducibility practices, provide infrastructure for findable, accessible, interoperable, reusable reproducibility information, and evidence FAIR RMPs enable downstream uses. We align with FAIR while extending principles to different research object type (Wilkinson et al., 2016).

Our policy recommendations emphasize the need for funders to recognize RMPs as legitimate planning deliverables, establish clear expectations while allowing flexibility for epistemic diversity, coordinate requirements across funding programs to reduce researcher burden, and invest in interconnected support that long term is cost effective for both training and infrastructure.

Broader policy implications include recognition that reproducibility planning requires more resources than DMPs, and acknowledgment that cultural change toward routine reproducibility planning will take time and sustained effort.

Key recommendations vary by stakeholder group:

- **Researchers** are encouraged to initiate reproducibility planning at the proposal stage, using RMPs to document design, data, and methods systematically. Structured templates

and integrated guidance support reflection and consistency across the research lifecycle. Targeted training and support are needed to address knowledge gaps and technical challenges, particularly for early-career researchers who show the highest engagement.

- **Funders** should adopt and customise RMP templates to align with policy requirements, enabling automated monitoring through machine-actionable exports. Organisational readiness, including staff training and clear assessment criteria, is essential for effective implementation. RMPs should be recognised as formal deliverables, with coordinated policies that reduce administrative burden and strengthen open science compliance.
- **Institutions** can embed RMPs within existing Open Science frameworks to provide structure, shared language, and consistent support for reproducibility. Dedicated personnel, adequate resources, and integration with institutional systems are key enablers. Training and coordinated implementation foster cultural change, making reproducibility a routine research practice.
- **Service providers** should ensure interoperability through open APIs, standard formats, and clear documentation. Integrating RMPs with repositories, CRIS systems, and monitoring dashboards enhances data connectivity and workflow automation. These capabilities enable reproducibility information to circulate efficiently across research infrastructures.

The Pilot established clear pathways for continued work. ARGOS service commits to maintaining RMP functionality beyond TIER2, with OpenAIRE infrastructure ensuring technical sustainability. Monthly community calls continue as engagement forum, while RDA DMP Common Standard WG provides standards governance.

Future research directions include longitudinal studies of RMP use and impact, comparative effectiveness research on valuable elements, disciplinary deep dives, policy impact studies, meta-research using RMPs as data sources, integration studies, and cultural studies of norm influence.

3.5. References

Adamidi, E., Vergoulis, T., Momeni, F., & Papadopoulou, E. (2025, November 13). TIER2 D5.1 Tools and practices for researchers. https://doi.org/10.17605/OSF.IO/5NQH6

Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J.W., Wallach, H.M., Daumé, H., & Crawford, K. (2018). Datasheets for datasets. *Communications of the ACM, 64,* 86 - 92. https://arxiv.org/abs/1803.09010

Grossmann, Y. V., Lanza, G., Biernacka, K., Hasler, T., & Helbig, K. (2024). Software Management Plans – Current Concepts, Tools, and Application. *Data Science Journal, 23*(1), 43. https://doi.org/10.5334/dsj-2024-043

Leonelli, S. (2018). *Rethinking Reproducibility as a Criterion for Research Quality*. In L. Fiorito, S. Scheall, & C. E. Suprinyak (Eds.), *Research in the History of Economic Thought and Methodology (Vol. 36*, pp. 129–146). Emerald Publishing Limited. https://doi.org/10.1108/S0743-41542018000036B009

Manola, N., Papadopoulou, E., Kontopidi, M., Grypari, I., Spichtinger, D., Kakaletris, G., & Tziotzios, D. (2025). D4.2 Horizon Europe Report. Zenodo. https://doi.org/10.5281/zenodo.16753141

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

Plesser H. E. (2018). Reproducibility vs. Replicability: A Brief History of a Confused Terminology. *Frontiers in neuroinformatics, 11,* 76. https://doi.org/10.3389/fninf.2017.00076

Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J. W., da Silva Santos, L. B., Bourne, P. E., Bouwman, J., Brookes, A. J., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, C. T., Finkers, R., Gonzalez-Beltran, A., … Mons, B. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data, 3,* 160018. https://doi.org/10.1038/sdata.2016.18

4. Pilot 3 - Reproducible Workflows

Authors: Eleni Adamidi, Thanasis Vergoulis

4.1. Introduction

Pilot 3 customizes and evaluates tools and practices that enable reproducible computational workflows across two epistemic contexts, life sciences and computer sciences, by adapting and extending the open-source SCHeMa platform (Vergoulis et al., 2021) with containerization (IBM, 2024), workflow description languages (CWL, 2024), and experiment packaging specifications.

Faced with the complexity of analysis pipelines, the large number of computational tools, and the enormous amount of data to manage, there is compelling evidence that the reproducibility of computational workflows is of paramount importance (Cohen-Boulakia et al., 2017; Di Tommaso et al., 2017).

Without advanced workflow systems, scripts that work on a single computer are often not scalable to larger or cloud-based systems without significant modification. Workflow systems like Galaxy and CWL provide scalable solutions that maintain the integrity and reproducibility of workflows across different computational environments (Perkel, 2019). Specifically, regarding reproducibility in the computational research, the absence of systematic methods for managing data manipulation and version control can lead to non-reproducible outcomes. Automated and documented workflows help avoid these pitfalls by ensuring that all data manipulations are traceable and reproducible (Sandve et al., 2013). Moreover, detailed documentation and version control are critical for reproducibility, especially in computational research where outputs are highly dependent on specific software versions and configurations. Systems that track changes and manage versions of scripts and software settings help in maintaining the integrity of research outcomes (Sandve et al., 2013).

The Pilot targets two stakeholders, researchers in life sciences and in computer sciences where pipeline complexity and data scale collide with practical reuse needs. It extends SCHeMa (Vergoulis et al., 2021) by creating a new virtual lab tool called SCHEMA api in the back end and SCHEMA lab in the front end, that supports (i) containerized task execution, (ii) workflow execution (iii) computational experiment creation and (iv) packaging experiments via RO-Crates and validates these through stakeholder-driven iterations (questionnaires, webinars, GitHub ticketing).

Pilot 3 aimed to answer the following research question: to what degree are best practices for computational reproducibility supported and promoted in the new SCHEMA lab tool? The new SCHEMA lab tool demonstrates a strong alignment with best practices for computational reproducibility by integrating containerized execution, experiment creation and metadata packaging. The use of RO-Crates will enhance transparency and reusability by documenting inputs, parameters, software versions, and outputs according to FAIR principles. Co-creation activities such as questionnaires, webinars, and GitHub feedback further promote awareness and community engagement around reproducibility. Overall, SCHEMA lab and api supports and actively promotes reproducible research practices by combining technical rigor with user-centered

design, offering a practical and extensible solution for managing computational experiments across disciplines.

4.2. Methodology

Participant selection

Two primary stakeholder communities are included: life scientists (ELIXIR community) and computer scientists (Fleming).

Ethical approval

Ethical approval has been secured for the whole project for co-creation processes through TU Graz.

Research design

- Round 1 questionnaire to elicit requirements in life and computer sciences. Webinar 1 (Nov 2024): demonstrate 1st TIER2 SCHEMA lab and SCHEMA api deployment and collect feedback (Results report can be found here Enhancing_Reproducibility_in_Research_Round1_questionnaire_Analysis_Report)
- GitHub ticketing: ongoing feature requests, bugs, prioritization and monitoring.
- Round 2 questionnaire (Sept 2025) and Webinar 2 (Nov 2025): present enhancements and receive feedback.

Measures

The evaluation of Pilot 3 combined both quantitative and qualitative methods to assess the functionality, usability, and overall adoption of the SCHEMA lab and SCHEMA api prototype.

- Quantitative:
  The quantitative evaluation focused on measurable indicators of system use and engagement. These included the number and frequency of task and workflow executions performed through the platform, the number of active users during the pilot period, and the level of community interaction as reflected in GitHub activity, such as the number of issues reported, feature requests submitted, and pull requests merged.
- Qualitative:
  The qualitative assessment captured user perspectives and experiential feedback. Data were collected through webinars, where participants discussed usability aspects, potential barriers to adoption, and desired new features.

Analysis

Synthesis of webinar feedback and GitHub narratives to inform iterations.

Evaluation plan

- **KPI 1** Reproducibility: The number of successful reproductions of computational experiments.
- **KPI 2** Adoption Rate Across Domains: Rate of adoption measured by the number of active users, workflow executions, categorized by different domains (life sciences and computer sciences).
- **KPI 3** User Satisfaction Scores: Average satisfaction scores obtained from user assessment surveys at different stages of the Pilot.
- **KPI 4** GitHub Interaction Metrics: GitHub activity metrics, including the number of issues raised, feature requests, and contributions from the user community.

Our evaluation method for Pilot 3 included measuring the execution of computational tasks and workflows in SCHEMA lab as well as the creation of computational experiments. Moreover, we are collecting feedback from the research community through our questionnaire rounds and webinars at different stages of the Pilot.

  - Dec 2024: 1st deployment milestone (run tasks; create experiments).
  - Mar–Jul 2025: development based on Round 1; Round 2 survey & webinar.
  - Sep–Oct 2025: integrative analysis and iterative refinements; documentation.
  - Nov–Dec 2025: assessment and dissemination.

Synergy

We have been in synergy with Pilot 4 to explore the execution of a computational workflow that comes from the social sciences. The Reproducibility Checklists developed by GESIS (Pilot 4) could be added in the future directly in SCHEMA lab to promote the adoption of such reproducibility tools and practices.

4.3. Results

Pilot 3 resulted in the development and public release of a prototype version of SCHEMA lab, fully integrated with the SCHEMA api backend. These components establish an open-source framework that allows researchers to design, execute, and monitor computational experiments in a transparent and reproducible way. The first TIER2 SCHEMA tool deployment was released in December 2024.

The underlying architecture connects the SCHEMA lab front-end with the SCHEMA api, which interfaces with the Task Execution Engine (TESK), the Kubernetes orchestration layer, and an S3-based storage (SCHEMA DB) as shown in Figure 4.3.1. This modular setup ensures scalability for managing containerized computations across diverse environments.

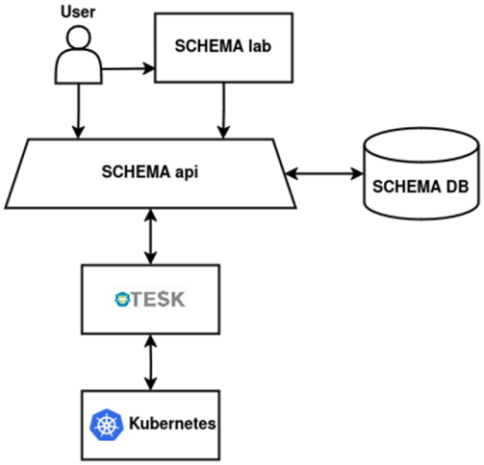

Figure 4.3.1. *SCHEMA lab and api architecture*.

Key functionalities introduced through this work include:

- Execution of containerized applications either as single tasks or multi-step workflows, providing portability of computational runs.
- Grouping of runs into computational experiments with automatic capture of metadata, parameters, and provenance, enhancing traceability and reproducibility.
- Real-time monitoring of executions, enabling users to follow task or workflow progress directly through an integrated dashboard.
- Export of completed experiments as RO-Crates (packaging research data with their metadata), ensuring standardized packaging for sharing and reuse of results.
- User-friendly management interface, offering a dashboard-based environment to create, edit, and review experiments and their associated computational runs.

The SCHEMA lab dashboard is shown in Figure 4.3.2, where we can see an overview of all project tasks, including workflow and single-task entries, their unique IDs, current status (e.g., scheduled or running), submission timestamps, last-update information, and available actions for managing each computational workflow such as the “re-run” functionality to re-execute the same task and workflow.

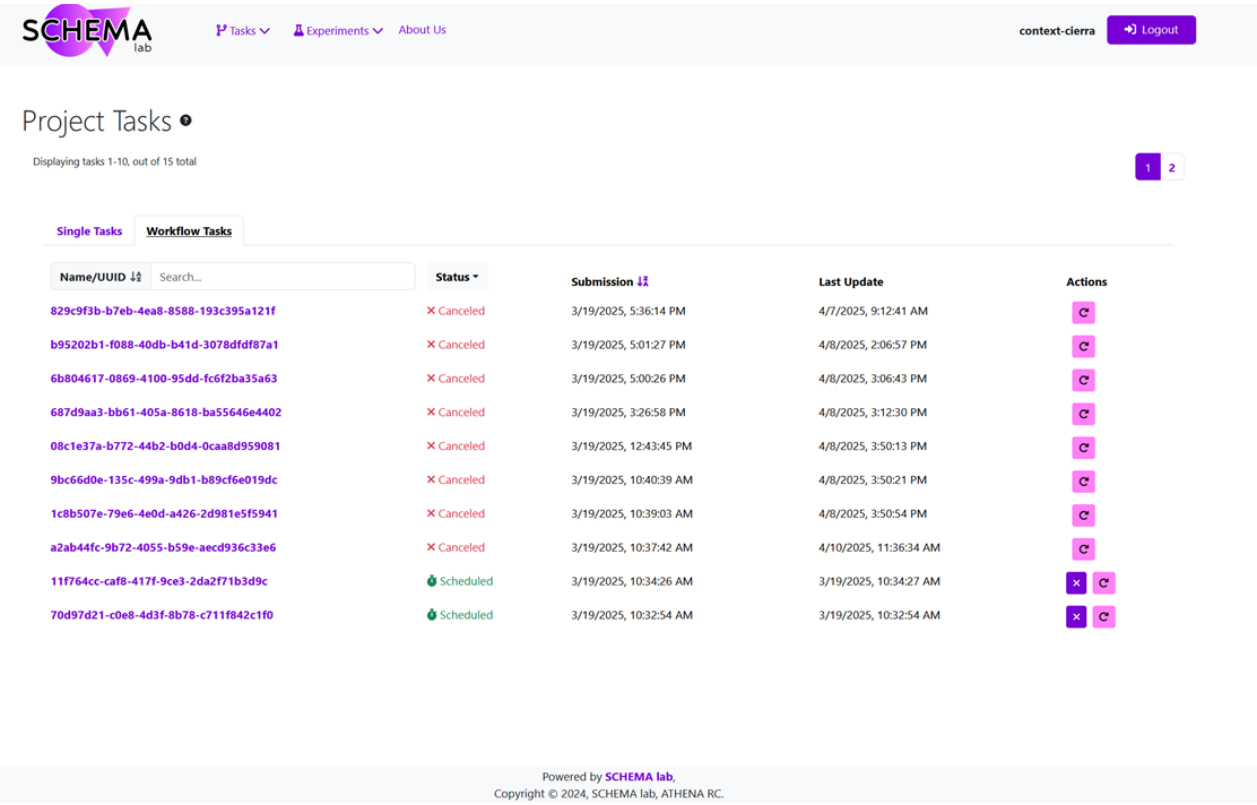

Figure 4.3.2. *The SCHEMA lab dashboard displaying a list of project tasks, including workflows.*

The prototype and related resources are publicly available through the following channels:

- SCHEMA lab homepage: https://schema-lab.hypatia-comp.athenarc.gr/
- GitHub repositories: SCHEMA API: https://github.com/athenarc/schema
- SCHEMA Lab: https://github.com/athenarc/schema-lab
- ACM SSDBM 2025 paper: https://dl.acm.org/doi/10.1145/3733723.3733743
- Technical documentation: SCHEMA API Swagger can be found here: https://api.hypatia-comp.athenarc.gr/
- User documentation: https://schema.athenarc.gr/

Efficacy and effectiveness

- Efficacy: The Pilot demonstrated the system's ability to execute containerized computational tasks and multi-step workflows, as well as to create structured experiments by combining these runs. Each experiment could capture and store relevant metadata, supporting transparency and reproducibility in computational research.
- Effectiveness: In practice, the platform showed early adoption in the life sciences domain, with researchers successfully applying it to real use cases. The packaging and management of computational tasks and workflows within the SCHEMA lab environment was feasible and beneficial for users seeking to organize and document their analyses more effectively.

Co-creation activities

To ensure that the platform reflected actual user needs, researchers from the Life Sciences Research ELIXIR communities were actively engaged in the co-creation process. Through questionnaires, online consultations, and interactive webinars, participants provided valuable input on usability aspects, metadata requirements, and the most common workflow use cases encountered in their research. This user-driven approach influenced interface design choices, metadata schema definition, and prioritization of features such as experiment grouping and monitoring dashboards. Reports on the questionnaire results can be found in the following link (video and analysis report):
Enhancing_Reproducibility_in_Research_Round1_questionnaire_Analysis_Report.

The evaluation results for Pilot 3 are presented below, structured according to the KPIs and qualitative measures defined in the methodology.

KPI 1- Reproducibility
Reproducibility was assessed through the number of successfully executed computational runs and the ability to repeat executions under the same conditions (through the re-run functionality of SCHEMA lab). Across the Pilot period, 181 computational tasks (single tasks and multi-step workflows) were successfully executed through SCHEMA lab. These included repeated executions of containerized workflows, demonstrating consistent outputs across runs.

KPI 2- Adoption Rate across domains
Adoption was evaluated by tracking active users and usage across the two epistemic communities. During the evaluation period, 17 active users engaged with SCHEMA lab, including both life scientists (ELIXIR communities) and computer scientists (Fleming). Usage statistics show early but meaningful adoption, with users applying the tool to real use cases.

KPI 3 - User satisfaction
User satisfaction was evaluated through the Pilot 3 webinar session and stakeholder survey. Participants rated core SCHEMA lab features, such as task submission, task monitoring, and workflow execution, as highly important, with average scores between 4.0 and 4.2 out of 5 in the live webinar feedback. Survey respondents expressed strong support for RO-Crate export capabilities, with 60% rating this feature as very useful (4/5) and 30% as extremely useful (5/5). All detailed feedback reports are available via the linked webinar here and survey analysis report here.

KPI 4 - GitHub interaction metrics
Community interaction and external contributions were measured through activity in the SCHEMA GitHub repositories (SCHEMA api, SCHEMA lab and all relevant branches of those repositories). Across the pilot period, 12 issues were opened, 2 feature requests (direct push and retrieval of workflows from open repositories, comprehensive documentation), 5 forks, and 6 contributors interacted with the codebase.

Qualitative feedback from the survey and webinar identified key strengths and areas for improvement. Participants emphasized the importance of improved documentation and enhanced metadata support to link data, methods, and workflow execution steps.

4.4. Discussion

SCHEMA api (https://github.com/athenarc/schema-api/tree/main) has been developed to provide a service for submitting and monitoring containerized task execution requests programmatically. **SCHEMA lab** has also been developed as an open-source platform aiming to assist researchers and scientists in managing and executing computational tasks in this virtual lab front end environment.

Implications

Pilot 3 contributed to the broader reproducibility agenda by demonstrating how lightweight, open-source infrastructures can lower the barrier for researchers to adopt reproducible computational practices. Rather than proposing new standards, SCHEMA lab and SCHEMA api build upon existing, widely adopted technologies such as containerization, task execution services, and orchestration systems, to translate reproducibility principles into a usable and accessible environment.

The Pilot highlighted that reproducibility is not only a technical challenge but also an organizational one. Through the co-creation process, participants emphasized that transparency, documentation, and ease of reuse are only sustainable when supported by intuitive tools. In this sense, SCHEMA lab serves as a proof-of-concept that user-centric design can significantly enhance the adoption of reproducible methods.

From a stakeholder perspective, although researchers are the main stakeholder group there are several groups (including those) who can benefit:

- Researchers and data analysts can streamline their computational workflows and ensure transparent documentation of their analyses.
- Infrastructure providers can use SCHEMA components as a model for integrating container execution and monitoring within larger research clouds.
- Policy makers and funders can reference such frameworks as examples of practical enablers of FAIR and reproducible science.

Implementability

The modular design of SCHEMA api and SCHEMA lab allows for gradual integration into different research contexts. The architecture is compatible with Kubernetes-based infrastructures and can be deployed locally or within institutional clouds. This flexibility supports applicability across diverse disciplines, from bioinformatics to computational social science, where containerized tools are increasingly being used.

While the current prototype was tested primarily in life-science settings, the underlying approach can be generalized. For instance, social scientists or environmental researchers could adapt the

same interface to execute statistical models, simulations, or text-analysis pipelines. Further evaluation in such settings would help confirm this broader relevance.

The Pilot benefited significantly from an iterative, co-creative process. Early feedback from researchers and developers guided the prioritization of features such as real-time monitoring, dashboard clarity, and metadata capture. Regular communication through webinars and questionnaires ensured transparency.

At the same time, the evaluation process revealed important learning points. The small but diverse group of pilot participants provided valuable qualitative feedback but limited opportunities for systematic, quantitative evaluation. Extending the Pilot to a larger and more diverse user community would allow for more robust measurement of usability and adoption over time.

Participants also noted that the process itself was valuable in clarifying their own reproducibility practices. Engaging directly with tool development prompted reflection on documentation habits, dependency management, and the need for standardized reporting. In this sense, the Pilot functioned not only as a technical test but also as a learning process that strengthened participants' understanding of reproducible research.

The outcomes of Pilot 3 align closely with existing literature emphasizing reproducibility as a foundational principle of computational science. Works such as Sandve et al. (2013) and Cohen-Boulakia et al. (2017) underline that workflow execution with metadata tracking is essential to reproducible research. SCHEMA lab operationalizes these principles by integrating containerization (as advocated by Di Tommaso et al., 2017) and metadata packaging through RO-Crates (Sefton et al., 2020) into a unified, researcher-friendly interface.

Strengths & Limitations

The strengths of Pilot 3 include:

- Cross-disciplinary validation: The Pilot engaged both life and computer scientists, ensuring applicability across epistemic domains.
- Integrated reproducibility model: SCHEMA lab and SCHEMA api together enable computational reproducibility from containerized execution to metadata-rich packaging, within a single ecosystem.
- Co-creation methodology: Iterative development through questionnaires, webinars, and GitHub ticketing ensured responsiveness to user needs and transparent prioritization.

Limitations of this Pilot are:

- Sample size and scope: While adoption is growing, the current user base (≈ 17 active users) remains limited, constraining quantitative evaluation.
- Temporal constraints: The Pilot's duration allowed only one major iteration cycle; longer evaluation would capture sustained usability and adoption dynamics.
- Integration with external workflow languages: Although compatible by design, full interoperability with established workflow languages (e.g., Nextflow, CWL) remains a target for subsequent development.

We will continue to maintain and adapt the SCHEMA api and SCHEMA lab tool to support the execution of tasks and workflows as well as the creation of computational experiments. The findings from Pilot 3 highlight the critical importance of supporting reproducible computational practices across research domains through (i) the use of standardized and containerized execution environments, (ii) workflow-based structuring of computational experiments that captures all parameters, dependencies, and configurations, and (iii) automated provenance recording of inputs, outputs, and software metrics. Based on our results and engagement with stakeholders, we propose the following recommendations for researchers:

- Adopt virtual laboratory environments such as SCHEMA lab to execute and document computational experiments. These environments facilitate reproducibility through containerized task execution, experiment creation and metadata tracking.
- Integrate RO-Crates or similar metadata frameworks in the research process to ensure that data, code, and results remain FAIR (Findable, Accessible, Interoperable, and Reusable).
- Participate in open co-creation processes and provide structured feedback to continuously refine reproducibility tools.

Pilot 3 successfully demonstrated how SCHEMA api and SCHEMA lab can bridge the gap between theoretical reproducibility principles and their practical application in real research settings particularly within life science and computer science workflows. Through co-creation activities involving both life and computer scientists, we implemented and tested a platform that enables:

- Execution of containerized tasks and workflows in a scalable manner.
- Creation and packaging of computational experiments enriched with FAIR metadata.
- Iterative community-driven refinement through surveys, webinars, and GitHub contributions.

The Pilot's success highlights that computational reproducibility—meaning the ability to re-run the same analysis with the same code, parameters, data, and environment to obtain same results—is attainable when supported by open collaboration, shared standards, and clear documentation practices. However, sustained support and integration with other research infrastructures remain essential to ensure long-term impact.

Looking ahead, we will continue to maintain and expand the SCHEMA ecosystem, connecting with other TIER2 Pilots, integrating domain-specific workflows (e.g., from the social sciences), and supporting new features such as reproducibility checklists and metadata quality validation.

In summary, Pilot 3 contributes a concrete and extensible model for computational reproducibility, offering a foundation upon which future research infrastructures can build to strengthen transparency, trust, and efficiency in scientific research.

4.5. References

Cohen-Boulakia, S., et al. (2017). Scientific workflows for computational reproducibility in the life sciences: Status, challenges and opportunities. *Future Generation Computer Systems, 75*, 284–298. https://doi.org/10.1016/j.future.2017.01.012

Common Workflow Language. (2024, February 12). *Common Workflow Language (CWL).* https://www.commonwl.org/

Di Tommaso, P., Chatzou, M., Floden, E. W., Barja, P. P., Palumbo, E., & Notredame, C. (2017). Nextflow enables reproducible computational workflows. *Nature Biotechnology, 35*(4), 316–319. https://doi.org/10.1038/nbt.3820

IBM. (2024, February 12). *What is containerization?* IBM. https://www.ibm.com/topics/containerization

Perkel, J. M. (2019). Workflow systems turn raw data into scientific knowledge. *Nature, 573*(7772), 149–150. https://doi.org/10.1038/d41586-019-02619-z

Sandve, G. K., Nekrutenko, A., Taylor, J., & Hovig, E. (2013). Ten simple rules for reproducible computational research. *PLoS Computational Biology, 9*(10), e1003285. https://doi.org/10.1371/journal.pcbi.1003285

Sefton, P., et al. (2020, October). *RO-Crate metadata specification 1.1.* Zenodo. https://doi.org/10.5281/zenodo.4031327

Vergoulis, T., Zagganas, K., Kavouras, L., Reczko, M., Sartzetakis, S., & Dalamagas, T. (2021, March). *SCHeMa: Scheduling scientific containers on a cluster of heterogeneous machines.* arXiv. https://arxiv.org/abs/2103.13138

5. Pilot 4 - Reproducibility Checklists for Computational Social Science Research

Author: Fakhri Momeni

5.1. Introduction

Reproducibility in scientific publications has been a concern across disciplines, including computational social science (CSS). Studies show that even when publications claim reproducibility, many fail to achieve it in practice. For instance, out of 19 publications claiming reproducibility, only 13 were found to be mostly or fully reproducible when re-evaluated through replication of figures, numerical results, and conclusions (Archmiller et al., 2020). Similar problems have been observed in biomedical research, where insufficient documentation of experimental environments, code errors, and discrepancies in results are common (Samuel & Mietchen, 2023).

It is now widely recognized that merely sharing data and code is not sufficient for reproducibility (Clyburne-Sherin et al., 2019). Essential elements include environment specifications (e.g., requirements.txt or YAML files), containerization files, and adoption of open formats (Archmiller et al., 2020; Nosek et al., 2022; Hardwicke et al., 2022; Hardwicke et al., 2020; Bednar, 2023). These practices, however, face persistent challenges such as changing software dependencies, version incompatibilities, and evolving computational environments (Lazer et al., 2020). The scarcity of key indicators for replication and verification further complicates reproducibility efforts (Nosek et al., 2022).

Non-computational resources are also critical transparency and reproducibility indicators. These include sharing raw materials, access protocols, funding statements, and conflict-of-interest declarations (Nosek et al., 2022; Hardwicke et al., 2022). Such metadata, sometimes termed “meta-research data”, reflects how reproducibility-friendly a research policy is, for example, when open access publications also provide statements, materials, and scripts (Hardwicke et al., 2020). However, systemic barriers persist, including institutional constraints, ethical limitations, and inadequate infrastructure (Lazer et al., 2020).

The literature highlights the importance of preregistration for enhancing research credibility but warns of potential misuse and long-term unintended effects (Pham & Oh, 2021). There have been proposals for pre-publication replicability assessments by journals, and for stricter post-publication reproducibility checks (Altmejd et al., 2019). In assessing reproducibility, a binary categorization (reproducible or not) is insufficient. More nuanced tier systems distinguish between reproducibility, replicability, robustness, and generalizability, aiming to counter “open-washing” (Schoch et al., 2023). Schoch et al. (2023) further define three degrees of computational reproducibility:

- 1°CR – reproducible by the original scholar;
- 2°CR – reproducible by a trusted third party (e.g., journal editor);
- 3°CR – reproducible by the general public.

Other frameworks extend this to eight levels, from non-reproducible research to containerized, fully automated setups requiring minimal user effort (Bednar, 2023). Training early-career researchers in reproducibility and open science practices is essential, but insufficient on its own. Institutional-level efforts—such as harmonized data access protocols, interdisciplinary policies, and support for version control and sustainable code-sharing practices—can help lower barriers (Lazer et al., 2020; Kohrs et al., 2023).

Existing frameworks such as the Transparency and Openness Promotion (TOP) Guidelines (Nosek et al., 2016), the Open Science Framework (OSF), and DIME standards by the World Bank provide valuable foundations. However, these do not fully address the broader CSS-specific challenges, particularly when working with sensitive, dynamic, or platform-dependent datasets (Playford et al., 2016). Challenges include lack of dedicated tools, inconsistent data retention policies, absence of sustainable web resources, and insufficient structured frameworks for research object documentation (Playford et al., 2016).

Despite increased attention in related fields like psychology and political science, there remains a notable gap in practical, stage-specific frameworks tailored to computational social science. Specifically, existing literature does not provide:

1. Clear, actionable guidelines for achieving reproducibility at different stages of CSS research;
2. Estimates of the effort required to implement these practices;
3. Mechanisms to integrate reproducibility practices into daily workflows without creating undue burden, especially for early-career researchers (Ferguson et al., 2023).

The Pilot was initiated to provide structured, actionable support for enhancing reproducibility in computational social science. From the outset, one strand of work focused on developing a three-phase reproducibility checklist covering planning and data collection, analysis and processing, and sharing and archiving. To ensure community relevance, researchers were consulted through surveys to evaluate and prioritize proposed checklist items. This process provided valuable insights into which practices were most strongly supported across the social science community, thereby contributing to a broader understanding of reproducibility needs.

At the same time, the Pilot recognized that the Methods Hub portal hosts granular computational methods rather than complete end-to-end research projects. For this reason, a simple, lightweight checklist was developed as a practical tool to support reproducibility with minimal additional effort. The checklist was designed to be both accessible and efficient, allowing researchers to document and share their computational methods in a way that facilitates reuse, replication, and integration into existing workflows.

To ground these developments, a survey of current reproducibility practices and challenges was conducted among computational social scientists. The survey, which received 180 responses across roles from PhD students to full professors, confirmed that while reproducibility is widely valued, its implementation is hindered by inadequate tools, limited documentation strategies, and practical barriers. These findings reinforced the need for a lightweight, checklist-based solution that integrates smoothly into research workflows rather than adding extra burden.

Building on this foundation, the Pilot continued with a second strand of work: empirically assessing the impact of Methods Hub, enhanced with the simple checklist, on reproducibility outcomes. This evaluation focused on measurable key performance indicators (KPIs) such as reproducibility success rates, data and code sharing rates, and user experience. In this way, the Pilot not only developed reproducibility tools but also systematically tested their effectiveness in practice.

By combining a validated three-phase checklist for comprehensive research processes with a simplified checklist tailored to granular computational methods, the Pilot addresses multiple levels of reproducibility needs. It provides both strategic guidance for research workflows and practical tools for everyday use, filling a critical gap between high-level frameworks and concrete, discipline-specific practices.

The preregistration for this Pilot can be found here: OSF | Reproducibility checklists

5.2. Methodology

Participant selection

The Pilot involved participants in two phases of data collection:

1. Surveys
    a. Recruitment: Invitations were sent to approximately 26,000 social scientists identified through the Scopus database (2016–2023) as corresponding authors of publications containing CSS-related keywords.
    b. Sample:
        i. *Survey on practices and challenges*: 180 responses were collected. Participants represented a diverse range of roles, including PhD students, postdoctoral researchers, and faculty members. Senior academics with more than a decade of experience, such as full and assistant professors, formed the dominant group.
        ii. *Survey on checklist evaluation*: 64 responses were collected from a subset of this group. Demographics closely overlapped with the first survey, although anonymity prevented direct participant matching.
2. Experiment (Methods Hub vs External Repositories)
    a. Recruitment: Participants were recruited via three channels: (1) internal GESIS networks, (2) outreach to partner universities with existing collaborations, and (3) direct email invitations sent to 2,040 corresponding authors of CSS-related publications indexed in Scopus in recent years.
    b. Sample: To date, 37 participants have enrolled, with a target range of 30–40 participants. Each participant is assigned two tasks: reproducing one method from Methods Hub and one comparable method from an external repository (e.g., GitHub). This design ensures balanced testing while minimizing redundancy.

Together, these participant groups provided both broad community feedback (through the surveys) and hands-on evaluation data (through the experiment), offering complementary perspectives on reproducibility practices and the practical impact of checklist-supported methods.

Ethical approval

All data collection activities in this Pilot were reviewed for compliance with ethical and data protection standards.

- Surveys: The two surveys were assessed by the legal affairs and data protection office at GESIS – Leibniz Institute for the Social Sciences. No sensitive personal data were collected. Participation was voluntary, anonymous, and based on informed consent. Respondents were explicitly informed about the purpose of the study and the type of information collected.
- Experiment: The experiment comparing Methods Hub with external repositories followed the same ethical standards. Recruitment was conducted through institutional networks, social media (Bluesky and LinkedIn), university collaborations, and email outreach to corresponding authors indexed in Scopus. Participants joined voluntarily, provided informed consent, and were compensated for their time. No personally identifiable or sensitive data were collected beyond what was necessary for task completion and evaluation.

As a result, the Pilot adhered to institutional requirements for ethical and data protection compliance, ensuring transparency, anonymity, and voluntary participation across all stages of data collection.

Research design

The development of the reproducibility checklists was grounded in a co-creation approach that actively engaged the computational social science community. Researchers were consulted through two surveys to provide feedback on proposed checklist items and to share their perspectives on current practices, challenges, and needs.

The Pilot combined multiple methodological components:

1. Surveys
    a. *Survey on reproducibility practices and challenges*: This survey assessed current practices, attitudes, and obstacles related to reproducibility. It included multiple-choice and Likert-scale questions covering open science practices, documentation strategies, and barriers such as time, incentives, and access to tools. A total of 180 participants responded.
    b. *Survey on checklist item evaluation*: This follow-up survey gathered feedback on 59 proposed checklist items derived from the literature. Participants rated each item on a 9-point Likert scale to indicate its necessity for inclusion in the final checklists. Responses from 64 participants informed the prioritization of items into three checklists aligned with different stages of the research process.
2. Experiment (Methods Hub vs external repositories)
    a. A controlled user study was designed to evaluate the impact of the simple checklist integrated into Methods Hub. Participants attempted to reproduce computational methods from two sources: one from Methods Hub (with the checklist) and one from an external repository such as GitHub (without the checklist).

   b. Each participant was assigned two tasks to ensure balanced testing across platforms and methods.

3. Workshop
   As part of the Pilot, a half-day workshop titled *"Social Science Meets Web Data: Reproducible and Reusable Computational Approaches"* (ICWSM 2025) was organized. The workshop provided a venue to introduce the Methods Hub portal and its integrated reproducibility checklists to the broader research community. During a hands-on session, participants used MyBinder.org to run one sample method directly from Methods Hub, guided by the organizers. This demonstration showcased the portal's functionality and highlighted the potential of lightweight checklists to support reproducibility. While the workshop did not formally collect structured feedback, it served to disseminate project outcomes, raise awareness, and demonstrate practical use of checklist-supported reproducible methods in a live setting.

Analysis

For the analysis we used the data from two surveys:

- *Survey on practices and challenges*: Responses were analysed using descriptive statistics (percentages, frequency distributions) to identify common practices, barriers, and desired features for reproducibility support. Comparisons were made across career stages (e.g., early-career vs. senior academics) to highlight potential differences in awareness and adoption.
- *Survey on checklist items*: Ratings of 59 items on a 9-point Likert scale were grouped into three categories (inclusion, neutral, exclusion). Items were prioritized for the final checklist based on aggregated agreement rates.

Also used the experiment test to evaluate the effectiveness of the Pilot. The controlled user study compared the reproducibility of methods on Methods Hub (with checklist support) and on external repositories (without checklist support). The following analyses are planned:

- Reproducibility success rate: Categorical comparison (success vs. failure) between platforms using chi-square tests.
- Time to reproduce: Average task completion time compared between platforms using t-tests (or non-parametric alternatives if distributions are skewed).
- Ease of use ratings: Likert-scale ratings of usability analyzed with Mann-Whitney U tests for cross-platform differences.
- Accuracy of reproduced results: Quantitative comparison of output similarity, assessed through descriptive statistics and error margins relative to the original results.

These analyses were designed to test whether checklist-supported workflows in Methods Hub produce higher reproducibility rates, reduced time, and improved usability compared to methods hosted in external repositories.

Evaluation plan

The pilot evaluation aimed to examine:

- Whether checklist-supported workflows in Methods Hub improve reproducibility compared to external repositories.
- The effect of checklist integration on reproducibility success rate, time to reproduce, ease of use, and accuracy of results.
- Broader adoption-related factors such as barriers and enablers identified through survey responses (e.g., lack of time, training, incentives).

Potential confounding factors considered include:

- Participants' prior experience with reproducibility tools and coding environments.
- The complexity and documentation quality of the assigned methods.
- Variation in computational environments (hardware/software differences across participants).

Changes to evaluation plan

While the original pilot plan envisioned only survey-based evaluations, the scope was extended to include a hands-on experiment. This shift allowed the Pilot not only to explore perceptions and priorities but also to empirically measure the impact of checklist integration on reproducibility outcomes.

5.3. Results

The piloting process followed three sequential stages:

1. **Community feedback and co-creation** through two surveys, assessing reproducibility practices and validating checklist items.
2. **Demonstration and dissemination** via a workshop introducing the Methods Hub and illustrating checklist-supported workflows.
3. **Empirical evaluation** through a controlled experiment comparing Methods Hub with external repositories.

This multi-stage process ensured that the Pilot addressed both the *perceived needs* of the computational social science community and the *practical effects* of checklist-supported methods on reproducibility.

Co-creation was implemented primarily through the survey-based consultations:

- **Survey 1 – Reproducibility Practices and Challenges (n = 180):**
  Participants represented diverse roles, including PhD students, postdoctoral researchers, and faculty members. Documentation of experimental steps (81%) and code sharing (71%) were the most frequent practices, whereas standardized checklists (29%) and specialized guides (16%) were rarely used. Main barriers included lack of time (58%), limited

incentives (48%), and insufficient training (28%). Respondents emphasized the need for accessible best-practice examples (58%), clear standards (56%), and step-by-step guides (51%).

- **Survey 2 – Checklist Item Evaluation (n = 64):**
  Respondents rated 59 proposed checklist items on a 9-point Likert scale. Of these, 76% were rated essential, particularly those related to computational methods (96.9%), data sources (95.3%), and data description (95.3%). The feedback informed the final simplified checklist integrated into the Methods Hub portal.

The evaluation assessed the impact of Methods with checklist support compared to other methods from GitHub and Hugging Face on reproducibility outcomes. The study included 37 tests across 20 representative computational methods (10 methods from Methods Hub and 10 from other repositories), combining quantitative and qualitative analysis. Participants were randomly assigned to reproduce one method from each platform, following the original documentation, while we recorded reproducibility success, time spent, errors encountered, and assistance needed.

Evaluation results and outcomes

The comparative experiment revealed consistent advantages for Methods Hub in reproducibility and usability.

- **Reproducibility success rate:** Methods Hub 84.2% vs. External Repositories 72.2%.
- **Average time to reproduce:** Methods Hub 59.6 min vs. External 54.5 min (slightly longer due to structured reading and documentation).
- **Repository-related errors:** 55 total errors for Methods Hub vs. 65 for External Repositories.
- **Assistance required:** 48 instances for Methods Hub vs. 60 for External Repositories.

These findings show that Methods Hub users encountered fewer issues, achieved higher success rates, and required less external help while maintaining accessibility and usability, but this came at the cost of slightly longer reproduction times due to the more structured and detailed documentation workflow.

Results of the evaluation (KPIs)

The key performance indicators (KPIs) evaluate how effectively the curated methods accompanied by checklists in Methods Hub support reproducibility compared to external repositories. The indicators focus on reproducibility success, error frequency, required assistance, and time efficiency. Table 5.3.1 summarizes the main KPIs, including total error counts, while Table 5.3.2 details the distribution of error types.

Table 5.3.1. *Key Performance Indicators (KPIs) – Methods Hub vs. External Repositories*

KPI	Indicator	Methods Hub	External Repositories	Effect
Reproducibility rate	Successful task completion (%)	**84.2**	72.2	12 % improvement
Total errors	Total number of errors during task execution	**55.0**	65.0	15 % fewer total errors for MH
Assistance frequency	Number of external assistance instances	**48.0**	60.0	34 % fewer for Mh
Time to reproduce	Average minutes per method	59.6 min	54.5 min	~9.4% more time for MH

Table 5.3.2. *Error Code Distribution by Platform*

Platform	Code	Count	Proportion	95% Confidence Intervals
External	A	13/65	0.200	[0.104, 0.297]
External	B	46/65	0.708	[0.596, 0.821]
External	C	6/65	0.092	[0.020, 0.164]
Methods Hub	A	19/55	0.345	[0.216, 0.474]
Methods Hub	B	30/55	0.545	[0.414, 0.676]
Methods Hub	C	6/55	0.109	[0.025, 0.193]

Interpretation:

- **Code A – Participant issues:** Slightly higher for Methods Hub (19 vs. 13), indicating minor user-related errors such as unsequenced execution steps or syntax mistakes. These were generally recoverable and reflect normal learning variations among participants.
- **Code B – Repository issues:** Substantially lower for Methods Hub (30 vs. 46), demonstrating that curated methods with integrated checklists mitigate repository-related problems such as missing datasets, unclear instructions, and deprecated libraries.
- **Code C – System issues:** Identical across both platforms (6 each), encompassing environment-specific limitations, runtime configuration, or versioning issues.

Overall, the error code distribution confirms that Methods Hub significantly reduces repository-level reproducibility barriers, while participant and system-related issues remain comparable across platforms.

Results of measures of efficacy and effectiveness

The experiment confirms that the checklist integrated into Methods Hub significantly improves reproducibility and usability for computational social science methods. The 12% higher success rate demonstrates greater reproducibility efficacy, while the 15% reduction in errors and 34% lower assistance need indicate improved effectiveness and user autonomy. Although participants spent slightly longer reproducing methods (on average 59.6 vs. 54.5 minutes), this reflected the time invested in following clearer, structured guidance rather than troubleshooting errors.

Overall, the findings demonstrate that the lightweight checklist effectively supports reproducible, interpretable, and accessible workflows, establishing Methods Hub as a reliable, community-aligned infrastructure for computational social science research.

To complement these objective indicators, post-test questionnaire data captured participants' subjective assessments of usability and effectiveness. Results showed that participants rated their overall experience more positively for Methods Hub across all phases of the reproduction process—method exploration (3.4), code reproduction (3.1), and experimentation (3.3), compared to external repositories (3.0 / 2.9 / 3.2). Table 5.3.3 summarizes the quantitative outcomes from the questionnaire.

Table 5.3.3 *Post-Test Questionnaire – Quantitative Results (Methods Hub vs External Repositories)*

Measure	Methods Hub (Mean)	External Repositories (Mean)	Interpretation
Ease of following instructions	3.40	2.95	Participants found Methods Hub easier to understand and navigate.
Clarity of setup steps	2.95	2.83	Setup clarity was comparable, with a slight advantage for Methods Hub.
Result correctness and meaningfulness	3.30	2.83	Outputs on Methods Hub were perceived as more accurate and interpretable.

Qualitative Insights

Categorical and open-ended responses further highlight platform-specific differences:

- **Documentation and guidance:** Methods Hub documentation was perceived as coherent and centralized, while external repositories often lacked a single, complete source of setup instructions.
- **Code modification:** Methods Hub users typically made only minimal edits, whereas users of external repositories—especially GitHub—frequently needed to debug or rewrite multiple code segments.

- **Beginner accessibility:** Hugging Face was occasionally described as more beginner-friendly, but Methods Hub's structured workflow supported participants with a wider range of technical expertise.
- **Help and troubleshooting:** Users of external repositories relied more heavily on external help (AI tools or facilitators), while Methods Hub users usually resolved issues using on-page documentation.

In summary, both quantitative and qualitative measures confirm that checklist-supported workflows in Methods Hub substantially enhance reproducibility, usability, and interpretability compared with external repositories, reflecting stronger overall user satisfaction and reproducibility performance.

5.4. Discussion

Brief Summary of the Results

The Pilot demonstrated that Methods Hub, through curated and checklist-supported computational methods, substantially improves reproducibility outcomes in comparison with external repositories. Empirical evaluation showed a 12 % higher reproducibility success rate, fewer repository-related errors, and reduced reliance on external assistance, while maintaining accessibility for users across varying technical skill levels. The post-test questionnaire confirmed these results from a user perspective: participants rated ease of use, clarity of instructions, and output interpretability higher for Methods Hub than for GitHub or Hugging Face. Together, these findings validate the Pilot's approach of integrating lightweight, stage-aligned reproducibility checklists into a shared platform for computational social science.

Implications

The Pilot provides clear evidence that structured, checklist-supported approaches can make reproducibility practical and attainable within everyday research workflows. Its implementation in Methods Hub demonstrates that when reproducibility principles are embedded directly into digital infrastructures, they effectively bridge the gap between *awareness* and practice. This approach transforms reproducibility from a compliance task into an integrated research habit. For the broader reproducibility landscape, the Pilot illustrates how curated methods and transparent workflows can improve not only technical quality but also trust and accountability in computational research. It highlights the importance of coupling open science policies with usable, researcher-centred tools that minimize effort and cognitive load.

The Pilot's outcomes have direct usability implications for multiple stakeholder groups:

- Researchers benefit from a guided, low-barrier process to document and share methods reproducibly.
- Educators can use the platform to teach good computational and data-management practices through live, interactive examples.

- Funders and institutions gain tangible metrics—such as checklist adoption and reproducibility success rates—for evaluating open science commitments.
- Publishers and reviewers can adopt the checklists to standardize reproducibility expectations in the publication process.

Through these contributions, the Pilot reinforces reproducibility as a shared responsibility that can be supported through infrastructure, policy, and culture simultaneously.

Reflecting on Applicability Across Diverse Epistemic Contexts

The design of Methods Hub, focused on granular, executable methods rather than entire research projects, makes it inherently adaptable across diverse epistemic and disciplinary contexts. Different fields conceptualize reproducibility differently: while computational social scientists emphasize data access and algorithmic transparency, qualitative researchers may focus on interpretability and documentation of analytical steps. By providing a flexible checklist and metadata framework, Methods Hub accommodates these varying epistemic traditions.

In quantitative and computational disciplines (e.g., social network analysis, computational linguistics, digital humanities), the platform can host shareable workflows that ensure consistent execution across environments.

In qualitative and mixed-methods research, the same checklist structure can be used to document decisions, coding procedures, and data management processes, providing transparency without forcing standardization.

This adaptability positions Methods Hub as a boundary infrastructure, a shared space where researchers from different domains can exchange methods while maintaining their field-specific norms and standards.

Furthermore, by aligning technical reproducibility with epistemic diversity, the platform supports interdisciplinary collaboration. Researchers can better understand and reuse methods from other domains, while still contextualizing them within their own theoretical and methodological frameworks. Such interoperability strengthens the broader ecosystem of open, reproducible science and ensures that reproducibility tools evolve alongside disciplinary needs rather than imposing a one-size-fits-all model.

Methodological Reflection and Evaluation of the Pilot Process

The pilot process demonstrated that reproducibility is an evolving target rather than a static achievement. While the integration of checklists and curated methods in Methods Hub led to clear improvements—raising the reproducibility success rate to 84.2%—complete (100%) reproducibility was not yet achieved. The remaining cases of non-reproducibility primarily stemmed from environment-related challenges, such as inconsistent software versions, dependency mismatches, or limitations in cloud-execution settings. These issues are widely

recognized across computational research and underline that reproducibility success depends as much on technical infrastructure as on methodological documentation.

General Impression of the Process

From a process perspective, the Pilot successfully combined conceptual, empirical, and technical components. The iterative sequence (literature review, community consultation via surveys, experimental evaluation, and platform integration) proved effective in identifying both strengths and weaknesses of checklist-supported workflows. Participants' engagement during the experiment and workshop sessions provided essential feedback that guided refinements to documentation structure, checklist wording, and task clarity. The process thus functioned as a co-learning environment, where both developers and users contributed to improving the tool's functionality and accessibility.

Evaluation and Reflection on the Pilot Process (Including Perceptions of the stakeholders)

Evaluation of the pilot process went beyond measuring numerical outcomes to include continuous reflection on its design, implementation, and participant involvement. The Pilot team conducted internal review meetings and post-survey debriefings to assess progress and incorporate user feedback into iterative improvements. This reflective process confirmed that while the Pilot met its main objectives—demonstrating measurable gains in reproducibility and usability—there remains scope for further refinement.

Participant feedback played a central role in shaping this evolution. Across the surveys, user study, and workshop, participants described Methods Hub as a *structured, user-friendly, and trustworthy platform* that reduced barriers to reproducibility through curated methods and clear checklist guidance. At the same time, they identified recurring challenges such as environment inconsistencies and dependency issues, particularly relevant to computational workflows. These insights directly informed the next stage of platform development: the creation of an interactive, browser-based execution environment that allows users to run methods in preconfigured setups, minimizing configuration errors and version conflicts.

From a methodological perspective, the Pilot demonstrated that reproducibility improvement is best achieved through iterative cycles of testing, feedback, and adaptation. Participants did not merely evaluate the tool—they actively co-shaped it by identifying usability gaps and proposing enhancements, embodying a genuine co-creation process. This participatory and reflective approach strengthened both the Pilot's credibility and the long-term sustainability of Methods Hub as a continuously improving research infrastructure.

Overall, the Pilot achieved more than a proof of concept: it established an iterative learning process between developers and users, producing concrete improvements while remaining adaptable to future needs and broader community adoption.

Relation to Existing Literature

The findings of this Pilot align with prior research emphasising the persistent gap between awareness of reproducibility and its practical implementation in scientific workflows (Archmiller et al., 2020; Nosek et al., 2022; Lazer et al., 2020; Schoch et al., 2023). Consistent with earlier studies showing that sharing code and data alone is insufficient for true computational reproducibility (Clyburne-Sherin et al., 2019; Hardwicke et al., 2022), the Pilot confirms the necessity of clear environment documentation, dependency management, and structured metadata.

By empirically validating checklist-supported workflows, this work complements existing reproducibility frameworks such as the Transparency and Openness Promotion (TOP) Guidelines (Nosek et al., 2016)—already implemented across journals in multiple disciplines including social sciences, health, life and physical sciences—and the DIME Standards developed by the World Bank, demonstrating how these principles can be operationalized within computational social science through concrete tooling and platform integration.

Moreover, the Pilot contributes concrete evidence to complement recent calls for practical, community-driven solutions to reproducibility challenges in data-intensive disciplines (Lazer et al., 2020; Schoch et al., 2023).

Strengths & Limitations

Strengths:

- Combination of community-based checklist design and experimental validation.
- Integration into an existing open infrastructure (Methods Hub) ensuring immediate usability.
- Evidence from multiple data sources (surveys, experiment, qualitative feedback).

Limitations:

- The experiment sample size was moderate (n = 37) and may not represent all computational social science domains.
- Results reflect short-term usability; long-term adoption and sustainability remain to be assessed.
- Participant heterogeneity introduced some variation in technical ability and tool familiarity, potentially influencing performance metrics.

Despite these limitations, the Pilot provides a strong empirical foundation for reproducibility support tools in computational social science research.

Future Work

After the conclusion of TIER2, the team will continue assessing and enhancing the Methods Hub platform, now transitioning into its public launch phase. The official public release campaign will

be featured on the GESIS main page, across institutional social media channels, and in the GESIS newsletter. This marks a critical step from prototype to operational research infrastructure and introduces a new phase of performance monitoring and impact assessment.

Future work will therefore include:

- **Defining and tracking new service-level KPIs** to evaluate the success of the public rollout, focusing on user engagement, number of uploaded methods, checklist adoption rate, and cross-disciplinary usage.
- **Continuous usability evaluation** through user analytics and feedback from newly onboarded researchers.
- **Expanding the repository** of curated methods and reproducibility checklists beyond computational social science to related domains.
- **Developing onboarding and training materials** to support new contributors and educators.
- **Long-term sustainability planning**, ensuring open-source maintenance, community moderation, and institutional integration of Methods Hub within GESIS infrastructure.

This transition from a project Pilot to a publicly launched service represents the culmination of the TIER2 Pilot's objectives and the beginning of a new phase focused on scalability, monitoring, and real-world impact.

Recommendations

The Pilot's results provide a strong empirical basis for advancing policy and infrastructure support for reproducibility across the research ecosystem. They show that simple, well-structured tools, when embedded in research workflows, can substantially improve reproducibility, reduce technical barriers, and foster cultural change toward open and transparent science.

Based on these insights, the following recommendations are proposed:

For researchers:

- Adopt reproducibility checklists and structured documentation as part of standard research practice.
- Share executable methods through trusted platforms such as Methods Hub to increase transparency and visibility.
- Incorporate reproducibility assessments into peer mentoring and supervision for early-career researchers.

For institutions and funders:

- Recognize and reward reproducible practices by including checklist compliance or method sharing as *optional and context-sensitive* evaluation criteria, ensuring that such expectations do not disadvantage researchers working with sensitive data, proprietary datasets, or disciplines where full openness is not feasible.

- Support long-term maintenance and integration of open infrastructures like Methods Hub to ensure sustainability and interoperability with institutional repositories.

For journals and publishers:

- Implement reproducibility checklists as part of manuscript submission and review workflows to standardize expectations.
- Encourage or require the use of open, executable repositories for code and data, linked to persistent identifiers.

For educators and training initiatives:

- Use curated examples and checklists from Methods Hub as teaching resources for computational methods and open science courses.
- Integrate reproducibility training into curricula to build the next generation of researchers who are fluent in open, transparent practices.

Together, these recommendations highlight that reproducibility is not only a methodological standard, but a systemic responsibility shared by all actors in the research ecosystem. The Pilot's approach demonstrates how technical tools and policy frameworks can reinforce one another to make reproducibility both achievable and sustainable.

5.5. Conclusions

The Pilot demonstrated that reproducibility can be effectively operationalized through the integration of lightweight, checklist-supported tools within open research infrastructures. By combining empirical evidence, user-centred design, and community feedback, the Methods Hub platform has evolved into a practical and scalable solution for improving the transparency and reliability of computational social science.

The evaluation results (showing a 12 % increase in reproducibility success, fewer repository errors, and higher user satisfaction) underscore the value of embedding reproducibility directly into digital workflows. Equally important, the Pilot highlighted that achieving reproducibility is an iterative and collaborative process requiring continuous refinement, feedback, and infrastructure support. Moving forward, Methods Hub will enter its public launch phase, expanding access to a broader community of users and introducing service-level KPIs to monitor adoption and impact.

This transition marks the culmination of the TIER2 pilot's objectives and the beginning of a sustained effort to integrate reproducibility into everyday scientific practice. Ultimately, the pilot demonstrates that reproducibility is attainable when technical innovation, community engagement, and institutional commitment are aligned within a shared framework for open, responsible, and verifiable research.

5.6.References

Altmejd, A., Dreber, A., Forsell, E., Huber, J., Imai, T., Johannesson, M., ... & Camerer, C. (2019). Predicting the replicability of social science lab experiments. PloS one, 14(12), e0225826.

Archmiller, A. A., Johnson, A. D., Nolan, J., Edwards, M., Elliott, L. H., Ferguson, J. M., ... & Fieberg, J. (2020). Computational reproducibility in The Wildlife Society's flagship journals. The Journal of Wildlife Management, 84(5), 1012-1017.

Clyburne-Sherin, A., Fei, X., & Green, S. A. (2019). Computational reproducibility via containers in psychology. Meta-psychology, 3.

Ferguson, J., Littman, R., Christensen, G., Paluck, E. L., Swanson, N., Wang, Z., ... & Pezzuto, J. H. (2023). Survey of open science practices and attitudes in the social sciences. Nature communications, 14(1), 5401.

Hardwicke, T. E., Wallach, J. D., Kidwell, M. C., Bendixen, T., Crüwell, S., & Ioannidis, J. P. (2020). An empirical assessment of transparency and reproducibility-related research practices in the social sciences (2014–2017). Royal Society Open Science, 7(2), 190806.

Hardwicke, T. E., Thibault, R. T., Kosie, J. E., Wallach, J. D., Kidwell, M. C., & Ioannidis, J. P. (2022). Estimating the prevalence of transparency and reproducibility-related research practices in psychology (2014–2017). Perspectives on Psychological Science, 17(1), 239-251.

J. Bednar. “Reproducibility: Future-proofing your python projects”. https://www.anaconda.com/blog/8-levels-of-reproducibility (accessed 11-10, 2023).

Kohrs, F. E., Auer, S., Bannach-Brown, A., Fiedler, S., Haven, T. L., Heise, V., ... & Weissgerber, T. L. (2023). Eleven strategies for making reproducible research and open science training the norm at research institutions. Elife, 12, e89736.

Lazer, D. M., Pentland, A., Watts, D. J., Aral, S., Athey, S., Contractor, N., ... & Wagner, C. (2020). Computational social science: Obstacles and opportunities. Science, 369(6507), 1060-1062.

Nosek, B. A., Alter, G., Banks, G. C., Borsboom, D., Bowman, S., Breckler, S., ... & DeHaven, A. C. (2016). Transparency and openness promotion (TOP) guidelines.

Nosek, B. A., Hardwicke, T. E., Moshontz, H., Allard, A., Corker, K. S., Dreber, A., ... & Vazire, S. (2022). Replicability, robustness, and reproducibility in psychological science. Annual Review of Psychology, 73, 719-748.

Patarčić, I., & Stojanovski, J. (2022). Adoption of transparency and openness promotion (TOP) guidelines across journals. Publications, 10(4), 46.

Pham, M. T., & Oh, T. T. (2021). Preregistration is neither sufficient nor necessary for good science. Journal of Consumer Psychology, 31(1), 163-176.

Playford, C. J., Gayle, V., Connelly, R., & Gray, A. J. (2016). Administrative social science data: The challenge of reproducible research. *Big Data & Society*, *3*(2), 2053951716684143.

Samuel, S., & Mietchen, D. (2023). Computational reproducibility of jupyter notebooks from biomedical publications. *arXiv preprint arXiv:2308.07333*.

Schoch, D., Chan, C. H., Wagner, C., & Bleier, A. (2023). Computational Reproducibility in Computational Social Science. arXiv preprint arXiv:2307.01918.

6. Pilot 5 - Reproducibility Promotion Plans for Funders

Authors: Barbara Leitner, Friederike Elisabeth Kohrs, Alexandra Bannach-Brown, Joeri Tijdink

6.1. Introduction

Funders hold incentives which can impact cultural norms. They are in a unique position to foster systemic change in the research ecosystem by promoting Open Science, and incentivizing reproducible research practices (Liu et al., 2022). Funding organizations can require research groups and individual researchers to make explicit commitments to reproducibility. Therefore, funders should treat reproducibility as a key criterion for the research they fund (Bishop, 2015), as only then can they promote and push for a change in research culture which values transparency, openness, and reproducibility. It is beneficial to instill these values at the funding level as it creates a culture of incentivizing, educating, and empowering researchers instead of policing the quality of outputs at the end (e.g. peer reviewers or replication attempts) (Munafò et al., 2014).

There are some funders who are already paving the way towards adopting and requiring reproducibility and Open Science practices in the research they fund, for example the Netherlands Organization for Scientific Research (NWO) and Open Science NL. They not only require different Open Science practices but have also specific funding lines for replication studies in place. Other funding organizations, such as the National Institute of Health (NIH) and CHDI Foundation, have introduced policies and pre-incorporated requirements into criteria for grant applications or evaluation committees. Whilst there are increasingly available and visible initiatives within funding organizations, they are bound to specific research fields or geographical contexts, limiting their impact and usability for the wider research community. Additionally, between funding agencies, there is a lack of cohesion and information shared on their best practices and initiatives to promote rigor and reproducibility in the research they fund. Therefore, we have identified a need for an initiative that can be used by funders of different capacities, levels of readiness, and epistemologies. Within this Pilot, we aimed to co-create a tool together with funders that can meet their actual needs. Based on the insights shared by an international group of funders, the Reproducibility Promotion Plan for Funders (RPP) ensures that the policy recommendations fulfil and match the identified requirements of funders.

We proposed the following research questions to inform our co-creation activities with funders:

Research question: What can funders do to foster reproducibility?

Sub-questions:

1. What topics are important for funders to foster reproducibility within their funding practices?

   1a. What can funders do to internally foster reproducibility?

   1b. What can funders do to externally foster reproducibility?

2. How should policy in the form of a reproducibility promotion plan look?

The full pre-registration for Pilot 5 can be found on OSF (https://osf.io/tuz62).

6.2. Methodology

We recruited individuals based on the following inclusion criteria: international funders who are working or worked in research funding institutions and have demonstrable experience and/or expertise with issues pertaining to reproducibility and Open Science in the realm of funding. This includes contributing to projects or procedures that aim to improve reproducibility or address related issues. Participants were recruited through a three-fold recruitment strategy: firstly, through existing connections and networks (including the previously established TIER2 stakeholder community) of the TIER2 consortium, secondly through snowballing, and lastly, through advertising the workshops publicly on the TIER2 website. Over three workshops, including the evaluation workshop (see below), we had a total of thirteen participants from different international funding organizations of varying sizes, capacities, and levels of readiness. We had two funding organizations Pilot, one international and one national funder, the RPP. Additionally, we had two funders provide detailed feedback on the RPP through a survey.

Ethical approval

Ethical approval was obtained from Ethical Review Board at the AmsterdamUMC (2024.0215), under a non-WMO declaration as the research does not fall within the reach of the Dutch Law on medical research.

Research Design

We developed the output of Pilot 5 through a co-creation process together with funders. During this process, we held two interactive online workshops using the meeting platform Zoom and a virtual collaborative platform called Miro.

First Co-creation Workshop

The first co-creation workshop focused on developing the essential themes and elements of the RPP using various converging exercises. Prior to the workshop we asked two open-ended questions: "what does reproducibility mean to you?" and "what reproducibility-promoting practices are currently in place for funders?". Furthermore, example texts from existing funder policies were presented to help inspire and engage funders during the workshop. The first exercise during the workshop was a free writing exercise, where funders were asked to write down as many ideas of what they would like to include in their ideal policy template. Then each participant was asked to pitch their top three ideas to the group. Following this, participants collectively discussed and clustered their ideas to identify emerging overarching themes. The second exercise was built on the previous one, and now participants were asked to think of the specific practicalities of such a policy template. This allowed them to identify enablers and barriers within and outside of funding organizations.

Second Co-creation Workshop

Participants in the second workshop consisted of two from the first workshop and one new participant. Prior to the second co-creation workshop, participants were sent the clustered overarching themes identified in the previous workshop. During the second workshop, participants were then asked to reflect on the draft policy template which was based on the input from the first

workshop and created by the TIER2 team. Participants were asked to write down specific recommendations and feedback per section of the policy template. After this, they were asked to prioritize their most important recommendations and narrow them down at most one. Afterwards, participants were tasked to write down any potential barriers they could envision for the specific theme and note down ways to overcome individual barriers. Additionally, participants had the opportunity to provide best practices for specific recommendations. Following this, participants were encouraged to provide general comments, additions, and suggestions to the discussed content. This included redundancies, gaps, unclarities, conflicting ideas, etc.

Evaluation Workshop

Following the content analysis by the core research team (colleagues from AmsterdamUMC and Charite) within TIER2, the first draft of the RPP was completed. This draft was disseminated to the workshop participants prior to the evaluation workshop. Within the workshop, two groups were created, each of them commenting on two of the three sections of the RPP. For each section, participants were first asked for their general impressions and feedback on the recommendations and then encouraged to think of any additional barriers or enablers that were not previously identified. Throughout the evaluation workshop, additional best practices were collected for each recommendation. General feedback was collected at the end of the workshop.

Survey

Alongside the evaluation workshop, we disseminated a detailed survey within TIER2's funder stakeholder community. The survey offered the opportunity to include detailed feedback on each recommendation assessing its clarity and feasibility. Survey participants could further suggest additional recommendations not included in the RPP. The survey's structure and content can be found on OSF.

Pilot

We piloted the RPP with two different funding institutions over a six-month period. The TIER2 Pilot team helped assess the specific needs of each pilot institution and created an individual pilot plan to implement the relevant recommendations. The pilot phase was accompanied by a series of monthly thirty-minute meetings. The first two meetings specifically focused on identifying the needs of the funding institutions and assessing which themes and recommendations to focus on during the pilot phase. Based on this, the research team developed a pilot plan which was shared with the funding organization prior to the next meeting. The selected recommendations were then incorporated into existing grant documents, and content was refined through iterative rounds of feedback between the TIER2 research team and the funding organization.

Analysis

After the second co-creation workshop, we analyzed the data using deductive qualitative analysis with the 'analysis on the wall' method (Sanders and Stappers, 2012). The data included all materials produced by participants during the workshops: written notes on post-its, audio transcripts, and facilitator notes.

These themes were divided among four TIER2 researchers, three researchers worked on two themes each, and one on three. Individually, each researcher:

1. Identified subthemes for their assigned themes,
2. Drafted recommendations and explored connections between them,
3. Linked best practice examples to the recommendations, and
4. Identified barriers and enablers related to those recommendations.

The team then met in a live co-creative session to compare and refine their analyses, reducing overlap among the five themes.

Pilot Evaluation

The metrics employed for the evaluation of this Pilot are qualitative in nature. During the follow-up interviews with the funding organizations piloting the RPP recommendations, multiple qualitative metrics were used:

- Funder satisfaction with the RPP: Funders indicated their satisfaction with the use of the RPP on a Likert scale of 1 to 5 (where 1 is unhappy and 5 is very happy)
- Adoption Rate: 1-5 Likert scale (where 1 means not adopted at all and 5 adopted multiple times/multiple projects)

We also evaluated the piloting process with funders, diving into how they found working with the tool and the piloting process and what needs to be changed to assess the usability of the tool. We also discussed with them if they see themselves using the RPP again in the future. Discussions into whether they can see themselves using the RPP in the future.

We also planned to assess funder satisfaction with change in researcher's behaviour and the compliance of researchers after the piloting period. However, due to the timeline of the Pilot this was not feasible with any of our pilot institutions.

The full interview guide can be found on OSF.

Pilot Synergy

Pilot 5 collaborated with Pilot 2 and Pilot 6 as a synergy. We used the tools from the Pilots as best practice examples and facilitated meetings with the piloting funding institutions. We evaluated the synergies through the evaluation interviews, assessing the piloting institutions' views of the synergy and their views on the tools.

6.3. Results

During the first co-creation workshop, participants identified five key themes for the Reproducibility Promotion Plan (RPP): motivation, incentives and recognition, monitoring, definition, and “the

how?". In the second workshop, funders developed specific recommendations for each theme, including examples of best practices, and identified barriers and enablers.

As described earlier, the research team conducted a synchronous analysis session to synthesize these five themes into three overarching categories: policy and definitions, evaluation and monitoring, and incentives. The recommendations were re-organized to reflect the new categories and re-structured in order to build progressively on one another, from basic, easy-to-implement actions to more advanced recommendations.

Subsequently, we held a validation workshop, where we collected additional best practices and incorporated them into the RPP, and all feedback was implemented. By the end of the co-creation process, the RPP had evolved into a comprehensive, multi-page policy template containing specific recommendations linked to best practices, barriers and enablers. Further, it contained practical guidance to support implementation.

After the pilot period, separate evaluation interviews were conducted with representatives of the funding organizations. Both funders reported that the RPP was clear, user-friendly, and useful in clarifying what actions had been completed and what remained to be done within their institutions. They noted that the RPP can serve both as a preparatory tool for funders beginning to integrate language, mandates, or recommendations aimed at strengthening reproducibility into their funding calls and as a resource for those already more advanced in this process.

Both piloting institutions found the recommendations, featured in the monitoring and policy section of the RPP, particularly valuable and have begun implementing them. However, they also reported that internal barriers, such as limited time and bureaucratic challenges, continue to slow down further progress.

During the piloting phase, participants found it difficult to rate the adoption level on a Likert scale because within their smaller teams, they felt the RPP was already well integrated and close to full adoption. However, at the organizational level, they anticipated that broader implementation would require more time and further discussion.

6.4. Discussion

After the co-creation, piloting, and evaluation processes, the Reproducibility Promotion Plan for Funders (RPP) was refined to better align with and meet the needs of funding institutions. Based on these co-creative, stakeholder-driven processes, we believe the RPP is an important tool to drive necessary changes within funding organizations to enhance the reproducibility and openness of the research they fund.

While the RPP was co-created by funders and designed primarily for funding institutions, the evaluation process indicated that it may also be relevant and applicable to other stakeholder communities. These may include data managers, research policy makers within academic institutions (particularly those involved in research assessment), and others. For these groups, the RPP can serve both as guidance on how to integrate reproducibility into their own work and as a framework to assess what has already been achieved and what further steps may be needed.

Importantly, the RPP is not only suitable for different types of stakeholders but also adaptable across diverse epistemic contexts. Already in the early phase of the development of the RPP, ensuring epistemic applicability was a key objective, recognizing the wide variability among funding institutions and the types of research they support. This was reflected in our participant recruitment strategy, which targeted individuals with varied epistemic backgrounds. Additionally, we consulted with experts within the TIER2 project to ensure that the recommendations were broadly applicable. Further, we included best practices for the specific recommendations and themes that could be implemented across different disciplines. One of the institutions, piloting the RPP, highlighted this as a particularly valuable aspect, especially in relation to policy guidance and defining reproducibility.

Ease of use was another central design principle. The RPP was intentionally created to be user-friendly and accessible, structured as a policy template containing recommendations at multiple levels, from introductory to advanced, accompanied by implementation guidelines and best practices from existing funding institutions. The described barriers and enablers further support funders to anticipate potential challenges and identify ways to overcome them.

Feedback from the two pilot institutions confirmed that the RPP was intuitive and easy to use. Both institutions appreciated the guidance provided during the piloting process but noted that future users could likely apply the RPP independently from the TIER2 research team. To support the independent use, the final version of the RPP includes detailed “how-to-use” instructions and a visual overview to help users identify when and how specific recommendations can be implemented.

The stakeholder-driven, co-creation and piloting phases were essential for developing the RPP, as they allowed the project team to draft recommendations which reflect the actual needs and perspectives of funders regarding the embedding of reproducibility within their own funding activities. The TIER2 funder community and participating institutions played a crucial role here—not only in creating the RPP but also in refining and strengthening it through reflection and evaluation.

Despite these successes, some limitations of the co-creation and pilot processes need to be acknowledged. The participant sample was somewhat biased, as many participants already possessed knowledge of and interest in reproducibility and Open Science. To address this, the RPP includes several recommendations emphasizing the importance of reproducibility for funders, along with low entry-level recommendations for those with limited prior experience. Another limitation is the small number of pilot institutions; however, the inclusion of both an international and a national funder demonstrates the RPP’s relevance and adaptability across different institutional contexts.

Work on the RPP will continue in future European-funded projects focused on reproducibility, such as TRUSTparency (www.trustparency.eu). Efforts will aim to further promote its use among funding organizations and to keep it updated with latest best practices to ensure continued relevance and effectiveness. There is already a funder selected that will further pilot our guideline.

Ultimately, the RPP serves as a valuable tool for policy development, helping funding institutions identify and strengthen their internal policies related to reproducibility. We recommend that funders begin by assessing their internal needs and current level of engagement with reproducibility, then use the RPP to identify and prioritize areas of actions and implement the relevant recommendations accordingly.

6.5. References

Bishop, D. (2015, November 24). *Improving reproducibility: What can funders do?* (Guest post). Retraction Watch. Retrieved from https://retractionwatch.com/2015/11/24/improving-reproducibility-what-can-funders-do-guest-post-by-dorothy-bishop/

Liu, J., Carlson, J., Pasek, J., Puchala, B., Rao, A., & Jagadish, H. V. (2022). *Promoting and Enabling Reproducible Data Science Through a Reproducibility Challenge*. Harvard Data Science Review, 4(3). https://doi.org/10.1162/hasr_a_00319

Munafò, M. R., Nosek, B. A., Bishop, D. V. M., Button, K. S., Chambers, C. D., Percie du Sert, N., Simonsohn, U., Wagenmakers, E.-J., Ware, J. J., & Ioannidis, J. P. A. (2017). *A manifesto for reproducible science*. *Nature Human Behaviour*, 1(1), Article 0021. https://doi.org/10.1038/s41562-016-0021

7. Pilot 6 - Reproducibility Monitoring Dashboard

Authors: Haris Papageorgiou, Stefania Amodeo, Petros Stavropoulos

7.1. Introduction

The Reproducibility Monitoring Dashboard provides stakeholders (i.e., funding agencies, Research organizations) with tracking and monitoring capabilities to evaluate the adoption and implementation of reproducible research practices.

The purpose of this Pilot is to enhance transparency in research by offering a systematic way to monitor reproducibility metrics, supporting both policy development and compliance assessment.

This overarching goal is further decomposed in the following objectives addressing relevant and important research questions:

- Develop and test robust and explainable tools for tracking major research artefacts (e.g., datasets, software),
- Quantify and estimate Reusability indicators based on different types of artefacts,
- Develop good proxies of reproducibility & replicability, alleviating the relevant work of funding agencies and RPOs and at the same time providing evidence-based insights on the impact of their policies,
- Design & implement a dashboard enabling funding agencies & RPOs in tracking & monitoring reusability of research artefacts (datasets, software, tools/systems, etc) created in funded projects in an efficient and effective way.

7.2. Methodology

Our stakeholder recruitment strategy targeted individuals from Research Performing Organizations (RPOs) and Research Funding Organizations (RFOs) with specific interest in reproducibility monitoring:

- professionals from funding organizations responsible for overseeing funded projects and evaluating resource allocation
- representatives from research institutions who manage institutional research outputs and seek to enhance transparency in their projects

Twenty-three unique participants contributed to our co-creation process through two workshops (October 2024 and June 2025). The first workshop included 13 participants from RFOs, while the second featured 4 RFO and 10 RPO representatives. Four funders attended both workshops, ensuring continuity throughout the process.

Our registration process captured detailed information including:

- organizational type (research institution, funder, or other)

- strategic focus areas and primary interests
- current metrics and reporting practices

This participants' data allowed us to develop tailored use cases that addressed specific organizational contexts, following a structured storytelling format that resonated with participants' professional experiences (see more details in the r*esearch design* section below).

We used the following co-creation methods to incorporate stakeholder needs into our dashboard:

- **Interactive Workshops:** we hosted two workshop sessions with distinct purposes. The first workshop (October 2024) introduced key concepts, facilitated discussions about reproducibility indicators, and gathered stakeholder requirements. The second workshop (June 2025) presented the detailed prototype and collected suggestions for refinement.

- **Use Cases:** we employed a structured storytelling approach to help stakeholders articulate concrete organizational contexts by answering five key questions:

    - *Who am I and what are my primary responsibilities?* - Organizational Identity
    - *What specific reproducibility outcomes do we aim to achieve?* - Strategic Objectives
    - *Which metrics will effectively track our progress?* - Measurement Criteria
    - *Where can we implement improvements?* - Optimization Opportunities
    - *What specific dashboard features support our decision-making?* - Action Plan

- **Surveys:** we collected feedback through structured surveys distributed to webinar attendees during and after the webinars, using both quantitative rating questions and qualitative open-ended questions to capture nuanced feedback and user satisfaction. We also conducted individual follow-up discussions with key stakeholders to explore specific cases. Our surveys assessed several key aspects of the dashboard:

    - Requirements for documentation sources
    - Priorities among different research artifacts (datasets, software, methods, etc.)
    - Ratings of proposed reproducibility indicators
    - Evaluation of dashboard features and visualization styles
    - Open-ended feedback on missing features and improvement suggestions

- **Iterative Prototype Development:** we maintained a continuous development cycle where each dashboard iteration incorporated stakeholder feedback, ensuring alignment with evolving user requirements and use cases.

- **Data Analytics**: The analysis was based on research outputs from CORDIS projects, focusing primarily on publications. Using the SciNoBo toolkit, we automatically detected, clustered, and classified research artefacts (datasets, software, etc.) mentioned in these outputs. Those artefacts were then linked with project- and publication-level metadata from CORDIS, OpenAIRE, and Semantic Scholar to form the basis of the pilot dataset.

Within this processing pipeline, different SciNoBo components were employed: the **Research Artefact Analysis** tool identified and aggregated artefacts; the **Citance Analysis** tool examined citation contexts in terms of intent (why the work was cited), polarity (whether the citation was supportive, neutral, or critical), and semantics (which aspect of the work was referenced, such as a *method, result, or claim*); the **Field of Science Classification** tool placed outputs in their disciplinary context; and the **Citation Impact Analysis** tool provided bibliometric benchmarks such as Field-Weighted Citation Impact (FWCI). In combination, these tools enabled the estimation of impact, reuse, and reproducibility indicators, including the FWCI, the Field-Weighted Reusability Index (FWRI), the FAIR Index, and composite reproducibility metrics, all of which were normalised across disciplines.

The resulting interactive dashboard integrated those KPIs into views providing filters based on project, publication, research artefact, organisation, country, and field of science. This enabled stakeholders to move seamlessly between a high-level overview of reproducibility and reusability performance and the specific organisations, countries, or artefacts driving the KPIs. The dashboard also supported the exploration of temporal trends, facilitating comparisons across contexts and highlighting strengths as well as potential gaps. These capabilities were demonstrated in the pilot webinars as a foundation for stakeholder engagement and discussion.

Evaluation

Evaluation activities included gathering structured feedback during the workshops, where stakeholders were asked to assess the dashboards and underlying indicators in terms of usability, clarity, and relevance. This qualitative feedback was systematically analysed to implement improvements, enhance user-friendliness, and prepare the tool for broader adoption. The evaluation process will conclude with a final webinar and satisfaction survey in Q4 2025, where we will present representative dashboards for RPOs and funders.

The Pilot's KPIs measure three key dimensions: quantitative performance (KPI1), implementation success (KPI2), and user satisfaction (KPI3). Together, these metrics provide a holistic view of the dashboard's effectiveness across technical capabilities, adoption, and user experience. Our evaluation plan has remained consistent with the original design throughout the pilot lifecycle.

7.3. Results

The piloting process validated the core concept of the Reproducibility Monitoring Dashboard and identified areas for enhancement that were subsequently incorporated into the resulting prototype. The first survey explored **five key themes** related to reproducibility monitoring:

Project Portfolio Documentation

Participants identified publications and project deliverables as essential sources for tracking research outputs. Data Management Plans (DMPs) were noted as relevant but with limitations: in fact, they are typically created at project start, they could complement tracking if updated throughout the project lifecycle, aligning with some funders' existing requirements.

Research Artifact Identification and Metadata Extraction

Participants rated artifacts for assessing reproducibility: Datasets and Results (21% each), Software and Methods (19% each), and Claims (15%). Other mentions included analytic code and hardware specifications, highlighting reproducibility's multifaceted nature.

Reproducibility Indicators and Proxies

Participants rated four factors on a 1-5 scale: quality of documentation (3.9), positive reception (3.7) and reuse frequency (3.5). Citation count rated lowest (2.3), indicating it is insufficient alone. Discussion emphasized that good documentation does not guarantee reuse, and vice versa. Additional suggested factors included self-citations and citation-reuse correlations.

Quantitative Reproducibility Metrics

Six indicators were presented:

- Field-Weighted Citation Impact (FWCI): Citation impact relative to works in the same field/Research area,
- Field-Weighted Reusability Index (FWRI): Reuse frequency normalized by discipline
- Reusability Index: Composite of FWCI and FWRI
- FAIR Index: Metadata presence score (0-1)
- Repro Confidence Index: Based on supporting/neutral/refuting citations
- Reproducibility Composite Confidence Index: Combines multiple indices

All indicators scored above 3.5. Participants emphasized the importance of maintaining granularity when combining indicators.

Interactive Dashboard Features

Participants ranked their feature priorities from highest to lowest importance. The most valued feature was analytics and evidence presentation, which would allow users to interpret and communicate their findings effectively. Data download capability ranked second, enabling users to export information for further analysis or reporting purposes. Charts and graphs came in third, providing visual representations of the data. AI assistant integration received moderate interest, though participants expressed concerns about the need for transparency in how AI-generated insights are produced. Report generation functionality ranked lowest among the proposed features.

The second survey revealed strong positive reception for the dashboard prototype, with participants providing detailed ratings across multiple dimensions.

Dashboard Overview

The features presented in the overview page of the dashboard received particularly high marks. Field of Science classification scored highest (5.0), followed by category-based classification (4.9)

and geographic visualization (4.8). Country and organization filters both rated 4.4. Complex metrics received still positive but lower scores: composite indices (RI and RCCI) at 3.8, and field-weighted metrics (FWRI and FWCI) at 3.4.

Dashboard Pages

When evaluating the detailed pages, users found comparative analysis most valuable, specifically the ability to compare countries and fields within their own country while tracking changes over time. For visualization styles, distribution charts and graphs rated highest at 4.2, followed by detailed lists with metrics at 4.0, and rankings at 3.6.

User Experience

The overall user experience received favourable ratings, with ease of use scoring 4.2. Users identified one area for improvement: clearer guidance to distinguish between the "In" and "Out" versions of the dashboard. When asked about missing features for detailed views, participants offered no additional suggestions.

Overall, the survey demonstrates that the dashboard's core functionality resonates well with users, particularly its classification and filtering capabilities. To ensure clarity in the methodology and definition of metrics, comprehensive documentation was developed in the form of a handbook available for download directly from the dashboard. Additionally, interactive info boxes were added throughout the dashboard to guide users.

Results against KPIs

KPI1: Reusability Analysis Rate (see “quantitative performance” below)
KPI2: Adoption Rate: implementation of 2 dashboards: EC (RFO), Athena RC (RPO)
KPI3: User Satisfaction Scores: average satisfaction scores obtained from user assessment at different stages of the Pilot.

Quantitative performance

The **SciNoBo Research Artefact Analysis (RAA)** and **SciNoBo Citance Analysis (CA)** tools form the analytical foundation of the Pilot. Their evaluation aimed to confirm that the automated extraction, classification, and analysis steps supporting the dashboard’s indicators are reliable and accurate across diverse research domains.

All evaluations were conducted using publicly available benchmark datasets and purpose-built evaluation data. The following metrics were used to assess accuracy and reliability:

- **F1-score:** a combined measure that balances correctness (**precision**) and completeness (**recall**), showing how accurately the tool identifies and classifies information overall.
- **Exact Match (EM):** how often the tool retrieved an answer that exactly matches the correct value.

- **Lenient Match (LM):** how often the tool retrieved a value that was approximately correct (e.g., minor differences in formatting or phrasing).

These results show the quantitative performance of the SciNoBo components that underpin the reproducibility indicators and visualisations in the dashboard.

SciNoBo Research Artefact Analysis (RAA)

The RAA tool identifies mentions of research artefacts such as datasets and software within publications and extracts metadata such as their name, license, version, and online link. It also classifies whether the artefact was **created by the authors** (Provenance) or **reused from another source** (Usage). These capabilities are essential to calculate indicators such as the **FAIR Index** and **reuse ratios** that appear in the dashboard.

Validation on the SciNoBo RAA Evaluation Dataset

The first validation phase used the **SciNoBo RAA Evaluation Dataset**, a manually curated collection of publication snippets containing complex combinations of artefact mentions. This dataset was designed to assess how well the tool can handle both simple and complex cases, including unnamed artefacts or those with overlapping references.

Table 7.3.1. *Performance comparison of SciNoBo RAA model against Flan-T5 XL baseline on research artefact attribute extraction tasks.*

Task	Metric	SciNoBo RAA	Competitor Model	Competitor F1/Score
Artefact Mention	F1-score	**0.96**	Flan-T5 XL	0.82
Name	F1-score	**0.85**	Flan-T5 XL	0.60
	Exact Match (EM)	**0.84**	Flan-T5 XL	0.70
	Lenient Match (LM)	**0.91**	Flan-T5 XL	0.83
License	F1-score	**0.96**	Flan-T5 XL	0.95
	Exact Match (EM)	**0.69**	Flan-T5 XL	0.64
	Lenient Match (LM)	**0.82**	Flan-T5 XL	0.78
Version	F1-score	**0.98**	Flan-T5 XL	0.94
	Exact Match (EM)	**0.76**	Flan-T5 XL	0.69

	Lenient Match (LM)	**0.77**	Flan-T5 XL	0.87
URL	F1-score	**0.98**	Flan-T5 XL	0.97
	Exact Match (EM)	**0.57**	Flan-T5 XL	0.50
	Lenient Match (LM)	**0.60**	Flan-T5 XL	0.53
Usage (reuse)	F1-score	**0.92**	Flan-T5 XL	0.77
Provenance (ownership)	F1-score	**0.93**	Flan-T5 XL	0.65

The SciNoBo RAA performs very strongly on this evaluation dataset, showing high accuracy across all evaluated tasks. It consistently identifies research artefact mentions with a precision level that exceeds comparable large language model baselines, while also extracting detailed metadata such as licenses, versions, and URLs with strong alignment to reference data.

Particularly noteworthy is the tool's performance in identifying **Usage (reuse)** and **Provenance (ownership)**, where it correctly distinguishes between artefacts that were created within a study and those reused from external sources. This capability is fundamental for the reproducibility monitoring objectives of the Pilot, as it enables the differentiation between an organisation's produced and reused outputs; the two analytical dimensions visualised in the dashboard.

High Exact and Lenient Match scores in metadata fields further indicate that the tool not only detects the presence of artefacts but also retrieves the associated descriptive information with a high degree of accuracy, even when the metadata is inconsistently formatted or abbreviated in publications. This means that the RAA can effectively process large and heterogeneous corpora, generating reliable data for indicators such as the **FAIR Index**, **Field-Weighted Reusability Index (FWRI)**, and other reuse and reproducibility measures integrated in the dashboard.

Validation on Public Benchmark Datasets

RAA was also evaluated on two well-established benchmarks commonly used in research artefact detection studies:

- **DMDD-E+**, which focuses on dataset mentions in scientific publications.
- **SoMeSci_test+**, which focuses on software mentions and related metadata (license, version, URL, and usage).

These datasets are widely used in the research community, allowing a transparent comparison against previously published systems.

Table 7.3.2. *Performance comparison of SciNoBo RAA against best prior models on public benchmark datasets.*

Dataset	Metric	Task	SciNoBo RAA	Best Prior Model	Model Reference
DMDD-E+	F1-score	Dataset mention detection	**0.82**	0.75	SciBERT (Pan et al., 2023)
SoMeSci_test+	F1-score	Software mention detection	**0.81**	0.80	SoMeNLP (Schindler et al., 2020)
SoMeSci_test+	F1-score	License extraction	**0.96**	0.79	SoMeNLP (Schindler et al., 2020)
SoMeSci_test+	F1-score	Version extraction	0.89	**0.93**	SoMeNLP (Schindler et al., 2020)
SoMeSci_test+	F1-score	URL extraction	**0.96**	0.97	SoMeNLP (Schindler et al., 2020)
SoMeSci_test+	F1-score	Usage (reuse)	**0.89**	0.87	SoMeNLP (Schindler et al., 2020)
SoMeSci_test+	F1-score	Provenance (ownership)	0.74	**0.80**	SoMeNLP (Schindler et al., 2020)

On these public benchmark datasets, the SciNoBo RAA demonstrates performance that is consistent with or superior to established research systems previously reported in the literature. In both dataset and software mention detection, it achieves results that are on par with the strongest existing models, while showing particularly strong outcomes in metadata extraction, including license, URL, and reuse information.

These results indicate that the RAA tool can accurately identify and characterise research artefacts across different scientific domains and publication formats. Its ability to extract reliable metadata and determine whether artefacts were reused or newly created provides a robust foundation for deriving the reproducibility and reusability indicators featured in the TIER2 dashboard. In particular, the accurate detection of reuse and metadata completeness directly supports the computation of the **FAIR** and **FWRI indicators**, ensuring that the information displayed in the Pilot's visual analytics is both empirically grounded and representative of real research practices.

SciNoBo Citance Analysis (CA)

The CA tool analyses how research papers reference or discuss other works. Each citation is examined along three complementary dimensions:

- **Intent**: the purpose of the citation (e.g. reuse, extension, comparison, or contextual reference).
- **Polarity**: whether the citation expresses support, neutrality, or criticism.
- **Semantics**: which aspect of the cited work is discussed (e.g. artefacts, methods, results, or claims).

This analysis provides the foundation for the **Reproducibility Confidence Indicators** displayed in the dashboard.

Evaluation Dataset

The CA tool was evaluated using the **SciNoBo CA Dataset**, a multidisciplinary dataset containing over 1,100 manually reviewed citation sentences drawn from publications in computer science, health and bioinformatics, and social sciences and humanities.

Evaluation Results

Table 7.3.3. *Performance comparison of SciNoBo CA against baseline models.*

Task	Metric	SciNoBo CA	Best Baseline Model	Baseline Reference
Semantics	F1-score	**0.73**	0.56	GPT-4o mini
			0.62	Llama 3.1 Base
Intent	F1-score	**0.81**	0.65	GPT-4o mini
			0.70	Llama 3.1 Base
Polarity	F1-score	**0.71**	0.67	GPT-4o mini
			0.65	Llama 3.1 Base

The SciNoBo CA tool clearly surpasses the strongest baseline models across all evaluated dimensions, demonstrating a high level of reliability in interpreting how research outputs are referenced and discussed within the scientific literature. Its ability to accurately classify the intent, polarity, and semantics of citations means that it can distinguish whether a cited work is being reused, neutrally mentioned, or critically discussed, and which aspect of the work, such as its methods, results, or claims, is being referenced.

This depth of contextual understanding is essential for generating the reproducibility and confidence-related indicators used in the TIER2 dashboard. In particular, the outputs of the CA

tool directly inform the **Reproducibility Confidence Indicator (RCI)**, which reflects how the research community perceives and engages with specific artefacts or fields. By capturing not only how often a work is reused but also the tone and purpose of those references, the CA tool contributes to a more nuanced understanding of reproducibility dynamics and community trust in research outputs.

User Satisfaction

User satisfaction levels were consistently positive throughout the Pilot. The overall ease of use rating of 4.2/5.0 indicates good design. Participants expressed satisfaction with classification and filtering capabilities, which enable the comparative analysis they value most. The dashboard's visualization options (distribution charts, detailed lists, rankings) received positive ratings, indicating that users can effectively interact with and interpret the dashboard's outputs. The absence of additional feature requests for detailed views suggests the current functionality meets user needs. Participants' willingness to provide detailed, constructive feedback and their engagement in rating numerous features across multiple dimensions reflected genuine interest and investment in the dashboard's development.

7.4. Discussion

Tech Evaluation Outcomes and Readiness

The evaluation results confirm that both SciNoBo components are technically mature and suitable for integration into the TIER2 Reproducibility Monitoring Dashboard.

- **SciNoBo RAA** demonstrated high accuracy in identifying and describing research artefacts, providing reliable data for measuring reuse, provenance, and FAIRness across projects, organisations, and scientific domains.
- **SciNoBo CA** proved effective in interpreting citation contexts, offering evidence on how research is received and reused within the community, and supporting indicators related to reproducibility confidence and scholarly perception.

Together, these validated components establish a robust analytical foundation for the TIER2 Pilot, ensuring that the dashboard’s indicators are grounded in transparent, data-driven analysis and can effectively support reproducibility monitoring and policy development at multiple levels.

Reproducibility Monitoring Dashboard

The Reproducibility Monitoring Dashboard addresses critical needs for transparency and accountability in research.

Funders represent a key beneficiary group, as evidenced by the participation of 13 RFO representatives in our initial workshop and continued engagement throughout the Pilot. The dashboard enables funders to:

- Monitor reproducibility practices across their funded project portfolios systematically

- Evaluate resource allocation effectiveness by tracking the availability and documentation quality of research artifacts
- Make evidence-based decisions about funding priorities and policy interventions

RPOs also benefit from the dashboard's capacity to:

- Provide evidence for institutional reporting and identify areas for improvement in reproducibility practices
- Compare reproducibility practices with other institutions or disciplinary standards

Our iterative development process, guided by structured use cases and stakeholder input, produced dashboard templates that can be tailored to different organizational contexts and role-specific needs. This flexibility allows stakeholders to focus on metrics most relevant to their strategic objectives, whether monitoring dataset availability, software documentation quality or methodology transparency, to cite some examples. The purpose of the dashboard is to transform reproducibility monitoring from an abstract concept into actionable insights. Stakeholders can identify specific optimization opportunities, such as projects requiring enhanced documentation or research areas where reproducibility practices remain below benchmark levels.A valuable suggestion emerged regarding long-term reusability assessment. Participants recommended implementing a mechanism to track recent reuse patterns rather than relying solely on cumulative reuse metrics. This approach would help account for research outputs that may become obsolete or less relevant over time due to rapid technological advancements or evolving methodological standards. By focusing on temporal patterns of reuse, the dashboard could provide a more accurate assessment of an artifact's current relevance and ongoing utility to the research community. Through collaborative engagement with both RFOs and RPOs, we established a shared framework for reproducibility monitoring that bridges organizational perspectives and enables coordinated efforts to enhance research transparency.

Future Steps

Work on various disciplines & research areas could reveal different perspectives and requirements enhancing our efforts in reproducibility. Specifically, additional artifact types depending on the discipline under study or metadata may be of interest to the reproducibility process. Moreover, further exploration for new metrics that could improve the assessment of reproducibility is needed providing a comprehensive analysis. Input on ways to better address and meet stakeholder needs will streamline the internal processes and strengthen the dashboard utility. Lastly, we are eager to identify priority case studies or fields that should be considered for future analysis.

7.5.References

Du, C., Cohoon, J., Lopez, P., & Howison, J. (2021). *Softcite dataset: A dataset of software mentions in biomedical and economic research publications*. Journal of the Association for Information Science and Technology, 72(7), 870–884. https://doi.org/10.1002/asi.24454

Heddes, J., Meerdink, P., Pieters, M., & Marx, M. (2021). *The automatic detection of dataset names in scientific articles*. Data, 6(8), 84. https://doi.org/10.3390/data6080084

Jain, S., Van Zuylen, M., Hajishirzi, H., & Beltagy, I. (2020). *SciREX: A challenge dataset for document-level information extraction*. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (pp. 7506–7516). Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.670

Luan, Y., He, L., Ostendorf, M., & Hajishirzi, H. (2018). *Multi-task identification of entities, relations, and coreference for scientific knowledge graph construction*. In Proceedings of the Conference on Empirical Methods in Natural Language Processing (pp. 3219–3232). Association for Computational Linguistics. https://doi.org/10.18653/v1/D18-1360

Pan, H., Zhang, Q., Dragut, E., Caragea, C., & Latecki, L. J. (2023). *DMDD: A large-scale dataset for dataset mentions detection*. Transactions of the Association for Computational Linguistics, 11, 1132–1146. https://doi.org/10.1162/tacl_a_00592

Schindler, D., Bensmann, F., Dietze, S., & Krüger, F. (2021). *SoMeSci – A 5 star open data gold standard knowledge graph of software mentions in scientific articles*. In Proceedings of the 30th ACM International Conference on Information & Knowledge Management (pp. 4574–4583). ACM. https://doi.org/10.1145/3459637.3482017

Schindler, D., Zapilko, B., & Krüger, F. (2020). *Investigating software usage in the social sciences: A knowledge graph approach*. In A. Harth et al. (Eds.), The Semantic Web (Lecture Notes in Computer Science, Vol. 12123, pp. 271–286). Springer. https://doi.org/10.1007/978-3-030-49461-2_16

8. Pilot 7 - Editorial Workflows to Increase Data Sharing

Authors: Thomas Klebel, Eva Kormann, Adrian Marangoni

8.1. Introduction

Sharing of research data is an important building block of research reproducibility (Leonelli, 2018). Reproducibility is a broad term with many meanings across domains (Plesser, 2018). At a basic level, ensuring computational reproducibility refers to the act of enabling others to compute the same results, based on data and code provided by authors (Leonelli, 2018). Even though it might seem that most studies should be reproducible if data and code were available, computational reproducibility is difficult even if data and code are provided (Crüwell et al., 2023; Hardwicke et al., 2018, 2021). However, many studies share neither data nor code, rendering attempts at numerical reproduction difficult to impossible (Hardwicke et al., 2021, 2022).

Beyond enabling reproducibility, data sharing has also been linked to various further benefits, such as enabling others to re-use the data (Reinertsen et al., 2021), or an increase in citations towards the paper linked to the dataset (Colavizza et al., 2020; Piwowar & Vision, 2013). Hence, research funders and publishers are increasingly seeking to increase rates of data sharing for their funded research (e.g., National Institute of Health, n.d.; Open Research Europe, n.d.). Within published research, data sharing practices commonly crystallise in so-called Data Availability Statements (DAS). Although the exact name for these sections might differ across publishers and journals, in broad terms, journals increasingly require authors to make transparent which data have been used in the creation of the manuscript, and how they might be obtained by others (Grant & Hrynaszkiewicz, 2018; Jones et al., 2019).

Data availability statements are in principle a good way to disclose data use and how to obtain the data, and most authors comply with the requirement of providing a DAS (Federer et al., 2018). Nevertheless, stakeholders have identified a range of issues around the current practice of providing information on data in DAS. Of the issues identified by publisher representatives and other stakeholders in a session hosted by the Data policy standardisation and implementation interest group of the Research Data Alliance, two directly influenced our research design. First, stakeholders reported authors being unfamiliar with what a DAS is, and what is expected in terms of writing a good DAS. Second, stakeholders also reported that it is also very common for authors to state that data “are available upon request”, or a variant thereof (Colavizza et al., 2020; Graf et al., 2020). There are certainly legitimate reasons why data might not be made publicly available, such as restrictions around sensitive data or concerns around data privacy (e.g., Bonomi et al., 2020). However, when requesting data that is available upon request, data is often not shared (Hussey, 2023; Tedersoo et al., 2021), possibly indicating that this statement is often used as an easy option to fulfil the requirement of providing information in a DAS while the authors have no intention to actually share the data. Alternatively, authors might struggle to collect and share their data months or years after their manuscript has been published and thus decline to do so.

All this combined leads to the current situation where only few DASs include a direct link to the data in a trusted repository (Colavizza et al., 2020; Graf et al., 2020; McGuinness & Sheppard, 2021; Serghiou et al., 2021). Fewer still reference data in their manuscripts, which is increasingly considered best practice (Data Citation Synthesis Group, 2014). While some scientific disciplines make their data available, others are more reluctant when it comes to sharing critical data (Tedersoo et al., 2021). We therefore see a clear need to further improve the status quo around data sharing and in particular Data Availability Statements. Given that policies mandating data availability can increase actual data availability (Hamilton et al., 2023; Hardwicke et al., 2018), we conjecture that explaining to authors why data sharing is useful, and how it can be done might improve transparency around data availability, but also additionally prompt researchers to consider the option of immediately sharing data in a repository, therefore indirectly increasing rates of data sharing.

To test this conjecture, we evaluated whether sending researchers information about why data sharing is beneficial and how to do it alongside the regular peer review process increases rates of data sharing in trusted repositories among manuscripts resubmitted after peer review. If effective, the intervention could be implemented across journals without requiring substantial efforts at the respective publishers and journals to increase rates of data availability.

We hypothesise that the intervention will lead to an increase in the share of DASs containing a working link to a trusted repository. Second, we hypothesise that the intervention will lead to a decrease in DASs that state “data available on request”.

8.2. Methodology

Design

The study employed a multi-journal randomised controlled trial (RCT) design. The study included two parallel arms and was designed as a superiority trial, testing whether the intervention leads to improved outcomes. Manuscripts were randomly allocated to either the intervention or a control (peer review as usual) group with an allocation ratio of 1:1, stratified by journal. This minimised the impact of potential confounding factors on the journal level. The study was conducted in collaboration with Taylor & Francis among six journals in the natural and engineering sciences: International Journal of Production Research, Engineering Optimization, International Journal of Digital Earth, International Journal of Logistics, Operations & Logistics, Geomatics, Natural Hazards and Risk . All participating journals operate on single-blind peer review and require authors to provide a DAS at submission. Ethical approval was obtained from the Ethics Committee of Graz University of Technology (EK-34/2024). The protocol was preregistered on the Open Science Framework: https://doi.org/10.17605/OSF.IO/D9V47

Eligibility Criteria

Eligible manuscripts were research articles reporting results based on some type of primary data that could technically be shared. Other formats such as letters, correspondence, and reviews were excluded. Since the intervention was administered automatically, we could not conduct eligibility

checks on each manuscript. Manuscripts stating at initial submission that they share their data were still included, because the intervention could also have a regressive effect on the rate of sharing.

Contamination between groups was unlikely, since there is no regular exchange between authors submitting to the same journal. There was a slight chance that the same authors would be enrolled in the study multiple times when submitting multiple manuscripts to participating journals within the study time window. Post-hoc checks were conducted to identify corresponding authors that occur multiple times within our sample. Only the first enrolled manuscript of those authors was included for analysis to avoid overlap between and within groups.

Intervention

The intervention was developed through three co-creation workshops, with representatives from major publishers. After adoption to the stakeholder feedback, we aimed at developing an intervention that was easy to implement but still showed promise to change researcher practice. Although multiple publishers contributed to the design phase, only one publisher agreed to participate in the study. The final intervention consisted of an automated pipeline that randomized manuscripts (1:1) to intervention or control, stratified by journal. Based on the input from the stakeholders and in collaboration with colleagues from Taylor & Francis, we developed an email to be sent out to manuscript authors, presenting reasons for why data sharing is beneficial and detailing steps on how to do so. This email was adapted from the preregistered version to fit the publisher's context and was delivered to authors at the time of their first post–peer review decision. The full text of the email is provided in Appendix 1. The control group underwent the standard review process without additional input. Furthermore, and in slight deviation to the protocol, the publisher included a statement on the websites of participating journals to inform authors. The final intervention began on December 24, 2024 and is ongoing.

Outcomes

Data on all outcomes was gathered on the level of the individual manuscript, which is the unit of analysis. In deviation to the protocol, we included all outcomes, regardless of time between submission and resubmission. The median time between submission and resubmission was 95 days, with only two manuscripts exceeding the initially set cut-off of 6 months (183 days). The reason for inclusion was threefold: First, the overall sample size was still low, second, the two manuscripts were resubmitted shortly after the cut-off (186 days and 205 days after initial submission), and third, we could not conduct any sensitivity analysis with only two manuscripts.
We focused on outcomes at first resubmission rather than later stages for two reasons:
Production times (time between acceptance and publication) might differ between journals, but this is not relevant to our research question.

Our aim is to assess the direct change authors make after having received our intervention. Editors might still request changes to DAS from authors after or in conjunction with acceptance (for example, asking them to formally cite the data instead of providing a link). This could represent a downstream effect of the intervention on the editors. However, we are solely interested in the

direct effect of the intervention on author behaviour. Since the intervention is applied after the first round of peer review, assessing outcomes at first resubmission allows an immediate investigation of its effect.

The primary outcome was whether the DAS contained a functioning link to a trusted repository at first resubmission. We further included a set of secondary outcomes.

Secondary outcomes were whether the DAS (i) stated that data were available on request, (ii) stated that data were not available, (iii) contained another type of statement, or (iv) was missing entirely. Another category was added for DASs that were indeterminable, e.g., due to links being redacted to be anonymized for peer review. These were excluded from the final analysis. Secondary (intermediate) outcomes included the DAS at initial submission and the following times (in days): (i) submission to AE assignment, (ii) AE to first decision, (iii) first decision to first resubmission and (iv) DAS text at submission (coded according to the final outcome categories).

Data Collection for Final Outcome

Final outcomes were assessed by two outcome assessors who were partially blinded to allocation. Disagreements were resolved by two coders discussing discrepancies to reach a consensus, and if necessary, a third assessor adjudicating.

The outcome coding was done as binary yes/no variables (e.g., on request yes/no). If the DAS contained multiple different statements, e.g., that parts of the data are in a repository and parts available on request, both these outcomes were coded as yes.

Assessment of whether a DAS contained a functioning repository link followed a stepwise procedure: (i) verification that the DAS included a link, (ii) testing whether the link resolved within 30 seconds (checked twice if necessary), (iii) confirming that the link directed to a trusted repository indexed in re3data.org, and (iv) checking that the repository page provided accessible files or instructions for access (in case it is under embargo). The outcome of this check was retained in a separate variable for further exploratory analyses. Only if all criteria were met was the outcome coded positively.

Data available on request was defined as authors stating that data were available from them on request. Variants coded under this category included: data available upon/on request, data available on reasonable request, data available from the author by email, and other forms of availability by request (sourced from Graf et al. (2020)).

The data was extracted from the manuscript system and entered into Excel. We used restricted fields for data entry with pre-specified categories.

Statistical Analysis

All outcomes were analysed using Bayesian regression models. Although the sample included manuscripts from multiple journals, we performed a pooled analysis, as estimating varying intercepts and slopes was not practicable with the small number of groups.

Given the interim state of our sample, we did not analyse time outcomes. For binary outcomes, whether the DAS contained a trusted repository link or stated “data available on request”, we used logistic regression to separately assess primary and secondary outcomes. The main model for the propensity of DAS linking to a trusted repository was of the following form:

$$trustedrepo_{post-intervention} = trustedrepo_{pre-intervention} + intervention$$

All included variables were binary variables. Including the measurement pre-intervention improved precision for the estimate for intervention.

We used the following priors:

- student_t(3, 0, 2.5) for the intercept (default prior in brms)
- student_t(3, 0, 2.5) for the beta coefficients

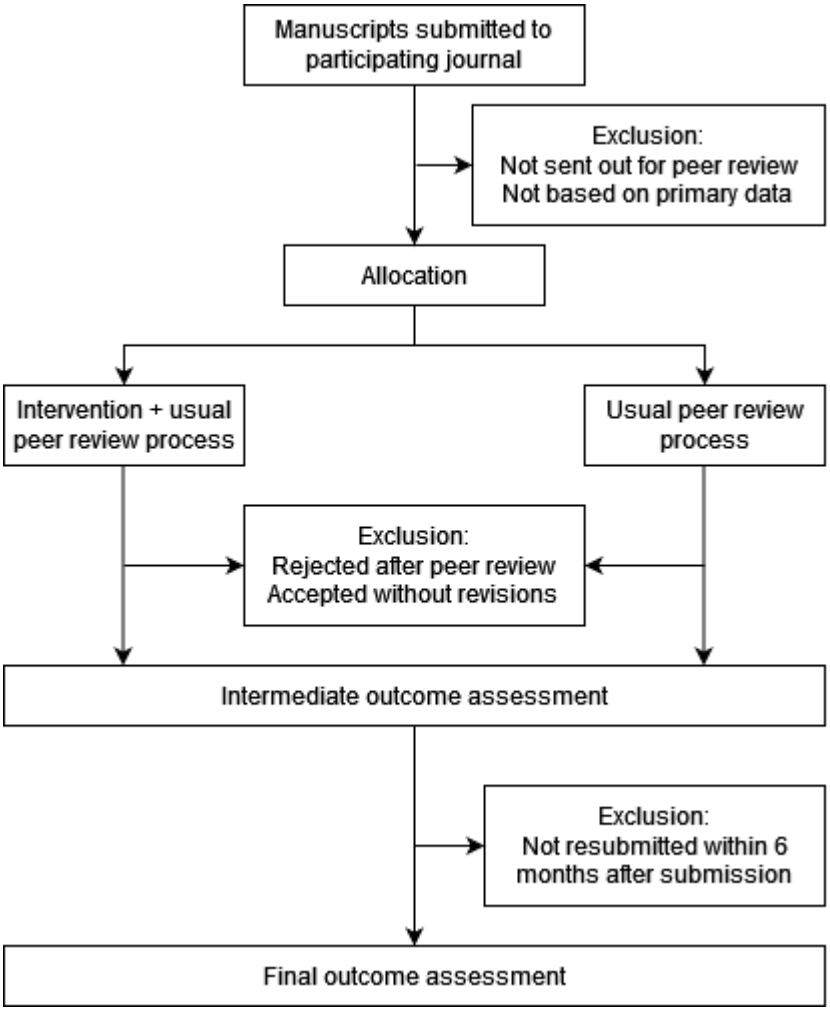

Figure 8.2.1: *Flow of manuscripts within Randomised Controlled Trial*

8.3. Results

The intervention is still ongoing, to ensure we can reach the target sample size of 600. Reasons for the delay include the enrollment starting later than anticipated and running time delays between submission and resubmission being slightly larger than initially assumed. We here provide a preliminary analysis of N = 231 manuscripts (n = 112 in the intervention, n = 119 in the control group) resubmitted by September 30, 2025.

Because of the small sample size, our estimates are still very variable. Below, we investigate (a) the absolute and (b) the relative increase in the probability of authors to make their data available in a trusted repository, comparing intervention with control group. Our preliminary estimate for the intervention's effect is 1.9% (median of posterior), with a credible interval (95%) of [-3.3%, 8.2%]. 50% of the posterior mass is between 0.0% and 3.9% (see also Figure 8.3.1).

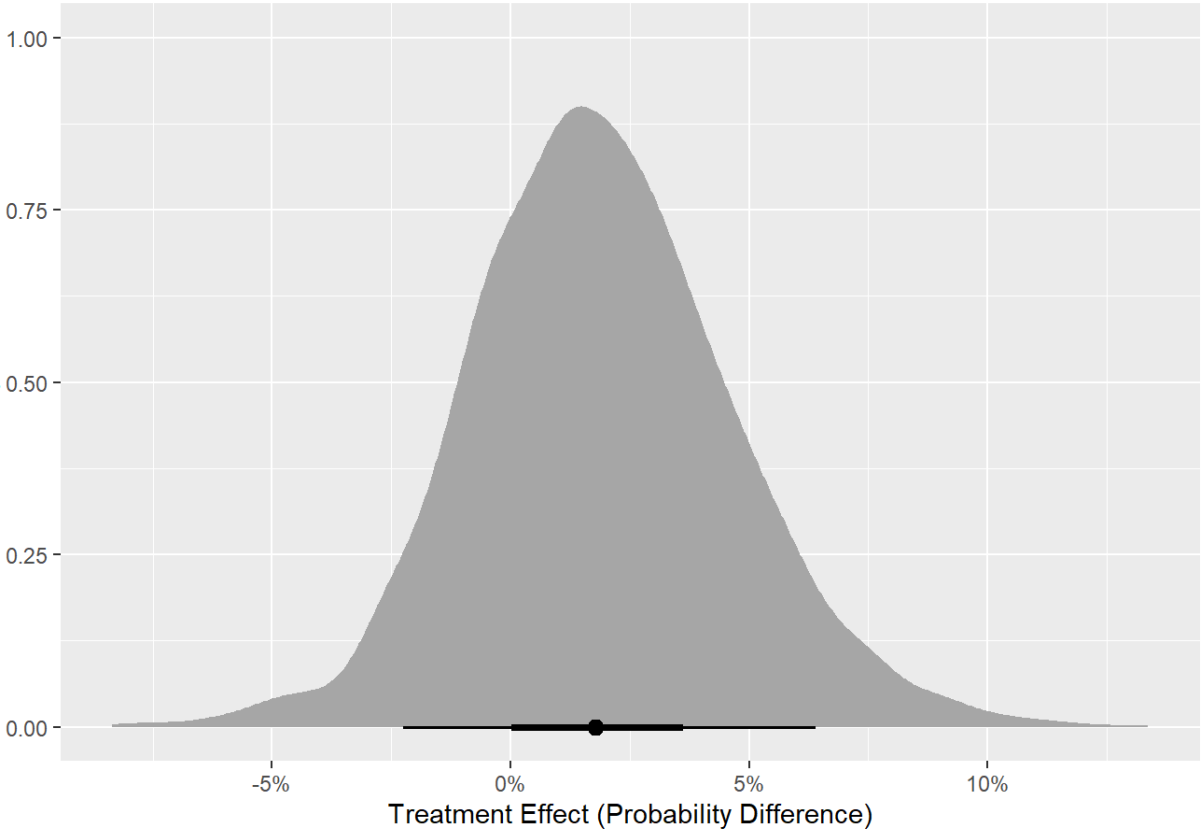

Figure 8.3.1. *Treatment effect (posterior) for absolute difference in share of authors providing data via a trusted repository.*

Overall, the share of authors providing their data via a trusted repository is very low: 3.9% in the control group. Subsequently, the estimated relative increase in the share of authors using a trusted repository due to our intervention is relatively large: the increase we observe is 47.8% (median of posterior), with a 90% credible interval of [-38.6%, 284.6%].

To summarise, although our results point in the direction of a small but positive effect, we require a larger sample size to estimate the effect with sufficient precision.

Investigating the intervention's effect on the rate of authors declaring that their data is available on request, our estimate is close to zero, with larger uncertainty than for the effect on the main outcome: 0.7% [95%-CI: -6.8%, 8.2%]. We would interpret this as the intervention having no substantive effect on the share of authors stating that data is available on request. This finding that the rate of authors sharing data in a repository increases, while at the same time the rate of authors declaring data to be available on request does so as well, is somewhat contradictory. However, this can be explained by the fact that we observed other "states" of data sharing as well: while the majority of DAS state that data is available on request, some stated that data could not be made available (due to privacy reasons, because the manuscript did not rely on data – for example in the case of theoretical mathematical models, and similar queries). In addition, some manuscripts were missing a DAS altogether. Cases where authors had redacted links to data shared in repositories to ensure blinding during peer review (despite the journals operating on single-blind peer review) were removed from the study since they could not be properly assessed.Figure 8.[OBJ] depicts the changes from before to after the intervention within the intervention group.

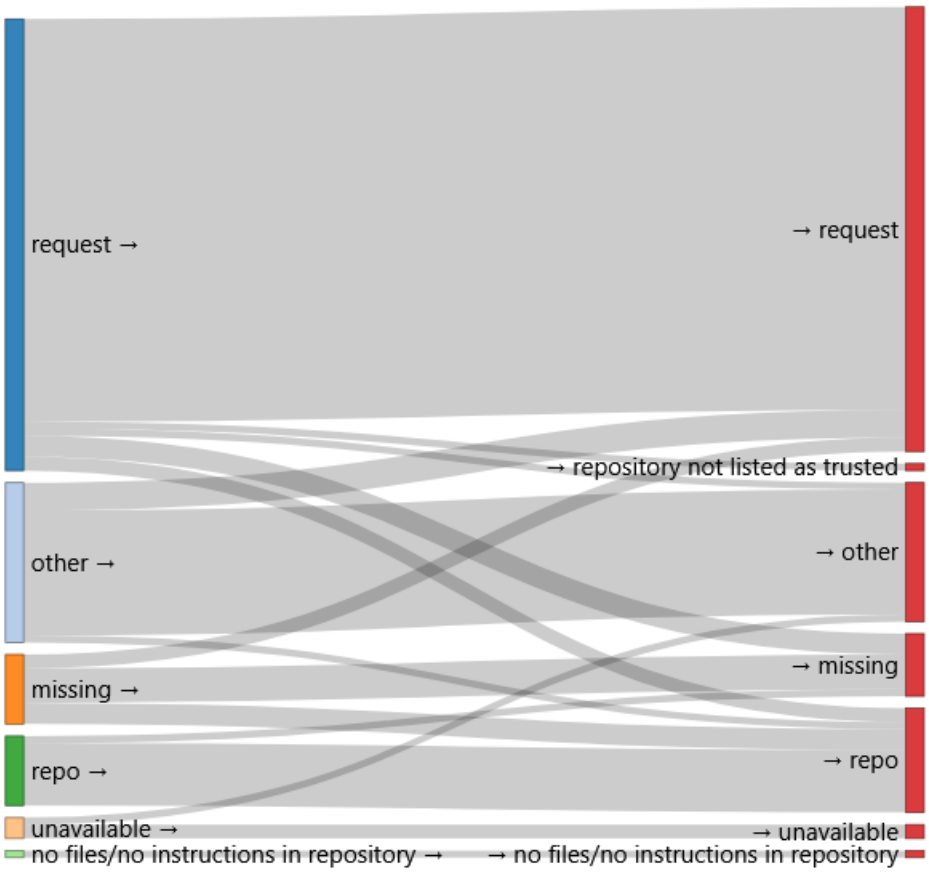

Figure 8.3.2: *Changes in data availability state from before and after the intervention within the intervention group. "other" includes statements such as "data contained in manuscript", or "no data*

used. "missing" denotes manuscripts with no DAS. "unavailable" refers to data not shared due to privacy or commercial reasons.

Inspecting Figure 8.3.2, we can see that there are cases where the DAS becomes more "open" – the rate of data shared in a repository increases. However, this might be due to pure chance, or due to other reasons. For example, in one case (in the control group) a reviewer requested the authors to make the data available (independently of our intervention, since reviewers are not aware of our intervention which only happens after the first round of reviews has been written). Ruling out such other events and determining the *causal* effect of the intervention is the reason for conducting an RCT.

Comparing the changes in Figure 8.3.2 with those observed in the control group, we hypothesised that our intervention might have an additional effect on changes in the DAS text. Certain authors seemed to align their DAS closer with the overall journal policy on resubmission. To substantiate the hypothesis, we investigated whether our intervention led to an increase in changes in the "state" of the DAS. Regression the intervention assignment (yes/no) on a dichotomous variable of whether the state of the DAS changed (for example, from "missing" to "on request") yields an effect estimate (median posterior) of 5.1% [95%-CI -3.6%, 14.1%]. There is thus weak evidence that the intervention increases the likelihood of authors making changes to their DAS.

8.4. Discussion

This randomized controlled trial represents one of the first empirical evaluations of a publisher-level intervention designed to increase data sharing rates among academic authors. Although enrolment continues toward our target sample size of 600 manuscripts, preliminary findings from 231 manuscripts provide initial insights into the intervention's effects and the broader challenges of implementing reproducibility-enhancing practices in scholarly publishing.

Summary of Preliminary Findings

Our interim analysis reveals a modest positive effect of the intervention on data sharing in trusted repositories. The intervention group showed a 1.9 percentage point increase in manuscripts with Data Availability Statements containing working links to trusted repositories, though the 95% credible interval [-3.3%, 8.2%] reflects substantial uncertainty given the current sample size. The baseline rate of data sharing in trusted repositories was remarkably low at 3.9% in the control group, substantially lower than rates reported in some previous studies (Colavizza et al., 2020) but consistent with findings from certain disciplinary contexts (Hardwicke et al., 2022). This low baseline means that even modest absolute increases translate into relatively large relative effects—our median estimate represents a 47.8% relative increase in repository sharing, albeit with wide credibility intervals.

The intervention did not substantially affect the rate of "data available on request" statements (0.7% difference, 95%-CI: [-6.8%, 8.2%]). However, approximately 5.1% more manuscripts in the intervention group showed changes in their DAS "state" between submission and resubmission [-3.6%, 14.1%], suggesting the intervention may prompt authors to reconsider their data sharing practices even when they do not ultimately deposit data in repositories.

Implications for Reproducibility

These preliminary findings situate our intervention within the broader landscape of publisher policies aimed at improving data sharing. Previous research has demonstrated that mandatory data sharing policies can increase data availability (Hardwicke et al., 2018), but enforcement and compliance remain persistent challenges. Our intervention represents a "light-touch" approach—providing authors with information and encouragement rather than strict requirements—that can complement existing policies without requiring substantial changes to editorial workflows.

The modest effect size aligns with expectations for an informational intervention in the context of weak baseline sharing norms. When data sharing is the exception rather than the norm within a research community, individual-level interventions face structural barriers related to training, infrastructure, and disciplinary culture. Our results suggest that while such interventions may move some authors toward better practices, they cannot single-handedly transform data sharing culture in fields where it remains uncommon.

The intervention's simplicity and low implementation burden make it feasible for adoption across diverse publisher contexts. Unlike more resource-intensive approaches requiring dedicated data editors or technical infrastructure, automated email delivery integrates readily into existing manuscript management systems. However, publishers should calibrate expectations accordingly. The intervention provides practical guidance to authors who may be unfamiliar with data repositories or best practices, addressing the information gap identified in stakeholder consultations. Yet it does not fundamentally alter the incentive structures or enforcement mechanisms that shape author behaviour. For publishers committed to achieving high rates of immediate data sharing, our findings underscore that stricter policies with active enforcement—such as mandatory deposition in approved repositories prior to acceptance—would likely be necessary.

Reflection on the Pilot Process

The development and implementation of this Pilot exemplified productive collaboration between researchers and publishing stakeholders while also illuminating practical barriers to conducting implementation research in scholarly publishing. The initial design phase benefited substantially from established relationships with publishers through the TIER2 consortium and early stakeholder workshops. Three co-creation sessions with representatives from major publishers generated valuable input on intervention design, emphasizing the need for solutions that would be "light touch" and not impose substantial additional work on editorial staff.

However, the transition from design to implementation revealed significant organizational complexities within publishing houses. Although engagement during intervention development was high, ultimately only Taylor & Francis committed to implementation within the randomized trial. This limited uptake stemmed from several factors that prospective implementation researchers should anticipate: capacity constraints from other ongoing initiatives, technical transitions in manuscript management platforms, and the need for coordination across multiple internal departments—open science teams, legal counsel, email template specialists, and backend technical staff.

These implementation challenges highlight an important consideration for future work: even interventions designed with parsimony and ease of adoption in mind may face organizational

barriers when deployed in complex institutional settings. The rigorous evaluation requirements of a randomized controlled trial, though crucial for generating credible evidence, can paradoxically impede adoption during initial testing phases.

The extended timeline from initial conception to interim results also merits reflection. Each phase required substantial time, and the full project arc demanded nearly the entire three-year duration. Future implementation research of this kind would benefit from realistic time allocation that accounts for manuscript resubmission and acceptance timelines that vary considerably across journals and disciplines.

Despite these challenges, stakeholder engagement remained strong throughout the Pilot, and the established stakeholder group provides a foundation for continued collaboration beyond the TIER2 project.

Strengths and Limitations

The study's primary strength lies in its methodological rigor. The preregistered randomized controlled trial design, coupled with detailed protocols for outcome assessment and blinded coding procedures, ensures high internal validity and minimizes bias in effect estimation. The intervention itself is clearly specified and readily transferable, enabling replication and adaptation by other publishers.

However, several limitations constrain our current conclusions. Most significantly, enrolment continues toward the target sample size, and the present interim analysis includes only 231 of the planned 600 manuscripts. This limited sample size results in substantial uncertainty around our effect estimates, as reflected in the wide credible intervals.

Even assuming our median effect estimate holds with increased sample size, the intervention is unlikely to produce radical changes in author behaviour. An approximately 2 percentage point absolute increase in repository sharing—while meaningful given the low baseline—would not transform data sharing culture in participating fields. This modest effect reflects the intervention's nature as an informational nudge rather than a structural change to incentives or requirements.

The study's restriction to natural and engineering science journals may also limit generalizability. Data sharing norms and infrastructure vary substantially across disciplines (Tedersoo et al., 2021), and fields with more established data sharing cultures might respond differently to the intervention. Additionally, all participating journals already required Data Availability Statements and operated under "share upon request" policies, meaning our findings specifically address how to encourage immediate repository sharing within that policy context rather than evaluating effects on journals without any data sharing requirements.

Future Directions

Data collection will continue until we reach the target sample size of 600 resubmitted manuscripts, enabling more precise effect estimation and assessment of heterogeneity across journals. Beyond completing the present study, our findings point toward several complementary research directions: investigating whether similar informational interventions prove more effective in disciplines with stronger existing data sharing norms; examining longer-term outcomes to determine whether the intervention produces durable practice changes or only affects individual

manuscripts; and testing enhanced versions that combine information with stronger incentives to identify the minimum policy stringency needed to achieve meaningful shifts in data sharing behaviour. Our stakeholder engagement suggests appetite within publishing communities for evidence-based guidance on these policy decisions, highlighting opportunities for future research examining trade-offs between different approaches to promoting data sharing.

Recommendations

For publishers operating under "share upon request" data policies, the automated email intervention represents a relatively straightforward mechanism for increasing rates of immediate repository sharing. Implementation requires minimal technical infrastructure and operates without ongoing manual effort once established. Publishers should view this as one component of a broader data sharing strategy rather than a comprehensive solution.

However, publishers should calibrate expectations appropriately. Our preliminary findings suggest the intervention may increase repository sharing by approximately 2 percentage points above baseline rates, a meaningful but modest improvement. Publishers seeking more substantial changes would need to implement stricter policies with active enforcement, such as mandatory repository deposition as a condition of acceptance, designated data editors to review Data Availability Statements, or restrictions on "available upon request" language.

The choice between "light-touch" interventions like ours and more stringent policy approaches depends on publishers' goals, resources, and stakeholder considerations. Our findings demonstrate that informational support alone cannot overcome structural barriers and weak incentives for data sharing in fields where it remains uncommon. Publishers committed to reproducibility as a core value may ultimately need to move beyond encouragement toward requirements, accepting that such transitions may require investment in infrastructure and editorial capacity alongside careful change management.

Conclusion

This ongoing randomized controlled trial provides preliminary evidence that a simple, automated intervention can modestly increase rates of data sharing in trusted repositories among authors submitting to academic journals. While our interim results require confirmation through completion of the full study, they suggest that providing authors with targeted information about the benefits and mechanics of data repository sharing represents a feasible, low-cost approach for publishers seeking to strengthen compliance with data sharing policies. However, the modest effect sizes observed underscore that informational interventions alone cannot transform data sharing norms in research communities where repository sharing remains exceptional. Publishers, journals, and scientific communities aspiring to comprehensive data availability must consider stronger policy measures alongside the practical supports this intervention provides. The collaborative development process, despite implementation challenges, demonstrates the value of co-creating evidence-based interventions with publishing stakeholders and establishes a foundation for continued work to embed reproducibility practices within scholarly communication infrastructure.

8.5.References

Bonomi, L., Huang, Y., & Ohno-Machado, L. (2020). Privacy challenges and research opportunities for genomic data sharing. *Nature Genetics*, *52*(7), 646–654. https://doi.org/10.1038/s41588-020-0651-0

Colavizza, G., Hrynaszkiewicz, I., Staden, I., Whitaker, K., & McGillivray, B. (2020). The citation advantage of linking publications to research data. *PLOS ONE*, *15*(4), e0230416. https://doi.org/10.1371/journal.pone.0230416

Crüwell, S., Apthorp, D., Baker, B. J., Colling, L., Elson, M., Geiger, S. J., Lobentanzer, S., Monéger, J., Patterson, A., Schwarzkopf, D. S., Zaneva, M., & Brown, N. J. L. (2023). What's in a Badge? A Computational Reproducibility Investigation of the Open Data Badge Policy in One Issue of Psychological Science. *Psychological Science*, *34*(4), 512–522. https://doi.org/10.1177/09567976221140828

Danchev, V., Min, Y., Borghi, J., Baiocchi, M., & Ioannidis, J. P. A. (2021). Evaluation of Data Sharing After Implementation of the International Committee of Medical Journal Editors Data Sharing Statement Requirement. *JAMA Network Open*, *4*(1), e2033972. https://doi.org/10.1001/jamanetworkopen.2020.33972

Data Citation Synthesis Group. (2014). *Joint Declaration of Data Citation Principles*. Force11. https://doi.org/10.25490/A97F-EGYK

Federer, L. M., Belter, C. W., Joubert, D. J., Livinski, A., Lu, Y.-L., Snyders, L. N., & Thompson, H. (2018). Data sharing in PLOS ONE: An analysis of Data Availability Statements. *PLOS ONE*, *13*(5), e0194768. https://doi.org/10.1371/journal.pone.0194768

Graf, C., Flanagan, D., Wylie, L., & Silver, D. (2020). The Open Data Challenge: An Analysis of 124,000 Data Availability Statements and an Ironic Lesson about Data Management Plans. *Data Intelligence*, *2*(4), 554–568. https://doi.org/10.1162/dint_a_00061

Grant, R., & Hrynaszkiewicz, I. (2018). The Impact on Authors and Editors of Introducing Data Availability Statements at Nature Journals. *International Journal of Digital Curation*, *13*(1), 195–203. https://doi.org/10.2218/ijdc.v13i1.614

Hamilton, D. G., Hong, K., Fraser, H., Rowhani-Farid, A., Fidler, F., & Page, M. J. (2023). Prevalence and predictors of data and code sharing in the medical and health sciences: Systematic review with meta-analysis of individual participant data. *BMJ*, *382*, e075767. https://doi.org/10.1136/bmj-2023-075767

Hardwicke, T. E., Bohn, M., MacDonald, K., Hembacher, E., Nuijten, M. B., Peloquin, B. N., deMayo, B. E., Long, B., Yoon, E. J., & Frank, M. C. (2021). Analytic reproducibility in articles receiving open data badges at the journal Psychological Science: An observational study. *Royal Society Open Science*, *8*(1), 201494. https://doi.org/10.1098/rsos.201494

Hardwicke, T. E., Mathur, M. B., MacDonald, K., Nilsonne, G., Banks, G. C., Kidwell, M. C., Hofelich Mohr, A., Clayton, E., Yoon, E. J., Henry Tessler, M., Lenne, R. L., Altman, S., Long, B., & Frank, M. C. (2018). Data availability, reusability, and analytic reproducibility: Evaluating the impact of a mandatory open data policy at the journal Cognition. *Royal Society Open Science*, *5*(8), 180448. https://doi.org/10.1098/rsos.180448

Hardwicke, T. E., Thibault, R. T., Kosie, J. E., Wallach, J. D., Kidwell, M. C., & Ioannidis, J. P. A. (2022). Estimating the Prevalence of Transparency and Reproducibility-Related Research Practices in Psychology (2014–2017). *Perspectives on Psychological Science*, *17*(1), 239–251. https://doi.org/10.1177/1745691620979806

Hussey, I. (2023). *Data is not available upon request*. https://doi.org/10.31234/osf.io/jbu9r

Jones, L., Grant, R., & Hrynaszkiewicz, I. (2019). Implementing publisher policies that inform, support and encourage authors to share data: Two case studies. *Insights the UKSG Journal*, *32*, 11. https://doi.org/10.1629/uksg.463

Leonelli, S. (2018). Rethinking Reproducibility as a Criterion for Research Quality. In L. Fiorito, S. Scheall, & C. E. Suprinyak (Eds.), *Research in the History of Economic Thought and Methodology* (Vol. 36, pp. 129–146). Emerald Publishing Limited. https://doi.org/10.1108/S0743-41542018000036B009

McGuinness, L. A., & Sheppard, A. L. (2021). A descriptive analysis of the data availability statements accompanying medRxiv preprints and a comparison with their published counterparts. *PLOS ONE*, *16*(5), e0250887. https://doi.org/10.1371/journal.pone.0250887

National Institute of Health. (n.d.). *NOT-OD-21-013: Final NIH Policy for Data Management and Sharing*. Retrieved 10 July 2024, from https://grants.nih.gov/grants/guide/notice-files/NOT-OD-21-013.html

Open Research Europe. (n.d.). *Policies | Open Research Europe*. Retrieved 10 July 2024, from https://open-research-europe.ec.europa.eu/about/policies#dataavail

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

Plesser, H. E. (2018). Reproducibility vs. Replicability: A Brief History of a Confused Terminology. *Frontiers in Neuroinformatics*, *11*. https://www.frontiersin.org/articles/10.3389/fninf.2017.00076

R Core Team. (2024). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. https://www.R-project.org/

Reinertsen, I., Collins, D. L., & Drouin, S. (2021). The Essential Role of Open Data and Software for the Future of Ultrasound-Based Neuronavigation. *Frontiers in Oncology*, *10*, 619274. https://doi.org/10.3389/fonc.2020.619274

Rowhani-Farid A, Aldcroft A, & Barnett A G. (2020). Did awarding badges increase data sharing in BMJ Open? A randomized controlled trial. *Royal Society Open Science*, *7*(3), 191818. https://doi.org/10.1098/rsos.191818

Serghiou, S., Contopoulos-Ioannidis, D. G., Boyack, K. W., Riedel, N., Wallach, J. D., & Ioannidis, J. P. A. (2021). Assessment of transparency indicators across the biomedical literature: How open is open? *PLOS Biology*, *19*(3), e3001107. https://doi.org/10.1371/journal.pbio.3001107

Tedersoo, L., Küngas, R., Oras, E., Köster, K., Eenma, H., Leijen, Ä., Pedaste, M., Raju, M., Astapova, A., Lukner, H., Kogermann, K., & Sepp, T. (2021). Data sharing practices and data availability upon request differ across scientific disciplines. *Scientific Data*, *8*(1), 192.

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), 160018. https://doi.org/10.1038/sdata.2016.18

9. Pilot 8 - The Editorial Reference Handbook

Authors: Allyson Lister, Susanna-Assunta Sansone

9.1.Introduction

In May 2023, a publishers' workshop indicated that strengthening journals' data policies and training in-house editorial staff were among the key priorities to improve the availability of data underpinning publications and foster good practices for sharing data, ultimately advancing open research.

TIER2 Pilot 8 was created as a result of this workshop to inform and support journals in operationalising a set of checks designed to enhance the FAIRness of research objects and promote good sharing practices. While some journals have internal guidance on promoting and enabling reproducible and FAIR data, there is little/no consensus among publishers. The resulting Educational Reference Handbook helped operationalize data checks to assist reproducibility and FAIRness, provided editors with a harmonized set of data checks, and served as advice for authors and reviewers (Lister *et al* 2025).

9.2. Methodology

Participant selection

The participants were drawn from the FAIRsharing Stakeholders Advisory Board. No ethical approval was needed, as the pilot members agreed to share the information and are listed publicly as co-authors of the Handbook, and will be co-authors of the final publication (in progress, preprint available at https://doi.org/10.17605/OSF.IO/FB9QW). The preparation and planning for this Pilot started in Dec 2023 and ended in Feb 2024; the co-creation phase from Mar to July 2024, which included the launch of the Handbook; and the intervention phase from Aug 2024 to Oct 2025.

Co-creation phase

Co-creation consisted of 2 workstreams across 9 online 1-hour sessions with pilot members. An iterative process was used to collect feedback and complete the workstreams, with offline work between calls, via google docs/sheets and email; ample time was provided for publishers' internal review and approval prior to the intervention.

First workstream: developing the core checklist and accompanying guidance
Although journals employ a variety of internal checks, their type, scope, and rigour vary considerably, and there is little consistency across policies. To address this, we reviewed 25 existing resources and initiatives relevant to our scope in order to identify the most frequently recommended elements that were also broadly applicable across disciplines and editorial contexts. Tables 9.2.1 and 9.2.2 lists those resources included in (Table 9.2.1) and excluded from (Table 9.2.2) further discussions and drafts. In addition to refining the checks, these discussions

enabled us to distinguish between those checks that depend on author or journal expertise and those that would require additional support for effective operationalisation.

Table 9.2.1: *Existing resources and initiatives considered relevant to the scope of Pilot 8 together with information on which components of those resources were included within the Handbook.*

Resource	Location	Notes relating to Inclusion
ARRIVE	https://doi.org/10.25504/FAIRsharing.t58zhj	Section 19: Protocol registration; Section 20: Data Access
F1000 checks summary	n/a	n/a
FAIR4RS	https://doi.org/10.1038/s41597-022-01710-x	Utilised F1, F2-F4, R1.1, R3. The other sections excluded as too granular.
GigaScience Minimum Standards of Reporting Checklist (was: BMC Better reporting for better research: a checklist for reproducibility)	https://academic.oup.com/gigascience/pages/Minimum_Standards_of_Reporting_Checklist (was https://doi.org/10.1186/s13742-015-0071-8)	Utilised: Resource subsection; Availability of data and materials subsection. Most other sections out of scope for this handbook.
The MDAR (Materials Design Analysis Reporting) Framework for transparent reporting in the life sciences	https://doi.org/10.1073%2Fpnas.2103238118; see also https://doi.org/10.25504/FAIRsharing.d56cdd and the MDAR checklist at https://osf.io/bj3mu/; https://www.science.org/content/blog-post/improving-reproducibility	Specifically, the MDAR checklist for authors. This is also explicitly intended for editors. Excluded MDAR's study-level protocols, study design, statistical tests, attrition. See also https://www.ncbi.nlm.nih.gov/books/NBK556610/, incl comments about how 'less is more'
Nature Portfolio Editorial Policy Checklist	https://www.nature.com/documents/nr-editorial-policy-checklist-Apr-2023-flat.pdf	Only certain sections are directly relevant (Code and data availability).
NIH Principles and Guidelines for Reporting Preclinical Research	https://grants.nih.gov/policy/reproducibility/principles-guidelines-reporting-preclinical-research.htm	Excluding replicates guidelines. Note that the NIH recommends the use of checklists during editorial processing that are then made visible to authors.
Promoting Reusable and Open Methods and Protocols (PRO-MaP)	https://doi.org/10.31219/osf.io/x85gh	Specifically, Section and Table 3 (Publishers and editors) was used in this review. No study-level design or protocols in PRO-MaP
REAPPRAISED checklist for evaluation of publication integrity	https://doi.org/10.1038/d41586-019-03959-6	Utilised: E - Ethics; A - Analysis and Methods; and D - Data duplication and reporting. Only certain sections are directly relevant
STORMS	https://www.stormsmicrobiome.org/	Section 8 (Reproducibility); 16 and 17 (Supplements and supplementary data). The majority of this checklist is too fine-grained for use within the Handbook.
TOP/COS checklist for editors/reviewers (see also	https://osf.io/55eu7 (Checklist for Authors implementing Level 1); https://osf.io/87v93 (Checklist for Editors Levels 1 and 2)	Certain sections excluded as too granular or out of scope

https://osf.io/kgnva/wiki/home/)		
RDA / CURE-FAIR	https://doi.org/10.25504/FAIRsharing.4a9e19	Utilised: Things 1, 2, 5, 6, 8 All included sections ("Things") are high-level without particular implementation details as relevant for manuscript submissions. However conceptually many "Things" align with the Handbook, and are represented accordingly in the guidance.
Developing a Research Data Policy Framework for All Journals and Publishers - Data policy standardisation and implementation IG	https://doi.org/10.5334/dsj-2020-005	Utilised: Definition of exceptions, Data repositories, Data citation, Data licensing, Data availability statements (DASs), Data formats and standards While this work is focused on journal data policies as a whole, some segments of this framework do align with the Handbook.

Table 9.2.2 *Existing resources and initiatives considered out of scope of Pilot 8 together with reasons for exclusion.*

Resource	Location	Reasons for Exclusion
AAAI Reproducibility Checklist	https://aaai.org/conference/aaai/aaai-23/reproducibility-checklist/	There is some overlap with this work, but mostly too narrow in scope.
Checklist for an Open Research Action Plan	https://www.ukrn.org/2021/11/03/open-research-action-plan/	Not related to this work at all; discounted
CONSORT	https://doi.org/10.25504/FAIRsharing.gr06tm	Too narrow in scope
FAIR software checklist and tool	https://ardc.edu.au/article/new-self-assessment-tool-to-promote-fair-research-software/	An implementation of FAIR4RS that we already include
GCCP	-	Too narrow in scope
GIVIMP, SciRAP GD211	-	Too narrow in scope
MICCAI Reproducibility Checklist	https://miccai2021.org/files/downloads/MICCAI2021-Reproducibility-Checklist.pdf	Reproducibility checklist for authors, to then be used by reviewers etc. Too narrow in scope
PRISMA	https://doi.org/10.25504/FAIRsharing.gp3r4n	Too narrow in scope
Reliability and reproducibility checklist for molecular dynamics simulations	https://www.nature.com/articles/s42003-023-04653-0	Too narrow in scope
Reproducible Research Publication Workflow: A Canonical Workflow Framework and FAIR Digital Object	https://doi.org/10.1162/dint_a_00133	This paper was read and assessed, but was not directly relevant to this review. Describes an example canonical workflow for publishers to follow.

Approach to Quality Research Output		

Second workstream: designing a generalised flowchart to situate the checks within an idealised manuscript submission workflow and associated staff roles.

The second workstream focused on understanding how internal processes operate in practice and on identifying the most appropriate roles and workflow stages for each checklist element. The workshop sessions were directed toward collecting information and experiences regarding when each check was likely to occur (or was already occurring, in the case of journals with existing practices), who would be responsible for carrying it out, and how it would be implemented. Through iterative discussion and refinement, this workstream culminated in the development of the flowchart component, which maps each checklist element to both a role that gained a broad consensus within the workstream members and a specific stage of the manuscript submission workflow (Taylor-Grant *et al* 2025).

Intervention phase.

The intervention stage involved three groups of participants: (i) intervention, (ii) positive controls and (iii) advisors. The latter contributed to the identification of the participants in the intervention group, and to the definition of the milestones and evaluation metrics. The publishers' and journals' representatives self-organised in groups taking up one or more of these roles. The intervention group comprised publishers and journals that applied the Handbook to evaluate manuscripts submitted during the intervention period, while those that had already implemented the Handbook's checks formed the positive control group.
We organized a dedicated online session for the intervention, also to identify in-house editors in journals willing to participate, as they differed from the members of the co-creation. The intervention phase was run in the 3 stages, described below. During stage 1 and 2, we provided multiple means for contacting and recording data, including brief interviews, email exchanges, surveys, and forms.

1. Preparation. Engagement with the pilot participants to understand what may need to change or improve to successfully implement the Handbook in terms of in-house ***capability*** (e.g., needing more knowledge about the Handbook), ***opportunity*** (e.g., needing support to apply the checks), and ***motivation*** (e.g., needing to prioritise the checks). We also discussed the output of an audit we have conducted, via FAIRsharing, on journal or publisher ***data policies***. In addition, we identified the type of data collected during the implementation phase; for example, the number of checks added to the participants' current practice, the time taken to undertake them, the response of the authors, and the overall impact on the manuscript submission workflow.
2. Implementation. Pilot participants used the Handbook in their manuscript submission workflows, over a period of up to 6 months, and collected data. We provided any additional support required, such as providing access to tailored training materials or technical expertise (e.g., to support those responsible for adding the checks to current workflows).

3. Evaluation. We collated, analysed and discussed the collected data to understand the experiences of those who participated in the preparation and the implementation, and the type of support we provided.

During the evaluation stage, questionnaires for both the positive control and intervention groups were finalised, distributed, and completed by participants. The positive control questionnaire was co-designed with participants to capture their motivations, enablers, barriers, and workflow modifications. Metrics assessed included the proportion of portfolios implementing checks, manuscript compliance rates, use of domain-specific repositories and formats, time required to complete checks, and the impact on editorial workflows and author correspondence. Participants were also invited to provide qualitative feedback on each checklist item, addressing implementation challenges, lessons learned, and recommendations for future adopters. The questionnaire was released to participants in April 2025.

Using a similar methodology, the intervention exit questionnaire was co-created with the intervention group. The questionnaire was organised into five Handbook-related sections: (i) administrative (e.g., journal details, checks employed), (ii) overall evaluation (e.g., motivations, enablers and barriers, modifications made), (iii) process outcomes (e.g., number of manuscripts evaluated, impact on submission times), (iv) FAIR-enabling outcomes (e.g., extent of sharing research objects), and (v) individual checklist elements (e.g., ease of understanding, roles and workflow positions assigned).

9.3. Results

Co-created with 14 journals and eight publishers, the Handbook establishes a shared understanding of a fundamental set of checks that help enable FAIRness, underpin reproducibility, and apply to all digital objects (e.g., datasets, code, materials) associated with a publication (Klebel and Lister 2025). The Handbook also maps these checks onto an idealised internal manuscript submission workflow. In practice, whether each check is performed—and, if so, how, when, and by whom—varies across journals, with implications for the consistency and effectiveness of outcomes.

The aims of the Handbook are to: (i) operationalise the agreed-upon checks as part of an ideal internal manuscript submission workflow; (ii) support journals in integrating the Handbook's concepts into their policies and editorial processes; (iii) assist with the implementation of open research policies; and (iv) ultimately enhance the reusability and potential reproducibility of published research.

The Handbook is structured into three interlinked components, which may be used independently or in combination to assess individual manuscripts or, more broadly, to inform the updating of journal policies and submission workflows. Each check is mapped to a specific role and position in the flowchart and, when implemented, may return one of three outcomes: Pass, Fail, or Not Applicable (N/A). Failed checks initiate corrective workflows in accordance with journal policies. In addition, checks are categorised by consideration level: core (applicable across all types of

digital objects, irrespective of research domain) or specialised (relevant only to particular research areas or types of manuscripts, e.g., domain-specific repositories, which may not be pertinent to all journals or digital objects).

In the intervention, all six participating journals implemented the Handbook to varying degrees, adapting it to their existing internal workflows. Participants began the intervention at a range of times, resulting in durations ranging from approximately two to six months. Journal submission volumes varied widely, from tens to thousands of manuscripts, with the number of manuscripts assessed during the intervention broadly correlating with intervention length and ranging from 12 to 86 per journal, for a total of 190 manuscripts.

Participants reported that the Handbook met their needs overall, noting that their journals had assumed a more active role in assessing the quality of digital objects and that policies had been strengthened as a result. Authorisation, prioritisation, and genuine commitment to, and belief in, the initiative emerged as the most common enabling internal factors. Strengthening policies and assessment processes was frequently cited as a key motivator for implementing the Handbook. The most frequently cited barriers to implementing the checks were the time required, competing editorial tasks, and variability in authors' beliefs, willingness, and skills.

All participants indicated that they intended to continue using the Handbook following the intervention. Very few authors became unresponsive during the submission process (ranging from zero to 10%) or withdrew their manuscripts (ranging from zero to three percent), and no significant impacts on turnaround time (either zero or low impact) were reported, although two journals noted increased correspondence with authors and higher rates of return for revision. Participants reported that all checklist elements were relatively straightforward to implement, with the exception of element 2. This element requires verifying that all digital objects that should be included in availability statements are in fact listed. Because this necessitates an evaluation of the entire manuscript to identify potentially missing digital objects, rather than simply reviewing those already included in availability statements, participants found it more challenging to integrate into existing submission workflows.

Six organisations—three publishers and three journals—participated in the positive control group. Their questionnaire responses provide valuable insights into how existing internal workflows and checks align with the Handbook. Several journals and publishers had integrated the checks to improve existing workflow procedures, while others had incorporated them from the outset. All positive controls reported that the majority of submitted manuscripts required additional work to achieve compliance. Rough estimates from positive control journals indicated that between 20% and 50% of manuscripts were compliant with the Handbook upon submission.

Overall, all positive controls expressed motivation rooted in the conviction that such checks represent good publishing practices. Training, education, and persuasion of staff were identified as the most important modifications required to enable change within journals. Although time and funding constraints were recognised as barriers, internal willingness, expertise, and commitment to good practices for research objects were highlighted as key enablers for implementation. Moreover, despite the additional requirements placed on authors, all positive controls reported that authors' belief in these practices further served as an important enabling factor.

9.4. Discussion

The Handbook integrates structured checks, narrative guidance, and visual workflows to bridge the gap between policy and editorial practice. It can assist journals and publishers in two primary ways: (i) for those without internal guidance to enforce an open research policy, it provides a workflow for assessing and improving individual manuscripts; and (ii) for those with existing guidance, it offers principles that can be used to validate and enhance current methodologies. It provides a model for embedding good research practices and FAIR principles into the scientific publication process, while also exemplifying the broader cultural shift toward open and responsible practices within scholarly publishing. By offering a shared, operational resource grounded in a consensus set of small but practical checks aligned with journal roles and workflows, the Handbook addresses a critical gap and supports scalable adoption.

The processes of co-creation and intervention represent significant collaborative and practical steps toward the operationalisation of good open research practices. The Handbook has also demonstrated potential for informing improvements in internal editorial workflows and for harmonising journal policies. Although primarily aimed at in-house editorial staff managing manuscripts, the Handbook additionally benefits reviewers, authors, and service providers by making fundamental checks and requirements transparent and accessible. The experiences of both intervention participants and positive controls demonstrate that the Handbook is sufficiently rigorous to be educational and practical, while also retaining the flexibility necessary for adoption across diverse journal contexts, tailored to local readiness and priorities.

Beyond its value as a practical resource, the creation of the Handbook was also a socio-technical initiative aimed at improving research culture, leading by example to influence and inform other publishers and journals. Its pilot use across a number of journals not only demonstrated its applicability but also highlighted areas requiring further development to ensure successful implementation. Specifically, participants identified needs in terms of in-house capability (e.g., greater knowledge of how to apply the checks), capacity (e.g., support in operationalising them), and motivation (e.g., prioritisation within existing workflows).

While the project has already engaged with a broad group of publishers and journals, there is potential to socialise the Handbook and support further adoption through relevant channels. Looking ahead, the Handbook and its community have the potential to serve as a foundation for broader initiatives in good research practices and reproducibility, policy harmonisation, and cross-publisher collaboration. Further integration with services such as FAIRsharing could enable the development of automated dashboards, policy audits, and metadata validation tools. We therefore invite relevant groups and initiatives to engage with the Handbook leadership to explore opportunities for extending and sustaining the success of this endeavour.

9.5.References

Klebel, T., & Lister, A. (2025, November 13). TIER2 D5.2 - Tools and practices for publishers. https://doi.org/10.17605/OSF.IO/S7GJV

Lister, A., Taylor-Grant, R., Cannon, M., Ahmed, R., Alfarano, G., Begum, R., Bright, J., Cadwallader, L., Cranston, I., Dunkley, L., Edmunds, S., Flammer, P., Hill, A., Hunter, C., Hyde, A., Klebel, T., Leary, A., MacCallum, C.J., McKenna, S., McNeice, K., Miorini, J., Nogoy, N., Patterson, K., Pulverer, B., Ross-Hellauer, T., Smith, A., Sonntag H., and Sansone, S. (2025, October 27). “Supporting FAIR Practices In Scholarly Publishing with the Editorial Reference Handbook” Preprint: https://doi.org/10.31222/osf.io/9vujt_v2

Taylor-Grant, R., Cannon, M., Lister, A., & Sansone, S.-A. Making reproducibility a reality by 2035? Enabling publisher collaboration for enhanced data policy enforcement. *International Journal of Digital Curation*, 19(1), 10 (2025). https://doi.org/10.2218/ijdc.v19i1.1064

10. Discussion

10.1. Overall reflection on the Pilots and the tools

Taken together, the eight TIER2 Pilots show that there is no single route to improving reproducibility and that many different problems need different solutions for different stakeholder groups. Instead, they illustrate just how multifaceted the challenge is. Each Pilot approached reproducibility from a different angle—whether by designing tools that support individual researchers in planning their work more transparently, experimenting with computational workflows to make analyses traceable, or engaging with funders and journals to shape and improve policy on reproducibility. What connects them is that each intervention/Pilot was grounded in some empirical understanding of real research practice, rather than in abstract ideals of how science *should* work.

One of the strengths of the Pilots is that they were not simply conceptual exercises: most of them were tested in real contexts, involving actual research teams, journal editors, institutional offices, and funding organisations. This allowed the project to observe not only whether ideas were theoretically promising, but also how they fared when confronted with the tempo, constraints and habits of everyday research work. The process of co-creation—where tools and materials were developed iteratively with the people who would eventually use them—proved particularly valuable. It ensured that solutions remained attuned to disciplinary norms and differences, and that the tools did not become overly prescriptive or detached from practice.

At the same time, the Pilots also revealed some limitations. Some interventions were only piloted in a small number of settings, making it too early to judge how well they will scale. Co-creation, while vital for stakeholder engagement, uptake and relevance, is intensive and depends on internal and external support from institutions and sustained engagement that not all Pilots could readily support. And because research practices and incentives differ across disciplines, several tools may require adaptation before they can be meaningfully taken up in other epistemic contexts. In other words, the Pilots did not produce a universal “recipe” for reproducibility—and perhaps that is precisely the point. Reproducibility is context-sensitive, according to our main point that epistemic diversity should be taken into account and successful interventions need to acknowledge and work with that diversity. Furthermore, one of the Pilots was not successful. The decision aid (Pilot 1) was not ultimately successful in creating a prototype due to several reasons. One reason was a lack of resources, and it remains difficult to assess whether the KPMs are now applicable to use in real life settings. The Pilot would have studied this as the goal was to develop an aid to make responsible judgement on which reproducibility practice/tool/intervention is eligible.

10.2. Next Steps

Looking ahead, a key challenge will be to embed the lessons and tools from the Pilots into the routines and infrastructures of the whole research enterprise. This will involve not only expanding access to the tools themselves but also ensuring that they integrate smoothly with existing workflows rather than adding new burdens. Sustained support—training, documentation, active communities and institutional buy-in—will be crucial. These elements are also highlighted by the

results from the future studies that revealed the main barriers and enablers for culture change towards more reproducibility (OSF) .

There is also an opportunity to strengthen alignment across different parts of the research ecosystem. Tools developed for researchers can be linked with monitoring dashboards for institutions or funders; journal editorial checklists can reinforce data and workflow transparency encouraged elsewhere. Increasing automation may help make reproducibility practices less reliant on individual effort and more simply “how things are done.” The next phase, therefore, is not only about scaling but also about weaving these interventions together, and about efficient and effective implementation. This will require time and efforts. However, we also are confident that future EU funded projects will take up this challenge and use our lessons learned and our recommendations to further this process.

10.3. Synergies between the Pilots

What becomes clear when examining the Pilots as a whole is how well they complement one another. Some focus on the individual researcher and the day-to-day decisions that shape the transparency of a project. Others operate at the level of institutional or publisher policy, shaping the environment in which those decisions are made.

This multi-level approach is important. Efforts to change research practice often fail when they focus solely on either individual behaviour or structural incentives. The Pilots suggest that progress is most likely when both are addressed simultaneously, and when shared tools, practices, interventions, policies and expectations circulate across the system. The co-creation approach—recurrent across several Pilots—acted as a connective tissue here together with the stakeholder communities that have been formed. It allowed different actors to develop a shared understanding of what reproducibility means in their particular setting, and how it can be practised without undermining disciplinary identity or research creativity.

10.4. Implications and recommendations

The Pilots show that reproducibility can be strengthened in concrete, practical ways if attention is paid to usability, context, and incentives. To build on this momentum, we recommend the following:

- Embedding reproducibility interventions at the earliest stages of research, such as in funders evaluation and monitoring practices, and at the start of project planning, rather than addressing them retrospectively.
- Continuing to involve diverse types of researchers, editors, and administrators in the design of tools and policies and keep the formed stakeholder communities active for future projects.
- Ensuring tools remain lightweight, flexible and adaptable, rather than prescriptive or overly standardising.
- Encouraging interoperability between systems, so that metadata, planning documents, and workflow traces can travel with the research.

- Supporting cultural change through recognition, incentives training, and community engagement, rather than relying solely on compliance.
- Maintaining iterative evaluation so that tools evolve alongside research practice rather than becoming outdated or unused.

Acknowledgements

We would like to thank all contributors to the Pilots. Without their help, efforts, engagement and support it would have been impossible to conduct so many different Pilots at the same time.

11. Appendix

Appendix 1 - Intervention email for Pilot 7

Subject: Benefits of Open Data sharing for your manuscript [manuscript ID] at [Journal name]

Dear [insert author name],

This email relates to your recent submission to [Journal name]. You should have received a separate email regarding the outcome of the peer review process for [Manuscript ID / Article Title]. If you have not received this email, or have other queries relating to your manuscript, please contact [journal editorial email address].

When submitting to [Journal name], you agreed to make available the data and materials supporting the results or analyses presented in your paper. The policy of the journal requires that data is shared upon reasonable request, when you are contacted by future readers. Because it is beneficial to you and to others, we would like to encourage you to share your data in a trusted data repository however, rather than sharing only when requested to do so by readers.

Benefits of immediate Data Sharing in Data Repositories:

- Increased Impact: Studies show that publications with shared data receive more citations.
- Cumulative Science: It enables other researchers to build on your work.
- Reproducibility: Sharing data allows others to verify and reproduce your findings (giving you more opportunity for credit and recognition).
- Easier to manage: No additional effort if a reader requests access to the data (which could be months or years after publication).

When preparing your revised manuscript, please consider sharing the data and materials supporting your results or analyses in a data repository, and indicate in your Data Availability Statement where the data can be accessed.
Note: Your choice to use a data repository, and any subsequent revisions of your Data Availability Statement will have no impact on the editorial decision regarding your submission.

How to Share Your Data:

- Identify your data: You should share all of the data and materials supporting the results or analyses in your paper, including the data used to build graphs, tables or other figures.
- Select an appropriate data repository: You can find trusted data repositories where you can upload your dataset including some descriptive information (metadata) at https://www.re3data.org.
- Protect sensitive data: If public sharing is not possible due to ethical concerns, consider whether it will be possible to anonymise your dataset or use repositories like Zenodo to grant access to individual researchers (if your participants have provided consent).

You can find more extensive guidance on data sharing at [Journal name] at [Insert link to publisher specific guidance if available].

If you have additional questions about how to share the data supporting your manuscript, just respond to this email.

Yours sincerely,
[insert Open Research team signature]
[insert Email address]