Appendix 4

Open Research Indicators: Report of a pilot on the downstream effects of research output

Contributors

Alice Howarth[1]*†, Anita de Waard[2]*, Ann Campbell[3], Bill Greenhalf[1], Etienne Roesch[4]†, Etienne Vignola-Gagné[2], Euan Adie[5], Evangeline Gowie[4], Harry Dimitropoulos[6], Iain Hrynaszkiewicz[7], Leonidas Pispiringas[6], Phillip Hall[3], Tim Vines[8], Valerie McCutcheon[4]

[1] University of Liverpool, [2] Elsevier, [3] Digital Science, [4] University of Reading, [5] Overton, [6] OpenAIRE, [7] PLOS, [8] DataSeer

*Alice Howarth and Anita de Waard are shared first authors.
† Correspondence should be addressed to Alice Howarth and Etienne Roesch; E-mail: alice.howarth@liverpool.ac.uk and e.b.roesch@reading.ac.uk.

Executive summary

This pilot explored how to better understand and measure the downstream effects of research output - impacts that go beyond direct citations, such as data reuse, influence on policy, or cross-disciplinary knowledge transfer. Recognising that traditional metrics fall short, the pilot brought together universities and data providers to develop a proof-of-concept indicator combining citation network analysis with narrative interpretation.

The pilot delivered:

- A prototype tool to visualise research influence over time via citation graphs;
- A hybrid methodology that combined both quantitative and qualitative interpretation of impact;
- A set of institutional use cases highlighting real-world needs for measuring downstream effects;
- Exploratory methodologies from data providers to track data reuse, showing both progress and persistent gaps;
- Insight into infrastructure challenges, especially the inconsistent tracking and citation of datasets.

Conducted by the Universities of Glasgow, Liverpool and Reading (co-Leads), the project showed how institutions can use these tools to support open research practices, inform strategic planning, and contribute to responsible research evaluation. This pilot provides a foundation for developing richer, more meaningful indicators that reflect how research makes a difference—across time, disciplines, and society.

Background and aims

Downstream effects of research output refer to the influence that scientific work exerts beyond its immediate citations, including indirect citations, incorporation into new areas of inquiry, adoption in technological innovations, and influence on policy decisions. These extended

effects are crucial to understand the long-term value of research, particularly in fields where knowledge dissemination occurs over extended periods and through diverse pathways. Despite their significance, existing citation-based metrics provide limited insight into these broader impacts, requiring the development of new indicators tailored to capturing downstream influence.

This pilot aimed to define and develop indicators to measure the downstream effects of research output using citation data. By leveraging citation network analysis, second-order order citations, and visualise knowledge flow, we tried to establish a framework that reflects the cascading impact of research over time, and may provide Institutions with a way to formulate useful narratives about the research produced. The pilot focused on exploring methodological approaches to visualising these effects and evaluating their applicability to research evaluation and policy, and produce a proof-of-concept to ground further work.

We envisioned an indicator that could provide a more comprehensive view of research impact than traditional metrics, moving beyond citation counts to capture the broader influence of research on society, policy, and practice. The pilot also aimed to identify and evaluated the data sources and processing steps needed to produce these indicators, the challenges that institutions and providers may face and to provide guidance, should others wish to engage in similar work.

The following sections outline the theoretical foundations of downstream research impact, describe the methodology employed in developing these indicators in this pilot, present the results of the pilot study, and discuss the implications of this work for research evaluation and policy. Through this work, we aim to contribute to a more nuanced understanding of scholarly influence, moving beyond direct citation counts toward a more holistic assessment of research contributions.

Goal and objectives

The goal of this pilot was to develop an indicator that would allow Institutions to measure the downstream effects of sharing data, and to provide guidance on how to use this indicator in practice.

The objectives of the pilot were:

1. for Institutions to identify the use cases for such an indicator;
2. for Providers to develop algorithms to identify data reuse in the literature;
3. for Institutions to develop a proof-of-concept, incorporating the above, and document the process, challenges and lessons learned;
4. for pilot partners to formulate recommendations for future work.

Scope

This pilot aimed to explore the impact of data sharing, investigating how shared research data is used, cited, and valued across research institutions. Our first meetings focused on defining the pilot's scope, setting up methodologies to work together, engaging with Providers, establishing a shared language and discussing the use of metrics to measure the benefits of data sharing.

Early discussions showed that focusing solely on research data may prove difficult, because data are not a kind of research output that is easily identifiable in the databases that were available to us.

Institutions in the UK typically maintain two kinds of public databases:

- a repository of publications, where researchers are mandated to upload research output for inclusion in research evaluation exercises, like REF2029 (https://2029.ref.ac.uk);
- and a repository of data: If it exists, researchers are not mandated to use it—and very few do, in fact use it when it is available. University Data Managers often recommend to researchers that they should use discipline-specific repositories instead, which may be organised by learned societies, journals, or other funded projects, and that can be more suitable for complex or large data, with technical requirements. In addition, data deposited in institutional repositories are not always assigned reliable URIs consistently across institutions, like a DOI, or leveraging global infrastructures that may assign persistent identifiers, like Zenodo or Crossref.

Additionally, in most disciplines, data are typically not published on their own, in a “data paper”, but are often part of a broader research output that can also comprise code, analyses and a broader scholar output.

Providers have access to other databases of publications and data, either that they themselves have constructed and maintain, or associated with journals that they publish. Journals do not consistently use data repositories, however. These databases are generally proprietary and not publicly accessible, are subscription-based or not for use in automated systems, e.g. through an API. In most cases, databases specifically about data remain rudimentary, and with no usable link to databases about academic papers. Novel combinations of publication datasets and open data datasets needed to track data reuse are just starting to emerge and have yet to be validated for use at at-scale.

Indeed, there is very little published literature available on the monitoring of data reuse, indicative of a novel research front for the “science of open science”. Of a handful of prior relevant studies, Sheehan et al (2024) have used the Dimensions database to track publications making mention of COVID-19 sequences-archiving repositories, GISAID and the INSDC family of databases. They characterize the repository-citing publications on a number of dimensions, including altmetrics, geographical spread of citations, or collaboration with low and middle income countries. Unfortunately, it is not clear from the publications’ methods whether the authors have aimed to specifically monitor data citations towards open datasets archived within the repositories; or have captured all publications making any mention at all of the repositories, whether as part of a data deposition statement or even when generally discussing research infrastructures.

In an arXiv preprint, Krause et al (2023) have used OpenAlex’s database of data citations, based solely on those data citations retrieved within the formal reference lists of journal publications and other document types indexed. From the more than 200 million works identified in OpenAlex, they identify roughly 85,000 documents including at least one data citation amongst their references within the full database. Over 95% of those open datasets that are indexed by OpenAlex appear to be uncited, and there is high skewness in citations towards just a few open datasets. Krause et al are able to show uneven distribution of data citation practices by discipline and geographical region.

Park and Wolfram (2017) carefully examine a small sample of 148 data-citing publications to derive observation on determinants conducive to this practice. The same authors subsequently used Clarivate’s Data Citation Index (DCI) to characterize citations towards

open research software depositions in Zenodo, ASCL, CRAN, Nanohub, ModelIDB, Figshare, among other repositories (2019). Earlier, Robinson-Garcia et al (2015) had also characterized the DCI, but found the tool was still in an early development at that point in time, with most data citations provided by only a single repository.

Hemphill et al (2022) used unique download user frequencies for 380 open datasets from the Inter-university Consortium for Political and Social Research repository, in their correlation analysis of usage as a dependent variable with data properties, curation decisions and repository funding models as independent variables.

In a different methodological register, Gregory et al 2023 have conducted a survey administered through a questionnaire sent to a representative sample of 2,492 researchers from all disciplines. They found that slightly more than 80% of respondents considered that they have reused data from other researchers and roughly the same proportion considered they shared their own data in the past. The authors do mention a possibility for self-selection bias influencing these answers. The study also highlights the variety of different usages for open datasets, but also a variety of motivations for not making active use of other researchers' datasets.

We thus decided to broaden our initial scope and focus on the return from participating Institutions to REF2021, to establish a starting point in time and leverage a period of a few years during which the research work returned may have led to measurable downstream effects. Whilst our general goal evolved to include the downstream effects of research output, including data, some of the Providers in the pilot were also interested in the development of algorithms that would permit the identification of data reuse in the literature, and we hoped these two strands of work would converge.

Methods

We followed a project management methodology loosely based on the Agile methodology in software engineering, with a focus on iterative development and flexibility. The project was divided into Sprints, each lasting one month, with a clear goal and deliverables. In between Sprint meetings, partners would meet in groups as needed, work on their tasks asynchronously, using a shared Google folder for documentation and a Communication File for discussion. All documentation was stored in a Google folder. Communication was initially centralised in the Communication File, to prevent cluttering of inboxes, but in the end, we did not make use of the file for asynchronous discussion as was intended, and instead used it to keep minutes of meetings and decisions. This was important, as it allowed partners to refer back to previous discussions and decisions on their own time, and to keep track of the project's progress if they missed meetings. Communication through email was kept to a minimum, and for the main purpose of general communication and occasional exchange of files.

Within the first couple of months, we designed a roadmap, describing the stages of work and the general "direction of travel" for the pilot. The roadmap was a useful tool to keep track of the project's progress, and ensure that the pilot remained aligned on the project's goals and objectives. However, this document constrained some of the industrial partners, who needed to engage in developments aligned with their priority and consolidate efforts for ongoing projects, such as output needed for their participation in other pilots. When needed, we also used the structure of the roadmap to discuss and clarify the pilot's scope, objectives and deliverables.

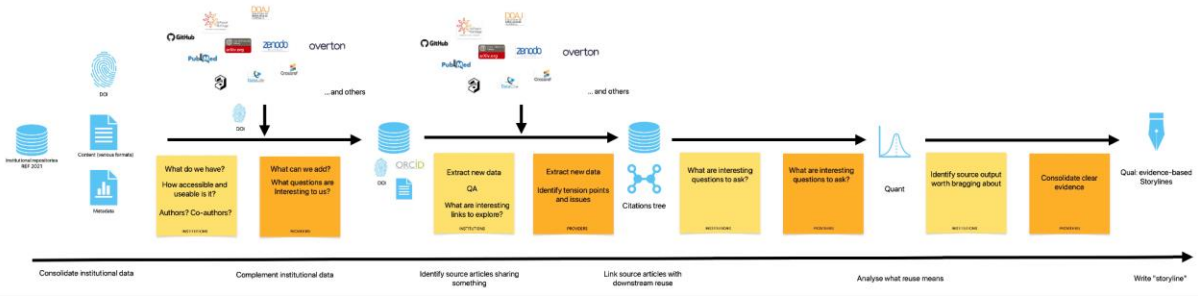

Figure 1. Roadmap

Our main goal was to produce an indicator that would rely on both qualitative and quantitative material for an Institution to formulate a narrative ("storyline") about a specific piece of published research: for a given list of research output, data or otherwise, an individual would be presented with a visualisation of the downstream citations that could be queried to answer specific questions, alongside metrics related to this citation graph; it would then be their task to contextualise this result into a narrative.

The decision to focus on a method that would augment quantitative results with qualitative interpretation was an explicit attempt to address the known limitations of quantitative indicators, which may not capture the full extent of research “impact”, favour a certain kind of output, measure the wrong things (Csiszarm, 2020), could easily "be gamed", or provide a false sense of robustness (Merton, 1970; O’Neil, 2017; Pardo-Guerra, 2022). A large majority of researchers and research supporters may indeed know some quantitative indicators that are used to evaluate research output, but very few actually understand how they are calculated or the contextual dynamics that influence their use (Rousseau & Rousseau, 2017).

As per the roadmap, we divided the pilot into four phases:

The first phase was dedicated to the identification and consolidation of data from Institutions and data providers. We collected data from the participating Institutions, and focused on research output and data returned to REF2021. Such research output is typically identified by a DOI, and may include scholarly output, data or code, and thus we focused on using DOIs as a common identifier across data sources.

The project as a whole was guided by the iNORMS SCOPE framework for research evaluation, and institutions committed in writing their motivation for contributing to the pilot, summarised below:

– The University of Glasgow are interested in the downstream effects of research sharing to track compliance with their Code of Good Practice in Research that encourages making data openly available for re-use. This will help plan future activities and support. Potentially, a number of re-used datasets could be added to our open research key performance indicators for internal planning purposes. Improved monitoring methods could increase confidence in our reporting, provide more coverage, and facilitate discussion between organisations on supporting a strong UK drive to openness. It is not intended to monitor individuals or schools but rather to inform and support them to grow open practice.
Of course, openness is included in promotion criteria for some grades so that is another use for records that have suitable indicators.

– The University of Liverpool are interested in monitoring open research practices in support of continuous improvement, service development, and strategic planning. This work aims to assess the impact of Libraries, Museums, and Galleries on research behaviours, measure engagement and awareness across all career stages and disciplines, and inform preparations for REF2029. Oversight is provided by the Open Research Leadership Group, with strong

links to internal networks and the wider academic community. Key values driving this initiative include collegiality, transparency, openness, team science, and a strong commitment to diversity and inclusion, supported by projects such as THRIVE. The focus is on promoting FAIR Data principles and the CRediT taxonomy, encouraging behaviours like data sharing and acknowledging contributions. Monitoring began in January 2021, aligning with REF timelines, and will continue to track progress, identify areas for improvement, and highlight best practices to strengthen the University's research culture and global standing.

– The University of Reading are committed to fostering a strong Open Research culture, guided by our 2018 Statement on Open Research and 2019 Research and Innovation Strategy, renewed since to cover 2024-30. Researchers are expected to adopt practices that ensure research is accessible, transparent, reusable, and reproducible. Policies mandate Open Access publishing, the use of data availability statements, and preservation of data in repositories adhering to FAIR principles. The University encourages openness for all research outputs, including code, software, digital resources, preprints, methods, protocols, and hardware designs, applying FAIR principles throughout the research lifecycle. Through our second Open Research Action Plan (2024-29), the University strategically increase staff engagement in Open Research practices to continually enhance research quality. Our contribution to the Indicators Pilots allowed us to evaluate our practice and identify gaps, and scaffold interaction with our community of researchers on setting up responsible research evaluation. Particularly, our Research Culture programme board is supporting the use of the SCOPE framework and the use of responsible indicators at all levels of management in the University.

In the second phase, we investigated how the data we gathered could be complemented with metadata from third-party sources, such as Crossref, ORCID, OpenAlex, Overton, DataCite, etc. We aimed to be in a position to produce a comprehensive view of the research output, to help focus further analyses. We met with representatives of OpenAlex, Overton and Crossref, to understand the data they provide, and how it could be used to complement the data from the Institutions. In addition, Institutions drew a list of questions that they would like to answer with these data, and we used these questions to guide further development. We also discussed the limitations of the data, and how these limitations could be addressed.

The third phase focused on the parallel developments of algorithms to identify data reuse in the literature by Providers, and to visualise impact of research output in citation graphs. Providers worked separately, as they aimed to align this work with their priorities and constraints, but shared progress and results with each other throughout the Pilot, but no code. The visualisation algorithm was developed by the Institutions, and available in the UKRN Github.

The fourth phase was dedicated to the integration of this work into a proof-of-concept for a single indicator, and write up of results.

Data sources and tools

REF2021 datasets – The Universities of Glasgow, Liverpool and Reading combined their return to REF2021, ca 143K publications covering the period 2014 to 2021. Most of these data were already accessible publicly. We created Excel spreadsheets containing the DOI, links to local repositories, types of publications and other metadata.

Proprietary tools – Providers used the databases and tools accessible to them, most of them only accessible through a subscription. Participants in the pilot had very limited access to each

other's tools. Providers shared tests and output to the partners, which we discussed in Sprint meetings.

Open source development – Institutions developed algorithms to explore citation graphs, calling on the OpenAlex API. These toys algorithms anchored reflections and demonstrated feasibility of some of the ideas discussed in Sprint meetings. This code is available at https://github.com/UKRN/ORP-Pilot3/releases/tag/v0.1.

OpenAlex, Crossref, Overton, DataCite – We engaged with providers external to the Pilot to understand how these tools could be used to complement our work. UKRN is a founding signatory of the Barcelona Declaration on Open Research Information (2014), and we were mindful not to include in our development tools that were not readily available to the public. The providers provided guidance and useful advice.

Results

Use cases by Institutions

As evidenced in their return on the initial iNORMS SCOPE survey, the Universities of Glasgow, Liverpool, and Reading engaged in this pilot for largely similar reasons: advancing open research through monitoring, strategic planning, and community engagement, in fulfilment of their public commitment to robust and transparent research. Each institution seeks to increase transparency, reproducibility, and data sharing, while aligning open research practices with broader goals such as responsible evaluation, REF preparation, and enhancing research culture.

In this context, we surveyed the relevant stakeholders at our Institution, including Librarians, Data managers, Policy Owner and individuals in management positions related to research and innovation, to understand how they would use an indicator of the downstream impact of research output, including data. We used these questions to establish the general direction of the Pilot project.

- What research is cited from the seed dataset?
- Why are they cited? -- i.e. venue, platform for sharing, communication associated with the research, etc
- Who is citing them?
    - Journals
    - Institutions
    - Countries, NGO
    - Authors of papers citing (important names in the field?)
- What OA features are associated with the research?
- Is the citation related to a change in institutional policy or communication to Colleagues?
- If funded, where does the funding come from?
- Context: timeline of keywords, to identify jumps in disciplines and fields of study
- Size of citing bubble, publication metrics, characteristics of the network
- Alignment with WHO sustainability goals
- Correlation with OA practice at the Institution?
- Correlation with Altmetric? (i.e. What are the features that drive citations?)
- Is there any evidence of compliance, e.g. OA requirements, funders?
- Differences by disciplines?
- Changes over time
- Impact of exhibitions in increasing citations

- Impact of books
- Impact of classics and arts

Graph exploration algorithm to answer the Institutions' questions

Citation networks provide valuable insights into the structure and dynamics of scientific knowledge diffusion. The Institutions were interested in producing a quantitative foundation to a human interpretation of impact over time, for a given piece of work identified by a DOI. We thus developed algorithms to query the OpenAlex API and created a simple citation graph to visualise changes in features of interest over time.

Figure 2 is an example of output exploring the jumps of citation between disciplines. Citation depth is indicated by the size of the nodes, and the primary field of investigation is indicated by the colour of the nodes, as reported by the OpenAlex API. In the example, the root of the graph is a published paper in Neuroscience. The graph demonstrates occasional jumps to the fields of affective computing and computer science.

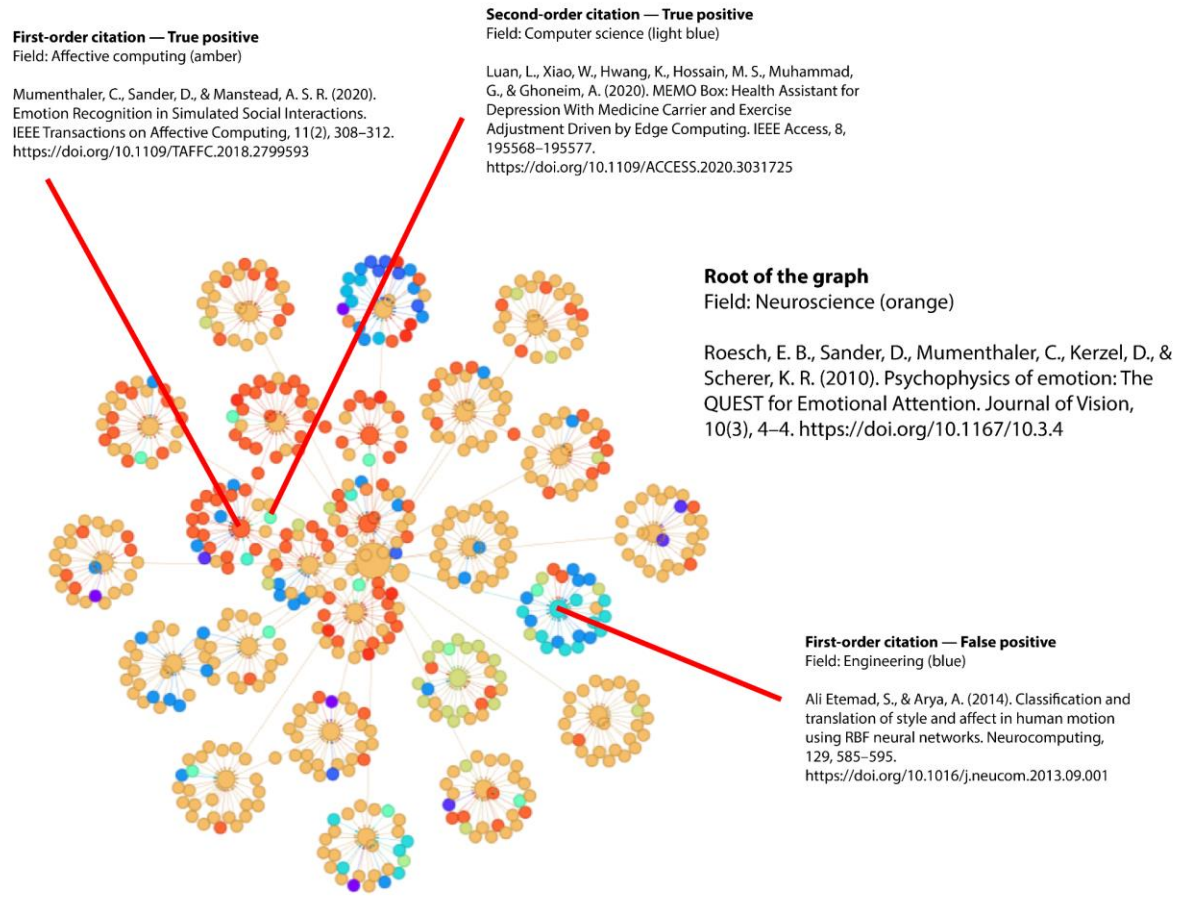

Figure 2. Example of application of the graph exploration algorithm produced by the Institutions from a specific DOI, using OpenAlex. The example illustrates jumps of citation between disciplines over time, as well as the reliance on quality citation data. Links point to specific nodes of the network, including first and second order citation, true and false positives (that do not in fact cite the previous node).

Several quantitative indicators have been developed to interpret such networks at both macro and micro levels. At the network level, basic measures include size (node and edge count) and density, which indicate the scope and connectedness of the literature (Waltman et al.,

2011). More sophisticated structural measures include clustering coefficients, which reveal the tendency of nodes to form tightly connected communities (Radicchi et al., 2012), and modularity, which quantifies the strength of division into clusters (Newman, 2006). Specific algorithms can be used to detect communities (Šubelj et al., 2016).

For assessing individual publications, degree centrality measures like in-degree (citation count) remain fundamental, while eigenvector centrality and PageRank variants offer more nuanced evaluation by weighting citations based on the importance of citing papers (Chen et al., 2007; Ding et al., 2009). Temporal indicators such as the Main Path Analysis (Hummon & Dereian, 1989) and burst detection algorithms (Kleinberg, 2003) can help identify critical pathways of knowledge flow and emerging research fronts.

Sophisticated bibliometric indicators like the h-index and its derivatives have been adapted to citation networks to evaluate the impact of both individual papers and authors (Hirsch, 2005; Egghe, 2006). Despite their widespread adoption, they face substantial criticisms and we wanted to advocate for the use of a bucket of measures, in support of more meaningful interpretation. Specifically, measures that summarise “impact” in a single statistic cannot account for disciplinary differences in citation practices (Bornmann & Daniel, 2009); e.g. in the field of History, having large citation counts indicate a strong level of opposition in the findings presented). They indiscriminately correlate with seniority (Costas & Bordons, 2007), are strongly correlated with publication quantity and sensitive to self-citation (Bartneck & Kokkelmans, 2011). Perhaps most fundamentally, these simplistic numerical indicators reduce the complex, multidimensional nature of scholarly impact to single figures, potentially encouraging gaming behaviors and distorting research incentives while failing to capture qualitative aspects of research value (Hicks et al., 2015; Wilsdon et al., 2015).

Our algorithm makes use of publicly accessible databases. It is an illustrative toy example that can be replicated on any database. However, as demonstrated in Figure 2, the quality of the graph produced is inherent to the quality of the citation data used to create the graph: Errors in the citation data will lead to erroneous edges, as demonstrated in the example.

DigitalScience

As part of Digital Science’s involvement in this pilot, structured metadata and citation linkages within the Dimensions platform were analysed to support the measurement of downstream effects on research data sharing. REF2021 submission data was imported into the Dimensions Google BigQuery (GBQ) environment and was matched to the Dimensions publication dataset through DOIs. This enabled the identification of sections such as Data Availability Statements, specific metadata fields and in-text references to persistent identifiers within the publication - an indication of when a dataset is listed, described or shared as part of a research output.

Dimensions incorporates two key datasets that can be used for this analysis: The **Publications** dataset which allows full-text mining of outputs, metadata and potential ‘trust markers’ such as data availability statements and author contributions; and the **Research Datasets dataset**, which sources and curates datasets from repositories such as Figshare, Zenodo, Github and PLOS. These two datasets are linked within Dimensions through persistent identifiers (typically DOIs) allowing the identification of:

1. When a dataset is mentioned or described in a publication
2. Whether it is later cited in subsequent publications, indicating potential reuse

By combining both elements, it is possible to assess not only the presence of data sharing practices (as covered in other pilots) but the extent to which shared data has been reused and

cited in later research, providing insight into the downstream effects of open research practices.

REF21 Data was extracted for 3 Institutions - University of Reading, University of Liverpool and University of Glasgow. This amounted to 6566 distinct DOI's. 81% of these (n=5350) were matched to publications within the Dimensions Database.

A quick analysis against a pilot dataset within Dimensions which identified the presence of certain Open Research Indicators and Research Integrity 'Trust Markers' indicated that a small proportion of these publications included a Data Availability Statement. 4722 out of the 5350 were processed. A DAS was identified in 14% (n=654)

Further analysis indicated that 178 used an online repository, 72 DAS's indicated that data was available in additional files and 187 DAS's indicated that data was available on request.

The second part of the analysis attempted to identify Datasets that were linked to these publications.

Again using Dimensions on GBQ, a script was written to flag any Research Data Datasets that were associated with the REF21 Publications. The REF21 Publications were restricted to those with a matching Dimensions ID.

1325 *Datasets* were identified from 256 Publications within the REF21 Publications (n=5350)

Further analysis will allow us to:

1. Search for data citations using the Dataset DOIs (above) and both the Datacite Corpus and OpenAlex
2. Explore trends by Field of Research
3. Analyse correlations between DAS quality and Dataset Citations

In parallel to Digital Science's involvement in this pilot, a separate internal study conducted within the Data and Analytics Hub - led by Joerg Sixt, Mark Hahnel and Kathryn Weber-Boer investigated global trends in data citation using the Dimensions platform on GBQ. This unpublished analysis spans the period 2004 - 2023 and reveals that while data citation is on the rise (with over 1.5 million publications citing datasets), only an approximate 10% of 30.5 million tracked datasets have been cited at all. Only 85, 000 of these have received more than 1 citation. These early findings highlight the progress made yet the significant challenges that remain in evidencing data reuse across the research landscape.

Participating in this data pilot alongside several institutions and peer data providers was a rewarding and eye-opening experience. It provided a golden opportunity to engage with domain experts, uncover untapped potential within the Dimensions provision, and gain a clearer understanding of where efforts should be focused to strengthen the open research ecosystem. The pilot highlighted the importance of evidencing the benefits and incentives of data sharing, promoting the use of trusted repositories, and ensuring data is both findable and reusable. However, it also underscored the complexity of the challenge—disciplinary norms and dataset characteristics vary widely, data is often unclean, persistent identifiers not always used and assumptions cannot be made. Balancing this exploratory work with day-to-day responsibilities was not without difficulty, but the experience reaffirmed the critical need for continued collaboration across all stakeholders, including private sector data providers, to drive progress.

Elsevier

Elsevier developed a tailored pipeline combining Scopus reference data, DataCite open dataset metadata and LLM synthesis to identify or summarize data reuse for this pilot project. Elsevier did not possess a prior analytical pipeline for open data re-use prior to participation in this project. A novel approach was therefore trialled with the hope to establishing

foundations for future development and scaling. Elsevier performed a novel combination of the DataCite database with Scopus records on secondary references to identify relevant data DOIs in indexed articles' references. The DataCite database indexes individual datasets deposited on a selection of open repositories, including a metadata field for a unique DOI identifier. Direct matching of those DataCite-DOIs was performed on Scopus records of secondary references, that is, references made within Scopus-indexed publications towards non-Scopus-indexed publications.

Between 2021-2024 within Scopus-indexed publications, a share of 1.3% of publications made one or more data citations towards the DOIs of DataCite-indexed datasets amongst their references. This number saw a slightly upward trend in time, falling below 1% in years before 2021.

It should be noted that the distribution of Scopus+DataCite-tracked data citations was heavily skewed by discipline. The figure reported above climbed up to 17.4% in the field of Meteorology & Atmospheric Sciences; 13.5% in Oceanography; or 10.3% in Ornithology. Data citation practices themselves may be shaped by disciplinary cultures; or these figures may alternatively or in addition capture disciplinary coverage biases in the databases used.

It can also be noted that 12.3% of Scopus-indexed publications issued between 2021 and 2024 made their references fully (100%) towards (Scopus-indexed) books and journal publications. In these publications, there is no chance to find a data citation amongst references.

On the data repository end (as indexed by DataCite), close to 10 million unique dataset DOIs were retrieved for the 2021-2024 period. A share of 0.7% of these datasets had been re-used at least once, as indicated by our approach to data citations. Re-use rates were quite skewed with a few repositories seeing very high level of re-use. More than 22,000 DOIs were identified for Dryad 2021-2024 datasets, of which 28.5% had seen at least one data citation. More than 2 million unique dataset DOIs were issued through the Global Biodiversity Information Facility between 2021 and 2024, of which 0.1% saw at least one data citation.

Precision could be considered problematic in this set, as in a manual evaluation of 100 data citations, only 55% of cases could unambiguously be considered to amount of re-use of datasets by external teams. In 18% of cases, one or more author(s) were shared between dataset and citing article. In 27% of cases, examination of dataset title and citing article title relevant great similarity in addition to author(s) overlap, indicating that a data availability statement had been made within the format of a reference.

The approach also has clear recall limitations in that it cannot capture data citations on the basis of accession codes/numbers; or data citations made in the body of the article text rather than in the reference section; or data citations towards the DOIs of open datasets not indexed in DataCite.

That said, these precision and recall limitations were mitigated through the production of small-scale, manually curated case studies of data reuse for datasets with a DOI and deposited with the institutional repositories of the pilot-participating institutions. Citation contexts from CC-BY (to allow for text mining) publications making a data citation towards the datasets of interest were manually extracted and provided to an Elsevier-proprietary implementation of ChatGPT for narrative summary. While a narrative summary could also have been produced manually, an AI-produced one highlights a pathway for future large-scale production.

Considering a dataset reuse narrative for the "MRC IEU UK Biobank GWAS pipeline version 1" which was deposited and published on the University of Bristol institutional repository in December 2017, the following AI-driven narrative summary was obtained:

> “The "MRC IEU UK Biobank GWAS pipeline version 1" dataset has significantly contributed to advancing genetic research in various health domains. One notable study, "A phenome-wide bidirectional Mendelian randomization analysis of atrial fibrillation" by Qin Wang et al. (DOI: 10.1093/ije/dyac041), utilized this dataset to explore genetic associations with atrial fibrillation (AF). The research identified several causal risk factors for AF, such as body mass index, height, systolic blood pressure, and genetic liability to coronary artery disease, and highlighted potential therapeutic targets, including proteins DUSP13, IL6R, and TNFSF12. This study underscored the dataset's role in understanding the genetic basis of AF and identifying potential drug targets. Another study, "The association of psychological stress with metabolic syndrome and its components: cross-sectional and bidirectional two-sample Mendelian randomization analyses" by Cancan Li et al. (DOI: 10.3389/fendo.2023.1212647), employed the UK Biobank data to investigate the relationship between psychological stress and metabolic syndrome (MetS). The research found associations between psychological stress and hypertension but did not establish a causal link between stress and MetS. This dataset facilitated the exploration of genetic predispositions to stress-related disorders and their impact on metabolic health. Overall, the MRC IEU UK Biobank GWAS pipeline has been instrumental in enabling studies to identify genetic risk factors, explore causal relationships, and propose potential therapeutic targets, demonstrating its broad applicability and significance in the scientific community.”

Using the DataCite-matched-to-Scopus method, this dataset was found to have received a total of 17 data citations (including the two summarized above), although including a number of author-level or institutional-level “self-citations”.

Alternatively, this dataset was also retrieved on OpenAlex, and the citations towards it extracted to create a tree map-style citation network graph (including second order citations). Note that manual validation was not performed in the OpenAlex set. It appears that, at least in the one case considered here, OpenAlex has much better coverage of data citations received by the University of Bristol dataset considered, than what could be retrieved from DataCite (see Figure 3).

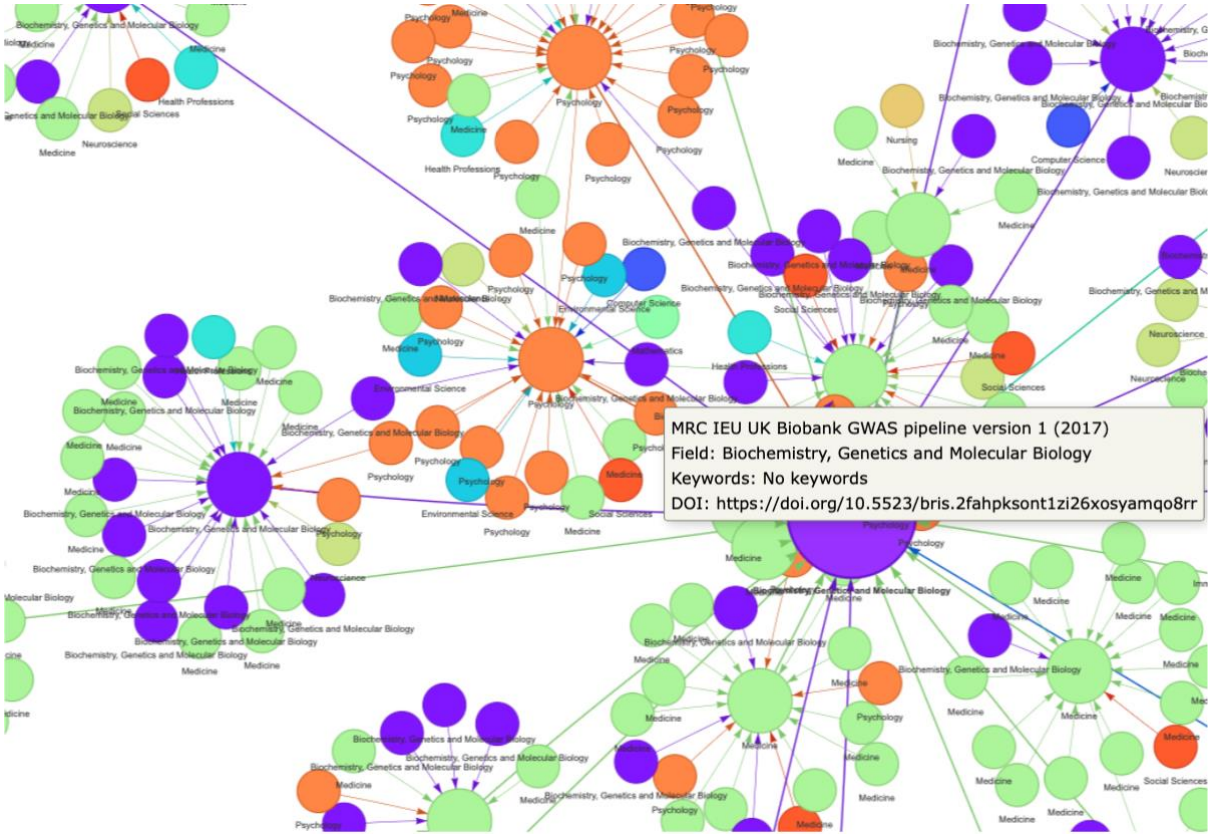

Figure 3. Application of the graph exploration algorithm produced by the Institutions to a selected case of data reuse as identified in the Elsevier output, using OpenAlex. The colormap indicates the primary discipline of the work reusing the dataset: orange for psychology, green for medicine, etc.

PLOS-DataSeer

PLOS-DataSeer focused on measuring data reuse, an important effect or consequence of data sharing, in publications. They first conducted a literature review and then a consultation of institutions, funders, meta-researchers (including and beyond institutions participating in Pilot projects) to establish the scope and requirements for an indicator of data reuse. This work highlighted concerns related to the definition and scope of what is understood as "data reuse" as it pertains to scholarly work, including replication. Further, the consultation emphasised that reuse does not imply data quality or robust procedures, and that the indicator should be able to differentiate between these aspects. The collaboratively developed requirements for a data reuse indicator were shared openly for public comment and consultation including with UKRN pilot institutions. At a high-level the requirements developed are:

1. Identify if a research article reuses research data
2. Identify which dataset(s) have been reused in the article
3. Identify how or why that(those) dataset(s) has(have) been reused by the article's authors

In response to public sharing of the requirements, 16 individuals provided a total of 85 comments, which have been summarised here.

To meet these indicator requirements, the project team developed a language model by fine-tuning Llama 3.1 8B on the task of identifying data reuse and sharing. Specifically, annotators at DataSeer labeled a corpus of 621 PLOS articles, which was split 80/20 into a training and testing set.

The language model was fine-tuned in three stages: supervised fine-tuning (SFT), reinforcement learning with verifiable reward (RLVR), and then reinforcement learning with human feedback (RLHF).

- **SFT:** Supervised fine-tuning was applied to learn the prompt structure. Using test-time scaling techniques to extend the model's reasoning trace, researchers sampled from the language model, prompted with each article, keeping completions that led to the correct prediction. Then, the language model was fine-tuned using next-token loss on these completions.
- **RLVR:** The language model was trained for three epochs using proximal policy optimization on the training set, without any inference-scaling technique applied to rollouts. Rewards were calculated based on the ground truth labels provided by annotators — specifically, the reward, between 0 and 1, was assigned as the percentage of the labels (data reuse true/false, data generation true/false, etc.) that the language model got correct.
- **RLHF:** After one round of review by PLOS Editors, the DataSeer team performed another round of fine-tuning to optimize the reasoning trace in the output of the language model for the preferences of the PLOS team. Specifically, the team learned a reward model (a linear layer applied to the final hidden state produced by Llama 3.1 8B) using a contrastive learning technique applied to examples of good and bad reasoning traces. The reuse LM was then updated 300 gradient steps under supervision by this model.

  The final model accuracy on the test set was 93% for data generation and 89% for data reuse. Future model checkpoints have exceeded this and the current accuracy on these fields is in the mid-to-high 90s.

Using this model, the project team analysed 4328 PLOS articles. Data reuse occurred in 2029 (47%) articles but only 176 persistent identifiers (83 accessions, 93 DOIs) for reused data were found. After further testing with users (editors, researchers and policy makers), they found that a modified version of the language model's reasoning trace provides usable information on how the data were reused. The results suggest that established bibliometric techniques for measuring data reuse may greatly underestimate data reuse. For example, measures of data reuse that detect links or citations to shared and reused datasets have estimated data reuse in just 1.6% of a sample of biomedical literature (Citation: Iarkaeva A, Nachev V, Bobrov E (2024) Workflow for detecting biomedical articles with underlying open and restricted-access datasets. PLoS ONE 19(5): e0302787.

https://doi.org/10.1371/journal.pone.0302787). This novel language model approach, which PLOS-DataSeer plan to publicly release preliminary results of in 2025, demonstrates that data reuse can be measured at scale using language models and generative artificial intelligence.

OpenAIRE

As part of this pilot's broader effort to explore indicators of data reuse, OpenAIRE adopted a complementary approach by investigating how text-mined links between research publications and datasets can support the development of narrative indicators for downstream data reuse, with a particular focus on direct & indirect citation patterns.
Approach

OpenAIRE leveraged its full-text mining pipeline to identify relationships between datasets and publications in the OpenAIRE Graph. Semantic relationships such as "IsDocumentedBy", "IsSupplementTo", "IsDerivedFrom", "Describes", "IsSupplementedBy", and "Documents" were used to infer potential cases of data availability or reuse. A curated subset of these

relationships was extracted to ensure a focus on high-confidence links likely to represent meaningful dataset sharing.

A sample dataset-initially shared in Excel format-included, for each publication-dataset pair:

- The publication DOI
- The related dataset DOI
- The number of direct+indirect citations of the related datasets

Measuring Downstream Effects

To assess the reuse impact of these datasets, OpenAIRE constructed a complementary dataset with citation counts:

- Direct citations: Number of publications citing the dataset DOI directly in the references list.
- Full (direct + indirect) citations: Includes citations to publications that cite the dataset directly in the references list and/or mention the dataset in the full text, serving as a proxy for indirect impact.

This analysis was based on OpenAIRE's internal version Graph 9.0.1, focusing exclusively on citations from publications (i.e., excluding dataset-to-dataset links). Despite deduplication efforts, some edge cases persisted, particularly in granular datasets, highlighting the limitations of current deduplication strategies and the need for better version control and dataset identity resolution.

A representative example from HEPData revealed multiple table-level DOIs (e.g., /v1/t137, /v1/t138) linked to a parent dataset, each with distinct citation counts. These variations, despite semantic similarity, underscore the importance of improved granularity management and dataset versioning.

Institutional Aggregation

Citation impact was aggregated at the institutional level based on a predefined set of publications provided by the UK universities participating in the pilot. Rather than inferring institutional affiliations from metadata, the analysis focused on this curated corpus to ensure alignment with the universities' own research publications. The aggregated citation metrics revealed varying levels of data reuse impact across institutions , which were grouped into three tiers based on total (direct + indirect) citation volume:

- High-impact institutions: Universities 7, 5, and 8, each with over 700 total citations, suggesting strong dataset visibility and reuse within the academic literature.
- Moderate-impact institutions: Universities 1, 3, 2, and 4, with citation counts ranging between 300 and 450, indicating moderate but potentially growing reuse influence.
- Low-impact institutions: Universities 9, 10, and 6, all under 300 total citations, possibly reflecting lower data sharing activity, limited discoverability, or metadata gaps.

University	Direct Citations	Direct + Indirect Citations
1	1	417
2	0	311
3	0	426
4	1	313

5	0	1248
6	0	137
7	10	1750
8	1	731
9	0	285
10	0	154

These figures support comparative insights and dashboard-ready summaries of institutional engagement with dataset-linked research outputs.

Narrative Indicators

The citation results offer a basis for deriving narrative-style indicators that help contextualise the patterns and potential implications of dataset reuse across institutions[1]:

- University 7 recorded the highest total citation count (1,750), including the most direct citations (10), which may suggest that its datasets are not only visible but also actively reused and cited within the scholarly ecosystem.
- Universities 5 and 3 demonstrated strong indirect citation performance (1,248 and 426 respectively), despite receiving no direct dataset citations. This may indicate that while their datasets are influential, they are more often acknowledged via the related publication rather than being cited directly.
- Universities 6 and 10 showed limited overall citation activity. This could be linked to lower levels of dataset sharing, reduced discoverability, or the presence of weaker semantic connections between the datasets and the associated publications.
- Several institutions appear to enable dataset reuse indirectly, as indicated by citations to their related publications. This highlights the role of these institutions in contributing to the research landscape, even when datasets themselves are not explicitly credited.
- The variation between direct and indirect citations across institutions points to potential differences in community practices, dataset citation norms, or visibility of the datasets in scholarly workflows.

These insights serve as early-stage narrative indicators and could be complemented in future work by triangulating them with additional signals, such as download counts, data mentions, or qualitative feedback from researchers. Furthermore, the integration of emerging tools such as large language models (LLMs) is expected to enable the extraction of richer, more comprehensive narrative evidence on dataset reuse, particularly by identifying implicit references, use cases, and qualitative context that may not be captured through citation analysis alone. For instance, LLMs could help classify types of reuse (e.g., methodological reuse, background data, validation), extract sentiment or intent from narrative mentions of datasets, and cluster publications by thematic reuse patterns. When combined with DAS metadata, LLMs may also identify inconsistencies or missing attributions, enhancing the completeness of dataset reuse indicators.

OpenAIRE's Conclusion and Outlook

OpenAIRE's output demonstrates a scalable methodology for tracking downstream effects of dataset sharing through a combination of text mining and citation analysis. While it does not

[1] Participating institutions in this pilot formulated a range of exploratory questions regarding downstream effects, such as who reuses shared datasets, what platforms facilitate visibility, whether open access practices correlate with reuse, and how data sharing aligns with funder requirements or disciplinary norms. While OpenAIRE's current approach focuses on citation-based reuse, it lays the groundwork for addressing such questions in future iterations. (See Methods section of the main report.)

yet quantify formal data reuse (e.g., replication or third-party analysis built directly on a dataset), it lays a foundation for developing narrative indicators of how shared datasets contribute to ongoing scholarly communication - particularly in fields where datasets are integrated into publications but rarely cited formally.

Looking ahead, future steps may include the application of large language models (LLMs) to enhance the extraction of richer reuse evidence. The latter could allow for more precise identification of reuse scenarios, detection of implicit attributions, and contextual interpretation of data sharing impact. Additional work could also explore aligning these indicators with FAIR data principles, connecting them to data management planning workflows, and incorporating community feedback to improve clarity and usefulness for decision-making and policy development.

Discussion: contributions, challenges and recommendations

This pilot addressed a persistent gap in research evaluation: understanding the downstream effects of research output—those broader, often indirect influences that extend beyond simple citation counts, including reuse in novel research contexts, policy adoption, and societal or technological application. Recognising that existing metrics insufficiently capture these effects, the project brought together universities, data providers, and infrastructure experts to co-develop indicators and methodologies capable of uncovering and contextualising the extended reach of research outputs, particularly shared datasets.

The primary contribution of this pilot was the development of a proof-of-concept indicator capable of combining quantitative citation network analysis with qualitative narrative interpretation, to trace and visualise the spread and influence of research over time. This framework enabled institutions to explore citation "storylines" for individual outputs, providing the basis for richer and more responsible narratives of research impact—especially in preparation for exercises like REF2029.

Crucially, the pilot demonstrated how data and methods can be combined to answer institutionally relevant questions, such as: who is citing shared research, in what contexts, through what platforms, and with what downstream implications? Participating universities—Glasgow, Liverpool, and Reading—shaped these questions based on their commitment to open research and responsible evaluation practices, and guided the technical development in alignment with their strategic goals. Their motivation extended beyond compliance: they sought actionable insights to improve support for researchers and enhance the visibility and value of open science practices.

The pilot produced several tangible outputs:

- A prototype graph exploration tool using OpenAlex data to trace first- and second-order citations and disciplinary shifts;
- A curated set of institutional use cases and indicator requirements;
- Explorations by providers (Digital Science, Elsevier, OpenAIRE, PLOS-DataSeer) into identifying and quantifying data reuse in diverse ways, each revealing critical methodological challenges and field-specific practices;
- Recommendations on data infrastructures and gaps, such as the inconsistent use of persistent identifiers for datasets, limitations in institutional repositories, and the opacity of proprietary tools.

Collectively, our efforts exposed both the potential and limitations of current infrastructures in tracking data reuse. Findings from Providers suggest that while data citation is increasing, it remains rare and unevenly distributed across disciplines, with significant under-reporting due to poor data, inconsistent referencing practices, and inaccessible proprietary data.

Furthermore, institutions and even most providers often lack the internal systems to monitor reuse systematically, and to produce the metadata that's needed, at a fine-grain level enough to both identify data that are shared and reused, and to provide contextual background for interpretation of their impact. That the scope of the pilot shifted significantly to include all research output precisely because we were not able to readily access enough metadata about data shared is telling. The success scenario that sees researchers consistently depositing (and reusing) data relies on both institutional and systemic structures and incentives, as well as a culture change within which data is part and parcel of the output that is valued, and not a mere by-product, as it is now (Lowenberg, 2019).

The quantitative results that we produced have to be taken with a grain of salt for several reasons. First, we were only able to assess information we were able to access ("the streetlight

effect"). As noted before, information about data shared is scarce, and information about data reuse is opaque due to inconsistent citation practices. Similarly, the moment that one relies on a database to extract statistics, one becomes dependent, locked-in really, to the focus and biases intrinsic to that database. The databases we used, proprietary and publicly accessible, also contain mistakes that can have significant effects but yet be invisible in aggregate statistics. As demonstrated in our example, false positives in citations will impact measures aggregate measures. Second, we relied on DOIs to identify both scholarly works and datasets; however, DOIs are predominantly assigned to specific types of outputs, such as journal articles, and often overlook other important forms of scholarly communication, including books and exhibitions.

Finally, we witnessed deep inequities in the research landscape as to how institutions are resourced to demonstrate the impact of their work. The ability to engage with advanced indicators of downstream research influence—such as data reuse tracking or citation network analysis—depends not only on the quality of research produced, but on access to costly tools, technical infrastructure, and specialist expertise. Many institutions, particularly those with fewer resources, face systemic barriers: they may lack subscriptions to proprietary platforms, have limited access to open infrastructures, or be without dedicated staff to analyse and interpret complex data. As a result, the capacity to surface and communicate research impact becomes unevenly distributed. This reinforces structural imbalances, where better-resourced institutions are more likely to be recognised and rewarded, while others—often serving more diverse or underrepresented communities—risk being overlooked. Addressing these disparities is essential for ensuring that research evaluation practices are not only robust, but fair, inclusive, and representative of the full spectrum of scholarly contribution. We invite institutions to evaluate their own practice and identify ways of improvement using Champieux et al. (2023), see: https://zenodo.org/records/7369811.

Conclusion

In sum, this pilot made the case for a hybrid, narrative-driven approach to research impact assessment—one that values downstream influence, engages institutions directly in indicator design, and leverages both open infrastructure and emerging technologies like large language models. The work lays important groundwork for future development of responsible, scalable tools that reflect the real-world pathways through which research—particularly open data—makes a difference.

References

Bartneck, C., & Kokkelmans, S. (2011). Detecting h-index manipulation through self-citation analysis. Scientometrics, 87(1), 85–98. https://doi.org/10.1007/s11192-010-0306-5

Bornmann, L., & Daniel, H. (2009). The state of h index research: Is the h index the ideal way to measure research performance? EMBO Reports, 10(1), 2–6. https://doi.org/10.1038/embor.2008.233

Champieux, Robin, Anthony Solomonides, Marisa Conte, et al. 'Ten Simple Rules for Organizations to Support Research Data Sharing'. PLOS Computational Biology 19, no. 6 (2023): e1011136. https://doi.org/10.1371/journal.pcbi.1011136, as well as: https://zenodo.org/records/7369811

Chen, P., Xie, H., Maslov, S., & Redner, S. (2007). Finding scientific gems with Google's PageRank algorithm. Journal of Informetrics, 1(1), 8–15. INSPIRE. https://doi.org/10.1016/j.joi.2006.06.001

Costas, R., & Bordons, M. (2007). The h-index: Advantages, limitations and its relation with other bibliometric indicators at the micro level. Journal of Informetrics, 1(3), 193–203. https://doi.org/10.1016/j.joi.2007.02.001

Csiszar, A. (2020). Gaming Metrics Before the Game: Citation and the Bureaucratic Virtuoso. In M. Biagioli & A. Lippman (Eds.), Gaming the Metrics (pp. 31–42). The MIT Press. https://doi.org/10.7551/mitpress/11087.003.0003

Ding, Y., Yan, E., Frazho, A., & Caverlee, J. (2009). PageRank for ranking authors in co-citation networks. Journal of the American Society for Information Science and Technology, 60(11), 2229–2243. https://doi.org/10.1002/asi.21171

Egghe, L. (2006). Theory and practise of the g-index. Scientometrics, 69(1), 131–152. https://doi.org/10.1007/s11192-006-0144-7

Gregory, Kathleen, Anton Ninkov, Chantal Ripp, Emma Roblin, Isabella Peters, and Stefanie Haustein. 'Tracing Data: A Survey Investigating Disciplinary Differences in Data Citation'. Quantitative Science Studies 4, no. 3 (2023): 622–49. https://doi.org/10.1162/qss_a_00264.

Hemphill, Libby, Amy Pienta, Sara Lafia, Dharma Akmon, and David A. Bleckley. 'How Do Properties of Data, Their Curation, and Their Funding Relate to Reuse?' Journal of the Association for Information Science and Technology73, no. 10 (2022): 1432–44. https://doi.org/10.1002/asi.24646.

Hicks, D., Wouters, P., Waltman, L., de Rijcke, S., & Rafols, I. (2015). Bibliometrics: The Leiden Manifesto for research metrics. Nature, 520(7548), 429–431. https://doi.org/10.1038/520429a

Hirsch, J. E. (2005). An index to quantify an individual's scientific research output. Proceedings of the National Academy of Sciences, 46(46), 16569. INSPIRE. https://doi.org/10.1073/pnas.0507655102

Hummon, N. P., & Dereian, P. (1989). Connectivity in a citation network: The development of DNA theory. Social Networks, 11(1), 39–63. https://doi.org/10.1016/0378-8733(89)90017-8

Kleinberg, J. (2003). Bursty and Hierarchical Structure in Streams. Data Mining and Knowledge Discovery, 7(4), 373–397. https://doi.org/10.1023/A:1024940629314

Krause, Geoff, Madelaine Hare, Mike Smit, and Philippe Mongeon. 'Who Re-Uses Data? A Bibliometric Analysis of Dataset Citations'. arXiv:2308.04379. Preprint, arXiv, 8 August 2023. https://doi.org/10.48550/arXiv.2308.04379.

Lowenberg, D., Chodacki, J., Fenner, M., Kemp, J., & Jones, M. B. (2019). *Open Data Metrics: Lighting the Fire* (Version 1). Zenodo. https://doi.org/10.5281/ZENODO.3525349

Merton, R. K. (1970). Behavior Patterns of Scientists. Leonardo, 3(2), 213–220. https://doi.org/10.2307/1572092

Radicchi, F., Castellano, C., Cecconi, F., Loreto, V., & Parisi, D. (2004). Defining and identifying communities in networks. Proceedings of the National Academy of Sciences, 101(9), 2658–2663. https://doi.org/10.1073/pnas.0400054101

O'Neil, C. (2017). *Weapons of math destruction: How big data increases inequality and threatens democracy* (First paperback edition). B/D/W/Y Broadway Books.

Pardo-Guerra, J. P. (2022). *The Quantified Scholar: How Research Evaluations Transformed the British Social Sciences* (p. 272 Pages). Columbia University Press.

Park, Hyoungjoo, and Dietmar Wolfram. 'An Examination of Research Data Sharing and Re-Use: Implications for Data Citation Practice'. Scientometrics 111, no. 1 (2017): 443–61. https://doi.org/10.1007/s11192-017-2240-2.

Park, Hyoungjoo, and Dietmar Wolfram. 'Research Software Citation in the Data Citation Index: Current Practices and Implications for Research Software Sharing and Reuse'. Journal of Informetrics 13, no. 2 (2019): 574–82. https://doi.org/10.1016/j.joi.2019.03.005.

Radicchi, F., Castellano, C., Cecconi, F., Loreto, V., & Parisi, D. (2004). Defining and identifying communities in networks. Proceedings of the National Academy of Sciences, 101(9), 2658–2663. https://doi.org/10.1073/pnas.0400054101

Robinson-García, Nicolas, Evaristo Jiménez-Contreras, and Daniel Torres-Salinas. 'Analyzing Data Citation Practices Using the Data Citation Index'. Journal of the Association for Information Science and Technology 67, no. 12 (2016): 2964–75. https://doi.org/10.1002/asi.23529.

Rousseau, S., & Rousseau, R. (2017). Being metric-wise: Heterogeneity in bibliometric knowledge. *Profesional de La Información*, 26(3), Article 3. https://doi.org/10.3145/epi.2017.may.14

Sheehan, Nathanael, Federico Botta, and Sabina Leonelli. 'Unrestricted Versus Regulated Open Data Governance: A Bibliometric Comparison of SARS-CoV-2 Nucleotide Sequence Databases'. Data Science Journal 23, no. 1 (2024). https://doi.org/10.5334/dsj-2024-029.

Šubelj, L., van Eck, N. J., & Waltman, L. (2016). Clustering Scientific Publications Based on Citation Relations: A Systematic Comparison of Different Methods. PLoS ONE, 11(4), e0154404. https://doi.org/10.1371/journal.pone.0154404

Waltman, L., van Eck, N. J., & Noyons, E. C. M. (2010). A unified approach to mapping and clustering of bibliometric networks. Journal of Informetrics, 4(4), 629–635. https://doi.org/10.1016/j.joi.2010.07.002

Wilsdon, J., Allen, L., Belfiore, E., Campbell, P., Curry, S., Hill, S., Jones, R., Kain, R., Kerridge, S., Thelwall, M., Tinkler, J., Viney, I., Wouters, P., Hill, J., & Johnson, B. (2015). The Metric Tide: Report of the Independent Review of the Role of Metrics in Research Assessment and Management. Unpublished. https://doi.org/10.13140/RG.2.1.4929.1363

Appendix 4: Downstream effects of research outputs. Annex 1: CWTS Review

Review of the Report on the Downstream effects of research output

The pilot study under review here (Appendix 4) aims at improving the understanding and measurement of research output and impact beyond direct citations. The study aims at an analysis to the extent to which other forms of scholarly outputs and products, such as data reuse, policy influence, of knowledge transfer across disciplinary boundaries can be made clearer, in other words, look at the societal impact or relevance of academic work, beyond citation analysis. The authors of the report claim that traditional metrics (i.e. citation analysis) fall short and therefore create a team consisting of universities and private and semi-private data providers, in order to create a proof-of-concept indicator whereby citation network analysis and narrative information is brought together.

With the above in mind, the pilot study suffers from a number of issues that are discussed below:

- First of all, the pilot did not succeed in separating the *academic quality* realm from the *societal relevance* or 'impact' realm (as it is called in the UK). Although the pilot claims to distance itself from *traditional metrics*, the use of citation analysis is still functioning in the realm of academic quality, and if the authors meant this to work differently, they did not succeed in making this clear, let alone be convincing here. For example, the analyses in the report focusing on data citations end with citation counts, while it would have been more convincing if the actual users of the data (so the citers towards those datasets) would have been analyzed here, so for example while conducting data citation analysis on datasets produced by institutes in biomedicine, being cited by pharmaceutical companies. Then, the transfer from the academic to the societal realm would have been clearer.
- A second issue is the fact that the *impact of data* is being determined by looking at citations to datasets, as if the datasets were scholarly publications. This opens the conceptual debate of whether counting citations to datasets can be used in the same way as citation counts for publications, and particularly whether counting citations to datasets is the most fruitful approach to measure the reuse of datasets created and shared. Citations to datasets still produce indicators that are similar to traditional direct citations (something the authors claim they want to leave behind), and the pilot could have reflected about the limitations of this choice. An alternative could have been to focus on the mentioning of datasets in full texts of scholarly publications, in order to inquire in which ways scholars make use of existing datasets produced by other colleagues in their fields and look at the textual context of the mentioning of used datasets.
- Third, in some of the approaches the team has chosen to develop, results are discussed, and claims are made, for example regarding the cooperation with Elsevier, using Scopus and DataCite material, on the resulting IA generated narrative indicators, where I would like to have seen these being corroborated by, for example peer review by people in the field, whether or not such AI generated narrative indicators do make sense to people in the field.

- Fourth, the overall targets by the team were way too ambitious, by making use of research information of four different suppliers of research information (five, if one would distinguish in the Elsevier based approach also DataCite as a separate supplier). Several approaches are developed, in close collaboration with private and semi-private providers of information on research, (Digital Science working on REF2021 data, Elsevier/DataCite to analyze data citation flows, PLoS-Dataseer to analyze data reuse, and finally OpenAire to link data to publications by using text mining techniques) which all remain in the developmental stages, and none of them is really consolidated. Given that the timeframe and money involved in the pilot was rather limited, it would have been better that the pilot had chosen for one of the approaches and developed it that much further.
- Fifth, one wonders to what extent this would be helpful in any research evaluation procedure, in particular given the point made under my fourth point (see above), as I think neither of these developed approaches show enough maturity to be eligible for application in a research assessment procedure. This is also what I hinted at under my first comment, the instruments developed in the pilot could just have been taken one step further, to make it more of a near-finished tool, which could be considered for piloting in a real-life assessment context.
- Sixth, it was surprising to see the focus on private providers of information on research. In the context of the Barcelona Declaration, I would have expected a stronger focus on open research information, or at least, as much as possible. As far as I can see, only Datacite qualifies as such an organization, while also OpenAire and PloS do provide open data, but driven from a somewhat different commercially oriented business model compared to DataCite, while all other organizations involved are providers of proprietary information on research and show their good will by being involved in pilots such as these while sometimes the developed tooling disappears behind a paywall (as has happened with the development of Rare Diseases Dashboard), ending up with public money being used to develop privately-owned tools and data.

Thed van Leeuwen, CWTS, Leiden, 19th of November 2025

Appendix 4: Downstream effects of research outputs. Annex 1: CWTS Review – Team response

We thank the reviewer for their careful and substantive engagement with the pilot report, including the detailed annotated comments. The reviewer's expertise in bibliometrics and research evaluation is clear, and many of the issues raised were also actively discussed within the project team during the pilot.

We would, however, like to clarify several aspects of the context, scope, and purpose of the pilot that we feel were not fully reflected in the review.

First, this work was explicitly designed as an exploratory, proof-of-concept pilot, not as the development of a mature indicator intended for immediate use in research assessment. The aim was to test feasibility, identify infrastructural and methodological constraints, and explore how institutions might interpret early signals of downstream effects of research output. The pilot was therefore oriented toward learning and diagnosis rather than consolidation or deployment.

Second, the continued use of citation-based approaches was deliberate, despite their well-known limitations. Citation data remains one of the few signals (if not the only one) that is widely available, interoperable, and institutionally legible across disciplines and infrastructures. In this pilot, citation networks and dataset citations were used primarily as a tool to leverage and explore alternative, more contextual and narrative-driven approaches, rather than as an endorsement of their adequacy for measuring societal relevance. For instance, rather than using citation counts, we used citation networks to explore how dimensions such as leakage across academic fields (example shown in the report); this approach can be applied to other dimensions, such as sources of citations, identifying cross-sectoral impact, etc. For that purpose, institutions developed a list of key topics they would be interested to report on.

Relatedly, the reviewer is correct that the pilot does not fully resolve the boundary between academic quality and societal relevance. This was not an oversight but a central finding of the work. One of the pilot's conclusions is that current infrastructures, identifiers, and data flows make this distinction difficult to operationalise in practice, particularly at institutional scale. Making these tensions visible is an outcome of the pilot.

Dataset citations were treated as partial and imperfect signals of data reuse, not as definitive measures. More context-sensitive approaches, particularly full-text analysis of data mentions and reuse contexts, are widely recognised as necessary but remain difficult to implement at scale due to access, interoperability, and sustainability constraints. Several providers participating in the pilot are actively exploring such full-text–based approaches with LLMs, and the pilot provided a structured setting in which these experimental methods could be tested, discussed, and refined through institutional feedback. A key contribution of the pilot is therefore both to demonstrate how much reuse and downstream influence remains invisible under current practices, and to help steer the development of more nuanced methods capable of addressing these gaps.

AI-generated narratives were explored as experimental, assistive tools to support human interpretation of complex citation and reuse traces, not as evaluative outputs in their own right. Formal validation of these outputs was out of scope for the pilot, given time constraints and the fact that participating providers developed their approaches independently rather than as a single integrated system. The pilot was intended to test feasibility and provoke discussion, rather than to establish validated narrative indicators.

The breadth of approaches and engagement with multiple data providers was also intentional. Rather than consolidating a single method, the pilot aimed to map the landscape of existing infrastructures, reveal differences in maturity and coverage, and expose dependencies on proprietary systems. Given the limited timeframe and resources, this breadth was prioritised over depth. A known limitation of all the pilots, in general, has been the early choice of engaging primarily with established providers, with whom institutions already had relationships.

Finally, we share the reviewer's concern regarding reliance on proprietary research information. It is important to note, however, that several decisions shaping the pilot were constrained by factors outside the control of the project team, rather than reflecting deliberate methodological choices. In particular, initial engagement with proprietary data sources reflected the availability and maturity of existing infrastructures at the time, as well as the independent development paths of participating providers. As the pilot progressed, we actively steered the work toward open and publicly accessible infrastructures, most notably through the exploration of OpenAlex for citation network analysis. The pilot therefore reflects both the reality that institutions currently operate within mixed ecosystems and an explicit attempt to assess how far open infrastructures can support such analyses, helping to identify remaining gaps and inform more grounded recommendations for future development.

Many of the reviewer's critiques align closely with the pilot's own conclusions. Where limitations are identified, we see these as evidence of the need for further, more focused work rather than as shortcomings of an exploratory exercise. We hope this response clarifies the intentions, constraints, and contributions of the pilot.