Appendix 2

Open Research Indicators: Report of a pilot on the Openness of data

Authors

Contributors: Kirsty Merrett †, Jade Godsall (joint lead authors), Christopher Warren, Lavinia Gambelli. All contributors from the University of Bristol.

† Correspondence should be addressed to Kirsty Merrett: J.K.Merrett@bristol.ac.uk

Executive summary

This report summarises the key findings of UKRN's Open Research Indicators Pilot on the Openness of Data. It details the specifications agreed by pilot partners; the methodologies used by the providers to supply data (where understood and permitted); the methodologies used by the pilot to test the data from providers; issues with locating and testing the Openness of data; issues with the accuracy and reliability of the data, and why this is problematic for developing an indicator.

To undertake the pilot, we looked at 470 DOIs of articles published in 2023 from data entries in three providers' datasets, checked the Data Availability Statements (DAS) fields, and then visually assessed and audited these individually against the article to determine if a DAS had been detected correctly, thus checking the accuracy of provider's data against the article. In some complex DAS instances, we reviewed an article across all three providers' datasets to establish whether any individual tool/methodology/algorithm produced data about the DAS which was more accurate and reliable than others.

The pilot found that overall, providers' tools *can* consistently identify and extract DAS from articles, but they are *unable* to reliably categorise cited datasets, such as classifying all named repositories, or access levels. Additionally, where multiple datasets, third party data, data and code, and/or multiple access levels are cited within articles, tools have *noticeable limitations* when differentiating between data generated by the project and other datasets (for example, datasets in references, or third-party data). Finally, the tools are *unable to check the dataset* for Openness and instead analyse and assess Openness from the wording of the DAS. This is of importance, as preliminary small-scale analysis of University of Bristol data from provider's datasets indicates that language used when constructing a DAS has implications for the tools to identify Openness in DAS, which is problematic as the language can be determined by discipline, journal, or by researcher.

The report will explain issues with providers' datasets and assess the tools for their utility against the pilot's specifications, where possible. These issues are identified thematically to highlight suggested areas for improvement and to emphasise how a collaborative approach with funders, publishers, institutions and researchers is the most practical approach to accelerate the uptake of open research practices. More specifically, with the Open Research Indicators 'data' Pilots, collaboration is necessary to increase the availability of data, and thus; to be able to measure its Openness; to measure the FAIRness of data; the prevalence and quality of DAS; and the downstream effects of data sharing. As this pilot also demonstrates a

conceptual and practical alignment with the 'non-data' pilots, it reinforces the urgent need for sector-wide agreement on common standards and their consistent application. The recommendations of the top-level report are expected to underscore this necessity, laying the groundwork for the future development of robust, reliable indicators across the research environment. The report will conclude by detailing the complex and overlapping issues which hinder the development of an ethical, reliable, and accurate monitoring process at this stage.

Background and aims

The concept of the openness of data originates from the principle that research data should be shared 'as Open as possible, as closed as necessary.'[1] It acknowledges that while some research data can be made freely available online to use and redistribute without restriction, not all data can or should be shared openly due to ethical, legal, or confidentiality concerns. Rather than placing openness on a hierarchal scale, this approach advocates for responsible and ethical data sharing practices.

Efforts to encourage data sharing have been reinforced through funder, institutional, and editorial open research policies, which require researchers to share their data as openly as possible in established repositories. Despite this, several studies show that the proportion of researchers who share their data remains low.[2] Hrynaszkiewicz, Harney and, Cadwallader's (2021) study assessing researchers' needs and priorities for data sharing states, 'best practices for data sharing are adopted by a minority of researchers in their publications. Problems with effective research data sharing persist and these problems have been quantified by previous research as a lack of time, resources, incentives, and/or skills to share data'.[3] Gabelicia, Bojčić, and Puljak's (2022) mixed-methods study on data sharing compliance in Open Access articles published in journals with mandatory DAS, discovered that the compliance rate of researchers who stated they would share data upon request was the same as researchers who did not include a DAS.[4] Many researchers still use 'contact the author' or 'data shared upon request' even when data sharing is mandated and enforced by publishers and funders. Yet, requests for data frequently have a low success rate which varies between disciplines: Tedersoo et. al. (2021) demonstrated that when contacting the authors of 310 papers, data was obtained with a range of 27.9 – 56.1% across distinct research fields, with social sciences, psychology, and humanities having lower success rates than natural

[1] European Commission, Directorate-General for Research & Innovation (2016). *Horizon 2020 guidelines on FAIR data management (Version 3.0).* Publications Office of the European Union. p. 4. Available at: https://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-data-mgt_en.pdf

[2] The background and aims section discuss some of these studies in further detail. Also see: Watson, C., (2022). Many researchers say they'll share data — but don't. *Nature*, 606, p.853. Available at: https://doi.org/10.1038/d41586-022-01692-1; Hamilton, D.G., Page, M.J., Finch, S., Everitt, S., Fidler, F., (2022). How often do cancer researchers make their data and code available and what factors are associated with sharing? *BMC Medicine*, 20, (438), pp. 1-12. Available at: https://doi.org/10.1186/s12916-022-02644-2; Thoegersen, J.L. and Borlund, P. (2022), Researcher attitudes toward data sharing in public data repositories: a meta-evaluation of studies on researcher data sharing, *Journal of Documentation*, 78, (7), pp.1-17. Available at: https://doi.org/10.1108/JD-01-2021-0015.

[3] Hrynaszkiewicz, I., Harney, J. and Cadwallader, L., (2021). A Survey of Researchers' Needs and Priorities for Data Sharing. *Data Science Journal,* 20, (31), pp. 1-16. Available at: https://doi.org/10.5334/dsj-2021-031.

[4] Gabelicia, M., Bojčić, R., Puljak, L. (2022). Many researchers were not compliant with their published data sharing statement: a mixed-methods study. *Journal of Clinical Epidemiology*, 150, pp. 33-44.

sciences.[5] These patterns underscore a key challenge for the field of open research - how do we meaningfully measure and evaluate the openness of shared research data, if a substantial proportion of research data is not shared?

The academic publishing landscape is centred around journal articles and offers limited recognition or reward for publishing standalone datasets, and DAS in published articles has become the primary access point for data, and is, therefore, a logical way to evaluate data openness at this stage. However, this reliance makes the DAS a single point of failure, as we will explore in the methods section of this report.

Brief overview of key literature on DAS and Open data

In 2023, Digital Science, Springer Nature, and Figshare collaborated on the white paper *The state of Open data*, based on responses from over 6,000 researchers worldwide from the state of open data survey.[6] The survey accessed practices and attitudes toward open data sharing and highlighted several obstacles: lack of specialised support for planning, managing and sharing research data, insufficient credit and incentives, and limited awareness or adoption of Large Language Model (LLM) AI programmes for data collection, processing, and metadata generation. Two additional reports were published in 2024; *From Theory to Practice* compiled case studies and commentaries from libraries, publishers, funders, and industry stakeholders, while *The Global Lens* focused on open data perspectives in Ethiopia, Japan, and the United States.[7] Both reports showcased real-world initiatives, such as publishers implementing data policies and tailored support from university libraries, that are advancing open research. Yet, persistent issues remain, including inconsistent policies, regional inequalities, and inadequate infrastructure. Addressing these concerns will require international coordination around standardising data management and sharing practices to make open data both equitable and practical.

A further report, *Bridging Policy and Practice in Data Sharing (2024),*[8] explored these issues from the perspective of a global adoption of open data policies and LLM tools, and analysed over 6,000 Data Availability Statements (DAS) from Springer Nature publications between 2019-2022. Using the DataCite Corpus and Digital Science's Dimensions, DAS were divided into seven categories based on common text and data terminology – including 'data available on request', 'data availability not applicable', and 'publicly available in a repository'. The analysis identified that countries with strong open research mandates showed higher levels of compliance and engagement, driven largely by funder and institutional support. However, it is important to note that the study had limitations: the scope was confined to Springer Nature publications, the accuracy and categorisation of the DAS were not manually verified, and there

[5] Tedersoo, L., Küngas, R., Oras, E., Köster, K., Eenmaa, H., Leijen, Ä., Pedaste, M., Raju, M., Astapova, A., Lukner, H., Kogermann, K., Sepp, T., et. al.., (2021). Data sharing practices and data availability upon request differ across scientific disciplines. *Scientific Data*, 8, (192), pp. 1-11.

[6] Hahnel, M., Smith, G., Schoenenberger, H., Scaplehorn, N. and Day, L., (2023). *The State of Open Data 2023*. Digital Science, Figshare and Springer Nature. Available at: https://doi.org/10.6084/m9.figshare.24428194.

[7] Adams, J., Barbosa, S., Campbell, A., Campbell, R., Chandramouliswaran, I., Chodacki, J., Clement, E., Curtin, L., Hahnel, M., Day, L., Holmes, K., Jones, B., Koers, H., Linacre, S., MacCallum, C.J., McIlwaine, P., Osório, J., Puebla, I., Ross-Hellauer, T., Sansone, S.-A., Stall, S., Grant, R., Van Gulick, A. and Wood, J., (2024). *From theory to practice: Case studies and commentary from libraries, publishers, funders and industry*. Digital Science, Figshare and Springer Nature. Available at: https://doi.org/10.6084/m9.figshare.25232899; Bahl, R., Gunawardena, K., Valdez, M. and Hahnel, M., (2024). *The Global Lens: Highlighting national nuances in researchers' attitudes to Open data*. Digital Science, Figshare and Springer Nature. Available at: https://doi.org/10.6084/m9.figshare.25569453.

[8] Hahnel, M., Smith, G. and Campbell, A., (2024). *The State of Open Data 2024: Special Report – Bridging policy and practice in data sharing*. Digital Science, Springer Nature and Figshare. Available at: https://doi.org/10.6084/m9.figshare.27337476.

was no consideration of the variability or clarity of DAS implementation across journals. The adoption of standardised templates without controlled vocabulary risks giving false or inaccurate quantifications on the openness of data.

Montague-Hellen and Montague-Hellen's paper *Publishers, funders and institutions: who is supporting UKRI-funded researchers to share data?* (2023) demonstrates that policy mandates have increased the inclusion of DAS in research papers.[9] However, the inclusion of a DAS alone does not guarantee that the data is accessible, reusable or, that it is curated with robust data-sharing practices. As they observe, 'recent research has shown that journals with strong data access policies have high compliance regarding the existence of a data availability statement, but that journal policies may be standardizing the use of statements which do not make it easy for other researchers to obtain the data'.[10] This concern is also raised in Colavizza et. al.'s 2020 article *The citation advantage of linking publications to research data*, which analysed 531,889 PLoS and BMC Open Access articles to develop an automated DAS classification system.[11] Their methodology demonstrates the systemic reliance on the DAS as an assessment for the openness of data, despite its lack of standardisation and clarity.

The limitations of DAS as a reliable mechanism for data sharing are further underscored in Tedersoo et. al.'s 2021 paper *Data sharing practices and data availability upon request differ across scientific disciplines*.[12] The authors evaluated the data availability in 875 articles across nine scientific disciplines and contacted corresponding authors of 310 studies that claimed data was 'available on request'. Only 39.4% provided full or partial data, while 19.4% declined, and 41.3% did not respond. While the study recommended the adoption of standardised data sharing policies, consistent language, and clear templates to improve data sharing practices across disciplines, it is clear that in the current research climate, we cannot reliable indicate if data is truly open and accessible via an uncontrolled reporting statement, an issue which is evidence in initiatives like this pilot and the TIER2 project, below.

The EU funded TIER2 project tackled the issue of data sharing in the context of research reproducibility. Through seven pilot initiatives, TIER2 developed tools such as Reproducibility Management Plans and decision aids. One of its key findings was that low data sharing rates continue to hinder reproducibility. The project recommended a cultural shift towards valuing quality over quantity in research outputs, cross-disciplinary understanding, and creating the incentives, guidelines, and infrastructure needed to support effective data sharing across the board.[13] Outputs from Pilot 7 *Editorial issues to increase data sharing*,[14] detail the complex dependencies of workflows between article submission, publication, and associated data publication show that publishers are aware of the importance of DAS for data sharing and reproducibility and are developing practical ways to address this.

This can be seen further in the DAS quality workshop organised by the Data Policy Standardisation and Implementation Interest Group, which brought together 20 stakeholders

[9] Montague-Hellen, B., and Montague-Hellen, K, (2023) Publishers, funders and institutions: who is supporting UKRI-funded researchers to share data?, *Insights*, 36, (4), pp. 1–17; Available at: https://doi.org/10.1629/uksg.602.

[10] Montague-Hellen, B., and Montague-Hellen, K, "publishers, funders and institutions…", p. 2.

[11] Colavizza, G., Hrynaszkiewicz, I., Staden, I., Whitaker, K. and McGillivray, B., (2020). The citation advantage of linking publications to research data. *PLoS One*, 15, (4), e0230416. Available at: https://doi.org/10.1371/journal.pone.0230416.

[12] Tedersoo, L., Küngas, R., Oras, E., Köster, K., Eenmaa, 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, (192), pp.1-11. Available at: https://doi.org/10.1038/s41597-021-00981-0.

[13] For further information and publications regarding the TIER2 project, see: https://tier2-project.eu/

[14] TIER2 Project. *Pilot 7 — Editorial workflows to increase data sharing*. Available at: https://tier2-project.eu/pilots/7.

from publishing and research organisation.[15] The workshop identified widespread issues with DAS quality, including lack of standards, poor resources, unclear author expectations, and the prevalence of vague statements like 'data available on request'. Among their proposed solutions were: earlier intervention in the publication workflow to assess DAS quality (which is partially addressed in a TIER2 pilot);[16] improvement of infrastructure and tools, such as standardised checklists for DAS evaluation and AI tools to verify data links; better alignment of publisher and funder policies with institutional training and support.

The key takeaway from research across the field is that while data-sharing policies and data availability statements are increasing, through both publisher and funder policy and researcher awareness, their effectiveness in facilitating data access and reuse remains limited by issues like standardisation, infrastructure, discrepant timelines for publication and data publications, and the terminology used by both publishers and researchers.

Methods

The following institutions and providers partnered in the pilot, volunteering time and expertise.

Pilot	Participating institutions	Participating providers
Openness of Data	**Bristol (Lead)**, Edinburgh, Exeter, Glasgow, Leicester, Liverpool, Reading, Sheffield, Surrey	Digital Science, Elsevier, OpenAIRE (did not return data), PLOS-DataSeer

Participating institutions used the INORMS SCOPE framework to consider the aims of the pilot. A SCOPE template was completed by the pilot lead, the University of Bristol, with participating institutions providing institutional-specific reflections and suggested areas of interest.

In monitoring open research, key considerations were:

- Transparent methodology - UNESCO Recommendation on Open science recommends 'the monitoring of open science should be explicitly kept under public oversight, including the scientific community, and whenever possible supported by Open non-proprietary and transparent infrastructures', including 'the necessary metadata for their efficient assessment'.[17]
- Ethical adoption - UNESCO Principles of Open Science Monitoring indicators must be 'relevant for the specific monitoring task at hand' and careful consideration should be applied before reusing the indicators for other purposes.[18]
- Collective commitment to enhancing open research practices - 'It is necessary for all of us involved in performing, managing, funding, and communicating research to commit to improving practices in our own organizations and disciplines as well as more broadly. Key areas of focus include institutional efforts to sustain research environments conducive to integrity, greatly expanded sharing of data and code, more

[15] Data policy standardisation and implementation IG co-chairs, (2023), *Report from DAS quality Workshop*. Available at: https://docs.google.com/document/d/1EYYzS71h58ZZn1bA-4fFDM0l9cXXEwn5Fi5zCHsJjY8/edit?tab=t.0
[16]TIER2 Project. (2024). *Flowchart - Editorial Reference Handbook*. Available at: https://osf.io/245zj. TIER2 Project (n.d.) *Editorial Reference Handbook*. Available at: https://publishers.fairassist.org/.
[17] UNESCO, (2021). *Recommendation on Open Science*. Available at: https://doi.org/10.54677/MNMH8546.
[18] Open Science Monitoring Initiative, (2025), *Principles of Open Science Monitoring*. Available at: https://zenodo.org/records/17143859.

complete reporting of results, more responsible approaches to scholarly publishing, better understanding of the causes and consequences of breaches in integrity, and clearer standards for authorship'.[19]

Using these guiding principles, the partners identified the main goals of this pilot as follows:

- To define the openness of data by collaborating with relevant services and working groups to create a standardised terminology and universal monitoring plan that can be applied across multiple locations.
- Measure and identify the extent to which cited datasets are openly available, on a spectrum from wholly and immediately open with a permissive license, through minimal metadata available after an embargo, to not available.

All partners' workshop 2024 – how do you develop an indicator for Openness of data?

In February and March 2024, the leads for this pilot, partner institutions and solution providers met to discuss and agree upon the detailed specifications of the openness of data indicator. These discussions covered defining terminology, identifying the data, determining the methodology, and agreed ways of working to guide the indicator development and collaboration across pilots.

From our current understanding, an effective indicator of the openness of data must be able to differentiate the degree to which data cited in a DAS is open. This ranges from fully open and freely accessible data, to data that is not available for sharing. It's important to note that these degrees of openness are not hierarchical, as appropriate access levels for data sharing must account for external factors such as ethical considerations and commercial sensitivities. For this pilot, institutional leads and partners agreed on the following definitions for the degrees of openness:

Open data	freely available for re-use and redistribution by anyone without any registration
Technically Open data	freely available for re-use and redistribution by anyone, but requiring some form of registration
Embargoed data	freely available for re-use and distribution by anyone, but after an embargo period
Controlled data	available 'on request' through standard data sharing agreement process
Author controlled data	available 'on request' through contacting the author directly
Closed data	data exists in a repository but not available to be shared
Dark data	exists outside a repository, cannot be shared or linked to

It was agreed that the DAS would serve as the primary focus for evaluations, based on the following considerations:

- **Technical feasibility**: Initial discussions with solution providers revealed that the majority of tools available to them were designed to scrape information from PDF documents. One solution provider indicated that extraction from repositories could be possible, but this was only explored after the decision to use the DAS as a starting point, and had commercial limitations.

[19] National Academies of Sciences, Engineering, and Medicine, (2017). *Fostering Integrity in Research*. Washington, DC: The National Academies Press. Available at: https://doi.org/10.17226/21896.

- **Project timeline constraints**: Given the limited timeframe for solution providers to generate outputs, it was necessary for them to utilise existing their tools and infrastructure, further supporting the focus on the DASs.
- **Academic publishing limitations:** The culture of academic publishing continues to prioritise and reward the publication of research articles over standalone datasets. This cultural bias reduces the availability and visibility of datasets that are independent of accompanying articles, making the DAS the only feasible access point.
- **Institutional repository limitations:** Many institutional repositories lack comprehensive internal linkages between publications and cited datasets. While not all institutions maintain dedicated research data repositories, all have scholarly works repositories, which include publications with accompanying DOIs and or PDFs they're able to extract the metadata for. These links and PDFs offered an accessible way for solution providers to access publication's DAS.
- **Commercial and legal constraints**: Solution providers faced limitations in accessing publications' metadata and/or PDF due to publisher agreements and if an article was Open Access or paywalled. Additionally, there were restrictions on their ability to disclose the proprietary methodologies and algorithms of their tools.

In retrospect, we acknowledge the methodological limitations of analysing the openness of data through a reporting statement instead of the cited dataset. These challenges, in particularly the variability and interpretability of DAS content, are addressed in the Discussion section.

Identifying the data – data pilots' collaboration

We required datasets from a current and complete academic year, to maximise the likelihood that data citations would be captured in the DAS. It was agreed that the dataset for evaluation would consist of journal articles published in 2023. Given the thematic overlaps between the pilots on FAIRness and openness of data and DAS, a common dataset template was agreed for all pilots. Solution providers were encouraged to adopt a staggered approach in their analysis, which included:

- Identification and extraction of the DAS content
- Evaluation of the completeness and quality of the DAS for that pilot[20]
- Assessment of the FAIRness and Openness of the cited dataset for those pilots

The staggered approach and thematic overlap of the pilots on the FAIRness and Openness of data, and on Data Availability Statements, was detailed in Pilot's specification documents as participating institutions agreed it was necessary that to produce an indicator to assess the FAIRness and Openness of data, preliminary work to identify and evaluate the DAS (detailed in the DAS pilot's specifications) needed to be completed.

The pilots leads and partner institutions agreed that each institution would upload a sample dataset to a shared Figshare project space. This workspace would serve as a centralised location for analysis by the solution providers supporting all three pilots. The sample datasets comprised of institutional-authored journal articles published in 2023 extracted from each institution's Current Research Information System (CRIS). To ensure consistency across datasets, the following mandatory fields, optional fields, and criteria were established:

[20] As DAS is the logical starting point, evaluating the prevalence and quality of DAS is a necessary precursor to producing indicators of the FAIRness and Openess of data; future examinations of the terrain should be structured so the work of Pilots 1 and 2 build upon the work of Pilot 4/5.

Mandatory fields	Description	Rationale
LocalID	A local repository id number	Useful for local reference
DOI	The publisher DOI	To provide the persistent link to the published item
PageLink	The link to repository splash page or similar	Useful for access to repository added bibliographic or other metadata about an item if required
Type	Item type	If we are extracting only articles then this is implicit. But if analysing more output types in future then type should be provided
Year	Year of publication	If we are extracting only 2023 items, then implicit. But if date scope gets widened then this should be provided
OrgUnit	The university faculty/college/school/department. This is the minimum viable unit that each RO can provide.	If analysing data by discipline is within scope, then this will be mandatory

Optional fields	Description	Rationale
FileURL	A link to a repository file - this could be the AAM or the publisher VOR depending on OA type	Useful for immediate access to repository held file if required
FileVersion	If there is a repository file (FileURL) then an explicit value flagging the version (publisher VOR or Author Accepted Manuscript)	Would facilitate some analysis of DAS prevalence at the various stages of the publication cycle if required
DAS	Local system value for DAS prevalence	Some institutions record the prevalence of a DAS and for pilot 4 it will be useful to include. Set to optional because not every institution has it available and possibly not every pilot requires it.
OAType	The Open Access type	This could facilitate some analysis of the data by Open Access type if required
Funding	Funder acknowledgements recorded in the local system	This could facilitate some analysis of the data by funder if required

Criteria	Rationale
Must be Open Access	This is to ensure supplier can easily access a file
Must be published in 2023	
Number of items in dataset	All of 2023 for initial count. Subset e.g. 300 can be chosen if manual detailed work needed to demonstrate solution.
Type must be article only	

In May 2024, the solution providers were provided with specifications for each pilot and datasets of 2023 published articles extracted from institutional repositories.[21]

Manual checks methodology

Leads for this pilot conducted a specification, classification, and terminology mapping exercise across data returns from Digital Science, Elsevier, and PLOS-DataSeer. This exercise was guided by the pilot specification (Annex 1), and aimed to cross match solution provider terminology against field terminology and the degrees of openness terminology provided. For manual checking, the leads used the common protocol for manual checks (see Overview Report. Annex 5), a template provided by UKRN. Initial assessments indicate that the returned data exhibited moderate levels of validity and reliability. To explore this further, a set of example DOIs was selected and analysed across all three providers to determine whether similar patterns of success and limitation emerged across methodologies. In total, 225 Digital Sciences records, 150 PLOS-DataSeer DOIs, and 101 Elsevier DOIs were manual reviewed. The DAS was used as the primary reference point for evaluation, as each solution provider's methodology relied on the reporting statement. A summary of results is provided in the Finding section, with a more detailed analysed available in the Discussion section.

[21] Warren, C., Godsall J. (2025). *Specification for Pilot 2 The Openness of Data v1.0*. Available at: https://osf.io/7bwvk and at Annex 1

Findings

The pilot leads received the data analysis from 3/4 of the solution providers, between October-December 2024. Due to commercial constraints, we are unable to provide a comprehensive explanation of the solution providers methodologies or the tools used in this paper. A snapshot of the methodologies used by the solution providers is provided in the table below:

	Digital Science	**Elsevier**	**PLOS-DataSeer**
Articles analysed	Gold and Green OA	Gold OA	Gold OA, CC BY
Number of articles returned	2672	299	1187
Where DAS extracted	HTML, PDF	Formal subsections of peer review articles	PDF
Information extracted from	Only DAS	In paper, DAS, and dataset(s) cited	In paper, DAS
Tools used	Dimensions	Research Data Monitor	DataSeer regex, DataSeer gazetter
Repository list	Edited list of repositories from www. re3data.org.	Used databases: Elsevier's Research Data Monitor Zenodo NCBI Database of Genotypes and Phenotypes (dbGaP) NCBI Gene NCBI Biosample NCBI Sequence Read Archive NCBI Bioproject NCBI GenBank NCBI Nucleotide	Public Library of Science (2022). 'OSI-Repository-List_v1_Dec22.xlsx ', *PLOS Open Science Indicators. Public Library of Science. Dataset*. https://doi.org/10.6084/m9.figshare.21687686.v9

		NCBI ClinVar NCBI Assembly Protein Data Bank European Nucleotide Archive (ENA) GitHub And LLM (GPT-4 via Azure OpenAI)	
Additional information	Returned ‘positive’ and ‘negative’ (including null) results.	Returned only ‘positive’ results, where a DAS and repository had been identified.	Returned ‘positive’ and ‘negative’ (including null) results.

Due to the varying methods and terminologies used, the commercial limitations, type of articles analysed, and the number of returned results, it is neither feasible or appropriate to directly compare or contrast the approaches use by the different providers. The findings presented are intended solely to address the challenges and successes of producing an indicator for accessing the openness of data using Machine Learning tools.

The tables below show a summary of the specifications successfully achieved by each provider and the outcome of the initial manual checks, where the content and links of 10 DOIs were checked for their validity and reliability. See the Discussion section for further analysis on the manual check of 470 DOIs across the three providers response datasets.

Digital Science

Specification	Validity	Reliability	Overview
Location (by repository) of any data, software, code or other supporting materials cited in DAS	8/15 repositories correctly identified (this includes databases but not programmes or archives)	3/10 DOIs correctly identified all named repositories with no errors	Can only identify repositories named in the DAS not in the URLs/DOIs. Potential to distort results. For example, GitHub will be identified as repository if named as such with an accompanying DOI from Zenodo. Uses re3data.org Registry of Research Data Repositories to identify repositories. However, University of Bristol's institutional repository was not identified. Possibility of undermining institutional/small non-commercial repositories as they will not be picked up in a measure for openness.
Location (by persistent identifier) of any data, software, code, or other supporting materials	14/67 URLs/DOIs identified correctly (this means the data was not corrupted and includes URLs/DOIs cited in DAS and in paper)	2/10 DOIs correctly identified and correctly extracted (not corrupted) all URLs and DOIs.	Can only extract a URL/DOI from a DAS. Cannot extract URLs/DOIs from hyperlinks unless the hyperlinked text is cited in the DAS as the full URL/DOI. If URL/DOI cited with a reference at the end of the link, tool will extract it as part of the URL/DOI.
Openness type: Open data / Technically Open data / Embargoed data / Controlled data / Author Controlled data /Closed data / Dark data	N/A	N/A	Did not use terminology provided. Algorithm is only looking at the DAS or Dimensions page, not the dataset landing page so cannot decipher the degree of data openness. Only if data shared in repository, on request, in paper or provided as supplementary information, or not public.

Elsevier

Specification	Validity	Reliability	Overview
Location (by repository) of any data, software, code, or other supporting materials cited in DAS	10/10 repositories correctly identified (this includes databases but not programmes or archives)	10/10 DOIs correctly identified all named repositories with no errors	Identified repositories and URLs by using Research Data Monitor (RDM) catalogued records and APIs from external data repositories, such as Zenodo, NCBI, GitHub etc. The data provided only included articles for which a DAS and repository had been identified, which will distort these results.
Location (by persistent identifier) of any data, software, code, or other supporting materials	10/10 URLs/DOIs identified correctly (this means the data was not corrupted and includes URLs/DOIs cited in DAS)	10/10 DOIs correctly identified and correctly extracted (not corrupted) all URLs and DOIs.	Identified repositories and URLs by using Research Data Monitor (RDM) catalogued records and APIs from external data repositories, such as Zenodo, NCBI, GitHub etc. The data provided only included articles that a DAS and repository had been identified which will distort these results.
Any supporting materials present in the article not included in DAS	N/A	N/A	We did not have time to investigate this specification in further detail.
Openness type: Open data / Technically Open data / Embargoed data / Controlled data / Author Controlled data / Closed data / Dark data	10/10 records checked returned a degree of openness	5/10 DOIs had degree of openness correctly identified	Our terminology was used. The degree of openness was based on the authors description in the DAS, not the dataset itself. Only one degree of openness was given which does not account for the citation of multiple datasets or points of access.

PLOS-DataSeer

The anonymised PLOS-DataSeer dataset is provided in Overview Report Annex 4.

Specification	Validity	Reliability	Overview
Location (by repository) of any data, software, code, or other supporting materials cited in DAS	10/15 repositories correctly identified (this includes databases but not programmes or archives)	3/10 DOIs correctly identified all named repositories with no errors	Identifies named repositories in DAS and from URL/DOIs provided. Potential to distort results. For example, GitHub and Zenodo will be identified as repositories for one dataset if DAS states it is deposited in GitHub but gives a Zenodo DOI. Uses own dataset to identify repositories (see Public Library of Science (2022) PLOS Open Science Indicators. Figshare. Dataset (version 8). https://doi.org/10.6084/m9.figshare.21687686) which includes commonly used repositories and popular domain repositories but is not an exhaustive list. Possibility of undermining institutional/small non-commercial repositories as they will not be picked up in a measure for openness. Provided columns for data and code repositories but cannot decipher the difference. Every data repository identified was duplicated as code repository, even if code was not mentioned in the paper.
Location (by persistent identifier) of any data, software, code, or other supporting materials	16/67 URLs/DOIs identified correctly (this means the data was not corrupted and includes URLs/DOIs cited in DAS and in paper)	5/10 DOIs correctly identified and correctly extracted (not corrupted) all URLs and DOIs.	Cannot extract URLs/DOIs from hyperlinks unless the hyperlinked text is the full URL/DOI. Only extracts a DOI for data and URL for code.
Any supporting materials present in	N/A	N/A	Needs further investigation as only two of the sample had data cited in the paper that was not

Specification	Validity	Reliability	Overview
the article not included in DAS			included in the DAS/had no DAS. The algorithm extracted a URL from 1/2. How would an algorithm identify the difference between in paper data sharing, citing, and methodology discussion?
Openness type: Open data / Technically Open data / Embargoed data / Controlled data / Author Controlled data / Closed data / Dark data	N/A	N/A	Did not use terminology provided. Used Colavizza et. al. terminology. See Table 1, in Colavizza, H., Hrynaszkiewicz, I., Staden I., Whitaker, K., and McGillivray, B., (2020). The citation advantage of linking publications to research data. *PLOS One*, 15(4). E0230416. Available at: https://doi.org/10.1371/journal.pone.0230416 The tool is only looking at the DAS or paper, not the dataset landing page so cannot decipher the degree of data openness. Only if data shared in repository, on request, in paper or provided as supplementary information, or not available.

Discussion

Pilot findings

The manual checks conducted by the pilot leads have determined that it is not currently possible to create an ethical, transparent, and reliable indicator to assess the openness of data. We have discussed these findings thematically to highlight suggested areas for improvement and the benefit of adopting a collaborative approach for future indicator work.

Specifications

Providers were unable to meet many of the specifications required by this pilot, and focused on those which they could map to existing tools:

- Location (by repository) of any data, software code, or other supporting materials cited in DAS;
- Location (by persistent identifier) of any data, software, code, or other supporting materials;
- Openness type: Open data / Technically Open data / Embargoed data / Controlled data / Author Controlled data / Closed data / Dark data.

An evaluation of these met specifications identified similar challenges and issues in the algorithm and tools used:

- **Repository identification:** although the extraction methods differed (only from DAS, URLs/DOIs & DAS, catalogue records and API from repositories), each solution provider worked from their own list of research data repositories. This list is not complete, and most institutional repositories were not identified.
- **Persistent identifier identification:** two-thirds of solution providers tools could not extract URLs and DOIs from hyperlinks unless the hyperlinked text was the full URL/DOI. This is problematic given hyperlinks are acceptable formatting in many publisher's data policies for DAS.
- **Evaluating the DAS not the data:** All solution providers algorithm categorised the data based on the DAS, not the cited dataset. This challenges categorisation on several points:
    - DAS are free text; there are no set templates or controlled language.
    - Certain words can produce false negatives/positives, for example, if a restricted access dataset is described as on request via application, this would be categorised the same as a statement that reads on request from author. Likewise, if a statement reads that an article is not 'openly available,' this could be categorised as being in a repository.
    - 'On request' as a category is too vague and could negatively impact certain disciplines as in the RDM community, 'on request' is typically read as on request from author and not compliant with data policies.
    - Two-thirds solution providers gave each DAS one 'openness' categorisation which does not account for the citation of multiple datasets.
- **Discipline disparity:** if the tools require a named repository and persistent identifier include in the DAS to categorise it as 'in a repository,' this leads to disparity across disciplines and how they cite their data. For example, complex cohort data studies with defined access routes, archives, and museums do not necessarily store data in a

'repository' as defined by the RDM community and use 'recognised' persistent identifiers like a DOI. These statements were often categorised as 'on request.'

Although the successful extraction rates of DAS were high across the board, ranging from 89-94%, categorisation accuracy (validity and reliability of returned data) was average ranging from 50-57%, and in specific areas, such as defining the openness of a cited dataset, was unattainable.

Openness type terminology

The terminology provided in our specification defining 'openness' was only used by one solution provider. Two solution providers used their own terminology guide which was basic in its categorisations in that it only reported if data was shared in a repository, on request, in paper, or in supplementary material. The solution provider that used our terminology noted that they did not feel like they had enough information from our specification document to make an informed decision for each record. All the solution providers provided one openness type based on the DAS; this was not a true categorisation of the cited dataset itself, nor did it account for statements which had multiple datasets cited. In the long-term, if agreed standardised terminology is not adopted, it makes producing a clear and reliability indicator that can be understood and cross-referenced across systems, tools, and platforms untenable.

Data cleaning and integrity

Across all service providers there were issues with data integrity and data cleaning of filters and fields to ensure that data was in the best condition for analysis. The inclusion of corrupt or missing data impacted accuracy ratings as it hindered machine and human readability. The limited timeline likely led to basic data cleaning not taking place. One example is in the PLOS/DataSeer dataset, where filtering by institution returned 300 individual entries where no institution was named. When cross-referenced with the 'address' and 'funding' fields in the same row most entries could be matched to institutions. Whilst this may be a result of poor metadata, either in systems used by the publisher or institution and/or data migration between the two, or how researchers provide their addresses for publications, it illustrates how a basic filtering check by the providers would have supplied the pilot with better quality data. All the pilot institutions provided their own institutional data, however it seems this may not have been used to cross match and assess the validity and robustness of the datasets produced by the providers. As the data cleaning is poor, and there is no clear documentation to guide us on how the systems pull and match the data, it is not possible to suggest improvements based on the 'by eye' method used by this pilot to check the data. It is recognised however, that the pilots may not have been provided with the full background methodology for the tools for commercial reasons, and that as we do not work in publishing, assumptions may have been made regarding our understanding of providers tools and systems, and how they work.

Commercial constraints, and lack of transparency of methods

Commercial constraints resulted in discrepancies between how solution providers obtained and thus returned their data and their accompanying reports. Two solution providers analysed DOIs from a wide variety of publishing houses and returned all data, including where DAS and cited datasets were not found and null results. One solution provider only analysed DOIs from one source and curated the returned dataset to only include results of DOIs where a DAS and named repository were identified. All solution providers' datasets were accompanied by a report on their methodology and findings. Although one solution provider aims to publish their report, the other reports discuss commercially sensitive methods and are only available internally.

Tools and methodologies must be fully transparent – this includes full documentation on what words and phrases has the algorithm been 'taught' to categorise and identify data and the data source used for each column. From our manual checks, we cannot decipher how an

algorithm identifies the differences between data citations (researcher and third party) in the paper, repository data, and citations in methodologies. In several cases, unrelated data were detected in the paper and cited as the paper's dataset. Code was not correctly identified consistently; code and data fields were populated with the same information, and in some instances, the word code had been picked up in the text from within words, for example, en*code*d.

Non-standardised DAS language, and researcher practice

A preliminary small-scale analysis of University of Bristol's DAS indicated the importance of adopting a controlled language in DAS to aid the tools in identifying data openness. We analysed 42 articles with data-bris datasets and identified that 83% of researchers used our standardised DAS text or a variation. Further analysis identified that the average machine readability accuracy for our standardised DAS was 93%, a variation of the standardised DAS was 86%, and for statements that did not use our standardised DAS, 64%. Although there are considerable issues with how restricted access datasets are categorised, this demonstrates how important the language we use in these statements is for assessment and monitoring purposes.

Whilst we recognise that the reliance of evaluating the 'openness of data' on the DAS is flawed as it is both a single point of access and failure, we need to find datasets to evaluate them. It is difficult to use quantitative tools to measure something that acts as qualitative data without controlled language. Recommendations were put forward by the pilot leads and institutional partners, such as creating a DAS taxonomy as was produced for CReDiT, universal terminology for defining the 'openness of data', and including DAS in the metadata of publications. These recommendations require collective action; funders, institutions, and publishers need to work together to 'speak the same controlled language.'

Although institutions can aid funder compliance by providing researchers with set text for DAS at the point of publication on institutional repositories, this is a minor change in the wider context of monitoring open research practices.

Ways of working: challenges and lessons learnt

The necessity to run all the pilots concurrently created challenges for the leads of this pilot. While the leads represented the University of Bristol as a participating institution in the pilots on DAS and FAIR data, the concurrent structure meant we were unable to use broader pilot data analysis to generate refined, accurate data subsets for the this pilot. This limited our ability to develop an indicator for the Openness of data further.

For example, had the pilot on the Prevalence of Data Availability Statements been run earlier in the sequence, institutions and solution providers would have had the opportunity to collaborate on multiple iterations of the response datasets, improving accuracy of the tools used to retrieve DAS. This would have provided a strong basis for the pilot on the Quality of Data Availability Statements to evaluate the reliability and quality of those statements. These efforts would have established a more robust foundation for the pilots on the FAIRness and Openness of data. The current limitations in accurately assessing DAS significantly impacted both the outcomes of these pilots and the feasibility of developing indicators that can reliably and ethically evaluate FAIRness and openness of research data.

The overall timeline was also difficult from a project management perspective. As all eight pilots ran concurrently, datasets and specifications were delivered to the solution providers at the same time (May – June 2024). This period coincided with the summer holiday season, during which many project participants and providers were on annual leave. Consequently, solution providers returned datasets later than anticipated. While analysis of the returned data had been scheduled for October–November 2024, datasets were not delivered until between October and December 2024. Due to these delays, the pilot 2 team were unable to include all solution provider datasets in its analysis. Sufficient time was needed to review and assess the

returned data, and this compressed timeline prevented comprehensive evaluation within the planned schedule.

This pilot found that scheduled fortnightly meetings were not necessary for long periods of the project. A lot of time was spent waiting for dataset returns, and relevant updates were covered during broader pilot meetings. This pilot decided not to hold dedicated meetings with solution providers to discuss data returns, as the team decided this would be a duplication of effort, given the overlap between several of the pilots criteria. In addition, sharing real-time updates and interim findings across pilot teams was not feasible. The related work was completed after the point in which those findings could have been meaningful applied to this pilot. This further reduced opportunities for cross-pilot alignment and perspectives of collective insight.

Recommendations for future work

In retrospect, the pilot programme would have benefited from adopting a staggered approach. This would have allowed institutions and solution providers more time to invest on each pilot in sequence, improving data quality, refining methodologies, and enabling a collective knowledge-building process across pilots.

Key recommendations for future iterations include:

- Appointment of a dedicated project manager and research team to oversee delivery and coordination.
- Clear definition of article types and date ranges for inclusion, ensuring consistency and comparability across pilots.
- A phased and realistic project timeline, with built-in flexibility to accommodate delays.
- Development of a communication plan, outlining frequency, channels, and stakeholders for project updates.
- Implementation of a data management plan from the outset, co-developed and adopted by all participating institutions and solution providers.

These measures would help enhance coordination, reduce duplication, improve data accuracy, and ultimately supported the development of robust and reliable indicators for evaluating open research practices.

Conclusions

As RDM professionals we are invested in ensuring that any tool used to monitor indicators of open research is transparent, ethical, and adds to the collective endeavour of increasing the uptake of open research practices. Whilst the indicators may be useful, we need to ensure that the underlying data is as accurate as it can be, and all accompanying dataset documentation provides the relevant information to make a dataset of this nature open and reproducible. What is required now is collective action across the sector to build a solid foundation for the tools to work from; this must be undertaken before we can ethically and reliably adopt machine learning tools to monitor open research with any degree of faith in the results provided. Fundamentally, within the current infrastructure and publishing environment, there is no machine as sophisticated as a human for interpreting the inherently qualitative nature of DAS, and, for monitoring the data related to open research.

Funders, institutions, and publishers need to 'speak the same controlled language' – we require consistent publisher templates and terminology within and across disciplines, openly accessible machine actionable DAS, and controlled language in data curation and citation. Until this action is taken, the indicators will only show improvement in detection rates, rather

than an increase on the 'Openness of Data', and we risk welcoming our own reproducibility crisis with open arms.

Data matters. Its accuracy and integrity matters. Underlying data, and the methods used by machine learning tools and algorithms need to be transparent and reproducible. We can learn from these pilots and previous endeavours, and make sure that until the fundamentals are fixed, we approach use of these tools with caution.

Annexes

- Annex 1: Specification for Pilot 2: The Openness of Data

References

Adams, J., Barbosa, S., Campbell, A., Campbell, R., Chandramouliswaran, I., Chodacki, J., Clement, E., Curtin, L., Hahnel, M., Day, L., Holmes, K., Jones, B., Koers, H., Linacre, S., MacCallum, C.J., McIlwaine, P., Osório, J., Puebla, I., Ross-Hellauer, T., Sansone, S.-A., Stall, S., Grant, R., Van Gulick, A. and Wood, J., (2024). *From theory to practice: Case studies and commentary from libraries, publishers, funders and industry*. Digital Science, Figshare and Springer Nature. Available at: https://doi.org/10.6084/m9.figshare.25232899.

Bahl, R., Gunawardena, K., Valdez, M. and Hahnel, M., (2024). *The Global Lens: Highlighting national nuances in researchers' attitudes to Open data*. Digital Science, Figshare and Springer Nature. Available at:

https://doi.org/10.6084/m9.figshare.25569453.

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

Data policy standardisation and implementation IG co-chairs, (2023), *Report from DAS quality Workshop*. Available at:

https://docs.google.com/document/d/1EYYzS71h58ZZn1bA-4fFDM0l9cXXEwn5Fi5zCHsJjY8/edit?tab=t.0.

European Commission, Directorate-General for Research & Innovation (2016). *Horizon 2020 guidelines on FAIR data management (Version 3.0)*. Publications Office of the European Union. Available at: https://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-data-mgt_en.pdf.

Gabelicia, M., Bojčić, R., Puljak, L. (2022). Many researchers were not compliant with their published data sharing statement: a mixed-methods study. *Journal of Clinical Epidemiology*, 150, pp. 33-44.

Hahnel, M., Smith, G. and Campbell, A., (2024). *The State of Open Data 2024: Special Report – Bridging policy and practice in data sharing*. Digital Science, Springer Nature, and Figshare. Available at: https://doi.org/10.6084/m9.figshare.27337476.

Hahnel, M., Smith, G., Schoenenberger, H., Scaplehorn, N. and Day, L., (2023). *The State of Open Data 2023*. Digital Science, Figshare and Springer Nature. Available at: https://doi.org/10.6084/m9.figshare.24428194.

Hamilton, D.G., Page, M.J., Finch, S., Everitt, S., Fidler, F., (2022). How often do cancer researchers make their data and code available and what factors are associated with sharing? *BMC Medicine*, 20, (438), pp. 1-12.

Hrynaszkiewicz, I., Harney, J. and Cadwallader, L., (2021). A Survey of Researchers' Needs and Priorities for Data Sharing. *Data Science Journal*, 20, (31), pp. 1-16.

Montague-Hellen, B., and Montague-Hellen, K, (2023) Publishers, funders and institutions: who is supporting UKRI-funded researchers to share data?, *Insights*, 36, (4), pp. 1–17.

National Academies of Sciences, Engineering, and Medicine, (2017). *Fostering Integrity in Research*. Washington, DC: The National Academies Press. Available at: https://doi.org/10.17226/21896.

Open Science Monitoring Initiative, (2024*), Principles of Open Science Monitoring*. Available at: https://docs.google.com/document/d/1eepqGt62NTdgy22jGp-nEsVC8zIvFZvexyIAVj3RdNA/edit?tab=t.0#heading=h.yzcbfhb49wsk

Tedersoo, L., Küngas, R., Oras, E., Köster, K., Eenmaa, H., Leijen, Ä., Pedaste, M., Raju, M., Astapova, A., Lukner, H., Kogermann, K., Sepp, T., et. al.., (2021). Data sharing practices and data availability upon request differ across scientific disciplines. *Scientific Data*, 8, (192), pp. 1-11.

Thoegersen, J.L. and Borlund, P. (2022), Researcher attitudes toward data sharing in public data repositories: a meta-evaluation of studies on researcher data sharing, *Journal of Documentation*, 78, (7), pp.1-17.

TIER2 Project. (n.d.) *Editorial Reference Handbook*. Available at: https://publishers.fairassist.org/.

TIER2 Project. (2024). *Flowchart - Editorial Reference Handbook*. Available at: https://osf.io/245zj.

TIER2 Project. (n.d.) *Pilot 7 — Editorial workflows to increase data sharing*. Available at: https://tier2-project.eu/pilots/7.

UNESCO, (2021). *Recommendation on Open Science*. Available at: https://doi.org/10.54677/MNMH8546.

Warren, C., Godsall J. (2025). *Specification for Pilot 2 The Openness of Data v1.0*. Available at: https://osf.io/7bwvk

Watson, C., (2022). Many researchers say they'll share data — but don't. *Nature*, 606, 853.

Appendix 2: Openness of Data. Annex 1: Specification for Pilot 2: The Openness of Data

V1.0 – last updated 16 May 2024
Authored by Christopher Warren, Jade Godsall (University of Bristol, pilot leads)

1. Headline

Open data describes those research data outputs published where content is made freely available online to use and redistribute, with no access control necessary. There is frequently a **Creative Commons licence** applied to the data indicating what level of commercial or non-commercial re-use is permitted.

An indicator of the Openness of data (i.e. the degree to which data is fully Open) must be able to distinguish whether data cited in a **Data Availability Statements (DAS)** is "Open" or "not Open". For this pilot, any identified data that is openly available without access control will be considered "Open data".

Pilot 2 ('the pilot') will build on the work of Pilot 4 looking at the prevalence and quality of Data Availability Statements (DAS), as a necessary precursor to producing indicators of the Openness of the data cited in those DAS.

2. Definition criteria

As a baseline, the following criteria will be measured, to produce an analysis of pilot results. Of necessity, some of these criteria will be overlapping the work done on studying the Prevalence and Quality of DAS (UKRN Pilot 4).

The pilot will:

2.1 Describe where and how research data, code, software or other materials supporting the results reported in a research output can be accessed. This should also describe where multiple instances of those data are cited in a single Data Availability Statement.

2.2 Include hyperlinks and persistent identifiers (e.g. DOI or accession number where available) if research data, code, software or materials are recorded or stored in a catalogue or online repository.

2.3 State that research data, code, software or materials are available in supplementary files if appropriate.

2.4 Identify the degree of Openness of cited data: **Open data** / **Technically Open data** / **Embargoed data** / **Controlled data** / **Author Controlled data** / **Closed data** / **Dark data** (definitions of each term in Glossary below)

2.5 If Open, describe the reuse license applied by each repository to each dataset cited: this might be Creative Commons licenses, public domain, or otherwise.

2.6 If not Open, explain why data cannot be shared openly, for example, to protect study participant privacy or respecting commercial contracts, if appropriate.

2.7 If data is only available after an embargo period, provide a specified date when the data will be made publicly available.

2.8 If data is not Open but is available on request, provide a clear point of contact (e.g. data access committee with the likelihood of timely response) for the process of requesting access.

2.9 Record a verification check of the accessibility of the data through its persistent identifier, i.e. prove that the link to the data works and the data is openly accessible. This may require machine reading or downloading the dataset(s) contents.

2.10 Identify whether the dataset contents are made up of open file formats, proprietary file formats or mixed.

2.11 Evaluate whether the Data Availability Statement identifies adequate data to back up the article's conclusion: whether the supplied data is whole and complete (checking dataset contents against a dataset inventory, if available), and also if possible verifying that the dataset is relevant to the subject of the article.

3. Methodology

Open data cited in DAS will be counted and analysed using the algorithms of industry partners ('partners' or 'providers') run through datasets provided by one or more of the participating institutions. Each provider will use their own software and unique method for doing this.

The pilot will:

3.1 Start with a small sample database provided by the lead or co-lead institutions based on research articles published in 2023 extracted from their institutional Current Research Information System ('CRIS'). CRIS data is in the public domain, and will not involve sharing commercially or personally sensitive information at any stage.

3.2 Identify as a % those items at an institutional level that contain DAS.

3.3 Categorise findings against the definitions outlined in Section 2 above, where possible. Categorise findings according to:

- 3.3.1 Location (by repository) of any data, software, code or other supporting materials cited in DAS
- 3.3.2 Location (by persistent identifier) of any data, software, code or other supporting materials
- 3.3.3 Any supporting materials present in the article not included in DAS
- 3.3.4 Openness type: Open data / Technically Open data / Embargoed data / Controlled data / Author Controlled data / Closed data / Dark data
- 3.3.5 License applying to each Open data, software, code or other supporting materials cited in DAS (if available)
- 3.3.6 Reason for non-availability of any non-Open data, software, code or other supporting materials cited in DAS (if available)
- 3.3.7 Specified date for accessing embargoed data, software, code or other supporting materials cited in DAS (if available)
- 3.3.8 Point of contact for accessing any non-Open data, software, code or other supporting materials cited in DAS (if available)

3.3.9 True/False indicator for live status of data record available through persistent identifier in 3.3.2
3.3.10 File formats in each data record checked
3.3.11 Check of dataset contents against data inventory file (if present)
3.3.12 Check of dataset appropriateness for the article subject

4. *Scope*

4.1 Institutional dataset(s) to be submitted will include articles only.

4.2 Institutional dataset(s) to be submitted will include articles from a limited period, to be agreed upon between each institution and each provider.

4.3 Institutional dataset(s) to be submitted will include the following fields as mandatory (see draft data template for more information: University of Bristol repository data for providers):

4.4.1 LocalID (local CRIS repository ID number)
4.4.2 DOI (publisher DOI)
4.4.3 Link (publication link)
4.4.4 RepoPageLink (link to repository splash page/similar)
4.4.5 Type (item type, i.e. journal article)
4.4.6 Year (year of publication i.e. 2023)
4.4.7 OrgUnit (the managing organisation unit record in CRIS)

The dataset linked above, drawing on a common dataset template originally provided by Pilot 4, will provide a baseline data extract that will facilitate consistent data analysis across participating institutions in Pilot 2, and enable cross-comparison with other pilots following the same basic template, i.e. Pilots 1 and 4. See 20240328 UKRN OR Coord Approach 1 2 4 5)

4.5 Some institutions may want to amend this basic template for their own purposes, according to differing research aims or depending on the quality of local data/capacity of local CRIS to produce data. It will also be possible to refine institutional data to focus on specific funders, research output types or date ranges. This document is looking at the pilot exercise only, and institutions should not be restricted in future exercises to these simple parameters.

4.6 These parameters, and the criteria listed in sections 2 and 3 above, are also subject to the abilities of providers to produce such findings, or equivalent findings, through their tools and algorithms. They may be able to 'value add' data using their specific tools.

4.7 The institutions submitting data may want to perform data quality checks on the findings produced by the providers, to verify the outcomes and to inform any further conversations with the providers aimed at improving their tools.

5. Issues

5.1 Not all full-text documents will be licensed for use by pilot partners. Alternative means of access to author accepted manuscript versions could be explored. Depending on CRIS functionality, an optional field linking to the CRIS manuscript PDF could be added to accommodate this. Any gaps will be accounted for in the analysis.

5.2 The ability of pilots to share their findings based on provider tools/algorithms will need to be addressed with each provider individually if there is the risk of breaching

providers' intellectual property rights by sharing datasets produced using their proprietary algorithms

5.3 This pilot is exploratory, and is open to the potential for any possible outcome, from the ability to drill down into the openness of particular file types within a non-Open dataset, through all possibilities, to the potential for a finding that we cannot create a reliable, consistent indicator for the Openness of datasets based on machine-readable data.

Glossary

Creative Commons licence: a license which allows for re-use and re-distribution of the underlying file(s). There are a range of these licenses according to the level of restriction applied (https://creativecommons.org/share-your-work/cclicenses/)

Data Availability Statements: Data Availability Statements, also known as Data Access Statements (DAS) describe where materials supporting the results reported in a research output can be accessed. Their primary function is to facilitate the replication or reproducibility of research by providing a simple means to access materials, improving trust and transparency in research.

Degrees of **Openness** for purposes of reporting:

> **Open data** (reporting): freely available for re-use and redistribution by anyone without any registration
> **Technically open data**: freely available for re-use and redistribution by anyone, but requiring some form of registration
> **Embargoed data**: freely available for re-use and distribution by anyone, but after an embargo period
> **Controlled data:** available 'on request' through standard data sharing agreement process
> **Author Controlled data:** available 'on request' through contacting the author directly
> **Closed data:** data exists in a repository but not available to be shared
> **Dark data:** exists outside a repository, cannot be shared or linked to

Open Access: Open access is free, unrestricted online access to research outputs. Research is published in journals and other scholarly forms to ensure results are recorded and communicated to the wider community. Publishers may require authors to transfer copyright to them, preventing researchers from distributing their work. Open access may be Green (an author accepted manuscript freely shared via a repository), Gold (paid for license to make the publisher's version freely available), or Diamond (similar to Gold, but with no payment involved for the license).

Open data (general definition): Open data can be freely used, re-used and redistributed by anyone - subject only, at most, to the requirement to attribute and sharealike. It may be in the public domain already or may carry a Creative Commons license.

Open file formats: Formats which are not dependent on proprietorial or bespoke software, but which can be opened by open source software, e.g. the spreadsheet format .CSV rather than Microsoft's .XLSX format, or the document format .RTF rather than .DOCX.

Public domain: Creative materials that are not protected by intellectual property laws such as copyright, trademark, or patent laws.

Research data: Any published data produced in support of research aims. Research data may have been published in support of a specific research output (article), or may have been published as a standalone work independent of any direct research output

Appendix 2: Openness of Data. Annex 2: CWTS Review

Report on UKRN pilot study on Open Research Indicators: Report of a Pilot on the Openness of data

This pilot project focused on Open Research Indicators on the Openness of Data. The work has been conducted by a team of experts form various UK universities. The study was based upon DOIs of articles published in 2023 in three databases, from three different providers. The work based on visual inspection of data Availability Statements (DAS), and assessed these against the articles behind the DOIs, in order to determine whether the link between DAS and articles functions well. Various tools used by the providers were tested, and the conclusions was that tools can, hypothetically do whatever is promised, while in practice this seemed much more difficult, and tools were not able to check the relation between openness of data and the DAS. The report lists various recommendations, aimed at funders, publishers, institutions and researchers, to accelerate the uptake of open research practices.

I am well aware of the fact that the reports have been ' closed down', so no more changes can be made, however, I hope my comments will be useful in case someone, either team members or others, decide to do follow-up work on the outcomes of the study.

Let me start by stating that this report follows a clear structure, with sections in the report. However, a table with an oversight of those sections would have been helpful to the readers. One aspect that confused me while reading it the first time, was that the authors have not paid attention to the fact that due to section breaks in the Word document, page numbering started two times from 1 again. This makes the report somewhat confusing to read. *[Note: page numbers now corrected]*

I appreciate the start with *Background and aims* section, which also includes a literature overview. This is helpful, as not everybody interested is equally well versed on this topic, and hence a proper introduction is very useful here.

The Methods section clearly outlines the use of a number of approaches or perspectives, such as the SCOPE method and the UNESCO point of view, arriving at clearly outlined guiding principles on page 6. What strikes me here is the fact that early in the research in this pilot, the central role played by DAS is commented on, as DAS are the stepping stone to the conditions of the data in any research report. Another remarkable issue here, and that returns on page 7, where legal constraints are discussed, is the lack of any reference to licensing of data. One can imagine that additional or supplementary data to for example journal papers, whereby the data are stored on the publisher's' website, do have other licensing constraints compared to data stored on a more neutral territory, for example in a data repository. Still in that same section, in the sub-section "Identifying the data – data pilots' collaboration", I would have liked to see also at least the mentioning of the issues related to the (lack of) uniformity of DAS, as that could have been collected form the colleagues running the pilot on DAS. This would have been helpful, for the readers of this particular pilot report.

If we then enter the first section with started numbering at 1, after page 10, the team does a great job by providing in a very detailed manner overviews of the providers of information. Due to the way the materials is presented, it is relatively easy to make comparisons across providers. While stating that, I would have appreciated a statement by the team on the fact that Elsevier has stopped its' **Research Data Monitor**, as a commercial tool. I know that, due to experiences within Leiden University with that same tool. Such a remark would be helpful, in case any of the readers of this report might be interested in using that Elsevier tool for their own internal purposes.

Then, following the section based on the landscape orientation, we get into the final section, which again starts numbering at 1. This Discussion section starts with a short summary of main findings of the research in the pilot, whereby I would have liked to see in the summing

up of impossibilities also the fact that automation is still far away, when it is about openness of data, and how to detect openness of data.

Furthermore, the next sub-section nicely identifies the street light effect the providers are confronted with, namely to be only able to use their exiting tooling to supply the pilot study with material and information. The third bullet point, Evaluating the DAS not the data contains some very important statements and conclusions. First of all, the fact that providers worked from DAS, not from the cited data. Apparently, there is something complex here, which remains non-discussed. Next, the fact that data availability is in some instances depending on dubious accessibility, namely via statements as " on request', or "approach the author", this vagueness is not compliant with (open) data policies, and hence needs to change.

On page 2 of the third part of the report, under the sub-section *Data cleaning and integrity*, the first line reads "there were issues with data integrity and data cleaning" I would have liked to see some deepening of what was meant with in particular data integrity in this sentence and paragraph.

On the Recommendations sub-section, I agree with the observation that the whole pilot project could have been approached differently, which might have had a positive effect on some of the pilots within the larger framework of the project. However, that is water under the bridge, and an important learning moment.

When it is about the key recommendations, most of that relates to process, it is my impressions, as a relative outsider, that some things were in place, such as a central project manager, but perhaps that did not work out well. What I get as an impression, again as outsider, is that the money felt short, time to conduct the research was too limited, and the fact that due to those conditions, work had to be squeezed into existing work schedules and planning, did not help the pilot very much. This work definitely deserves continuation, as the further study on data handling and sharing is supposedly an important element of the further development of open science research practices.

Review by Thed van Leeuwen, Berlin, November 12th 2025

Appendix 2: Openness of Data. Annex 3: Team response to CWTS Review

Team response to: Report on UKRN pilot study on Open Research Indicators: Report of a Pilot on the Openness of data

Overall comment

It is possible that as RDM professionals we look at the data provided by solutions providers from a different perspective than the reviewer, as we are looking at the accuracy of the metadata provided and any associated free text as research records. This is of paramount importance as the accuracy of this data is the foundation from which recognition and reward for Open Research practices can be built, and the validity and integrity of research can be tested. Our primary concern is that the data must be unimpeachable at a fundamental level if there is any potential for later iterations for any of the tools to be used for comparisons or league tables, whether this is the current suggested application, or not. Thus, we concentrated on the accuracy of the data provided by the solutions providers, as this needs to be sufficiently accurate before any improvements in indicators can be detected. If this is not right in the first instance, it will be difficult to unpick and correct further down the line as the algorithms evolve.

Formatting - structure and page numbers

UKRN reformatted the report after submission which led to the page number discrepancy. We have provided an updated copy that resolves this issue.

Data Integrity

We chose not to provide further details on the data-integrity issues to avoid unfair comparisons between providers, whose returned datasets varied substantially in both size and the extent of data cleaning performed. We have a large amount of analysis which identifies clusters where there are failures; this would be a larger project requiring more resources than were available for an unfunded pilot, though we may work on this on our own timeline, or obtain funding to undertake a more detailed analysis. Additionally, we felt our report needed to be in keeping with the other pilots, and inclusion of our rudimentary analysis produced a report which was too lengthy.

More generally, as noted in the report, while the available tools can successfully extract data where access is possible, the quality of the returned data was too poor to produce a reliable, transparent, and ethical indicator for assessing the Openness of data. This issue does not solely reflect a limitation of the tools; rather, further work is required at the publisher, funder, institutional, and researcher levels to standardise the language, location and readability of the data (in this instance, Data Availability Statements) being extracted.

Licensing of data – additional or supplementary data

An in-depth analysis of the types of licensing attaching to data cited in our report was not possible as the providers did not supply us with this data. We were reliant on the providers' information, and their algorithms needed to focus on DAS text, not underlying repository metadata. Our pilot was also focusing on the overall Openness of data in terms of accessibility, rather than the FAIRness of data (Pilot 1) Our question was whether a dataset was Open or not, using the criteria specified and agreed by the institutional partners in Pilot 2 (see table below). This was not adhered to by the solutions providers. The providers encountered difficulties in distinguishing between cited (external) datasets, and supplementary (included as article attachments) datasets, thus making any additional reading of the licenses pragmatically impossible.

Open data	freely available for re-use and redistribution by anyone without any registration
Technically Open data	freely available for re-use and redistribution by anyone, but requiring some form of registration
Embargoed data	freely available for re-use and distribution by anyone, but after an embargo period
Controlled data	available 'on request' through standard data sharing agreement process
Author controlled data	available 'on request' through contacting the author directly
Closed data	data exists in a repository but not available to be shared
Dark data	exists outside a repository, cannot be shared or linked to

Identifying the data – lack of DAS uniformity

The point here is valid; however, it underscores an additional point which, with the benefit of hindsight, has become clear – several of the pilots would have made more sense to have been run consecutively, rather than concurrently, and the lessons learned from earlier ones could have been passed onto the next. It would have made sense for the DAS Pilot (Pilot 4) to have started first, to enable definitions and uses of the DAS to be used by its natural 'child' pilots 1 and 2 on FAIRness and Openness of data, both relying on DAS statements. As it was, Pilots 1 2 and 4, although sharing a great deal of collaboration, were learning different lessons at different time. Chief among these, we found when comparing notes close to the end of the data-gathering section of the process, was the wide variation in the uniformity, reliability and readability of DASs. It would have been very helpful if we had known at the outset.

Discussion – automation and working from DAS

Providers working from the DAS wording, rather than datasets' metadata found in data repositories, has been an issue since the start of the project. We were not a party to the initial discussions pre-pilots, so were unaware of what was agreed in the pre-project phase between project leaders and the providers. When the project commenced, it was clear that, if asked to report on the 'Openness of data' in the context of how Open datasets cited in journal articles were being treated, providers would train their existing tools' algorithms to read DAS wording, not the wording of underlying repository datasets. This has meant, in effect, all pilot results have relied on machine reading of human-authored wording (i.e. journal authors writing their DAS text) – which indeed is something complex, but one that any pilot reading DAS texts (or even comparing DAS text with underlying repository metadata) would have to encounter.

Lack of project manager

Each pilot had 'leads' but there were no project managers. The project lead role would coordinate meetings and communications, but there was no %FTE allocated to project management across the pilots. Bristol was initially involved in several other pilots, but we had to withdraw due to lack of resources (see comment below). Our Pilot (and others) identified several aspects such as methods of sharing data, arranging meetings, vocabulary, etc. which would have benefited from a central project manager's 'umbrella'. As it was, this proved tricky to coordinate without one person at the helm guiding projects and standardising ways of working – thus, any ability to directly compare results at the end of the study was impacted. We excluded formal discussion on this in our pilot result for time and space limitations, but it has been our collective impression throughout the process.

Sole lead

Bristol was unique in being a sole pilot lead. Additionally, all staff on a small team (The Library's Research Data Service) committed to the project and at times were working on the pilots 80% of the week, which impacted our own service. Our participation in the project was

voluntary with no fiscal or resource compensation for us to work across 5 pilots. Consequently, we could not resource the amount of time warranted for a project of this scale.