Appendix 3

Open Research Indicators: Report of a pilot on the Prevalence and Quality of Data Availability Statements

Authors

Valerie McCutcheon 1†, Mick Eadie 1, Laurian Williamson 2, Radoslaw Pajor 2

1 University of Glasgow, 2 University of Leicester

† Correspondence should be addressed to Valerie McCutcheon; E-mail: valerie.mccutcheon@glasgow.ac.uk.

Executive summary

In this pilot, research organisation members were interested in testing the feasibility, effectiveness, and potential future methodology for 'tracking' both the presence and quality of Data Availability Statements (DAS). Monitoring the presence and availability of DAS is one way of measuring the growth and adoption of open research practices. Institutionally, monitoring DAS usage may have resource requirements and demand specialist skills and insights.

Key considerations for institutions are to be as open and transparent as possible without unwieldy cost or resource implications.

Data and methods used in this pilot were in some cases proprietary and restricted.A balance between these restrictions and a desire for openness and transparency was an important consideration in what to share. We believe the key findings cover the most useful information.

This small-scale time limited pilot study provided some useful findings and lessons learned for both partners and providers. However, participating in a pilot with limited resources (time, people, and tools) meant that there were activities and tasks that could not be done that may be useful for future research. For example, direct comparison of institutional data was not done. The pilot did enable all stakeholders to explore the feasibility of 'tracking' both the existence and quality of Data Availability Statements (DAS).

Whilst this pilot provided useful awareness raising and discussion on this topic, and fostered meaningful cross and inter-stakeholder relationships and better understanding of the challenges and possible solutions, we recognise that standardisation of DAS monitoring is still immature and further cross stakeholder discussion is desirable.

Timeline

Date	Action
February 2024	Glasgow and Leicester meet, set up recurring partner / leads meetings and arrange individual calls with suppliers.
February 2024	First draft of pilot specification shared for comment.
April 2024	Generic dataset template agreed and uploaded to Figshare.
May 2024	Pilot specification agreed with project partners and suppliers.
May 2024	Partner institutions begin adding datasets to figshare (May 21st deadline).
July 2024	Pilot specification added to OSF (osf.io/njcv7) and see Annex 1
July 2024	First Supplier dataset (Digital Science).
September 2024	Second supplier dataset (Elsevier aggregated data).
October 2024	Third supplier dataset (PLOS / Dataseer).
November 2024	Fourth supplier dataset (Elsevier itemised dataset).
December 2024	Report drafted and partner comments received.

Background (rationale and history)

Data Availability 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. They provide a useful indicator of good open research practice.

Underlying research data, code, analysis software and other materials can sometimes be separated out into different statements, and referenced with different names, e.g.: 'Code Availability Statements', 'Analysis Software', 'Availability of Data and Materials', 'Data Access Statements', and so on. For the purposes of this pilot, looking at the prevalence and quality of DAS, any statement referencing any type of underlying material was considered in scope.

It can be challenging to discover and track data availability statements. Sometimes this is done manually in a light touch way. Where automated methods are used, for example: machine learning, text mining and pattern matching across article full-text, there are challenges such as: difficulty in locating the appropriate document; the placement of the DAS within the text; the variety of ways that stakeholders use to tell readers where to find the data; or the lack of technical resources and skills to use available tools. Furthermore, in some instances DAS are considered a part of the full-text of research articles, and therefore protected by stricter Intellectual Property Rights (IPR) than standard article metadata. This can make them difficult to analyse and reuse. As a result, DAS analyses across articles from multiple publishing houses can discover the legal/IPR challenge is a much bigger bottleneck than the technical one.

Simply detecting the presence of a DAS is often not particularly useful as the quality can vary. For example, a DAS may simply direct a reader to ‘contact the author’; or provide a direction to ‘supplementary files’ or similar; or perhaps an author is simply adding some text to fulfil a funder or publisher requirement without taking the care to make sure the DAS contains an easy to follow instruction on how to access underlying data.

Although many of the requirements have been discussed and some progress made there is still no standard agnostic framework or set of definitions. Some tools, scripts, and databases have been used but more can be done to ensure reliable results and ease of use.

This viewpoint is backed up by the ODI Open Discovery Initiative that advocates improved references and connections between articles and research data.

We acknowledge all the many initiatives to explore this work and we hope that our work will help progress towards a more standard approach to counting and categorising DAS.

Overview

The overall aim of the pilot was to explore the co-creation of practical methods to:

- Monitor the prevalence of DAS in research articles
- Assess the quality of DAS and their usefulness

The openness and FAIRness of datasets were the subject of separate pilot projects which re-used the same input datasets. All 3 pilots worked together to contribute to the wider goals of the UKRN Open Research Project.

The purpose is to provide information of value to institutions to better support good practice in open research.

This final report illustrates the results of the work including:

- Methods
- DAS definition
- Establishing ways of working
- Evaluation and quality assurance
- Caveats
- Key findings and recommendations
- Lessons learned

As the goal is to consider indicators and methods to produce those, and resources were limited, we do not delve far into the meaningfulness of the results for specific organisations.

Methods

The following organisations participated in this pilot, volunteering time and expertise.

Pilot	Participating institutions (leads in bold)	Participating providers
Data availability statements	**Glasgow, Leicester**, Surrey, Sheffield, Reading, Newcastle, Manchester, Leeds, Edinburgh, Bristol, Exeter	Digital Science, Elsevier, OpenAIRE, PLOS/DataSeer

Research Organisations used the SCOPE Framework for Research Evaluation (2019) to consider the purpose of the work.

The main purpose was:

To encourage good open research practice in sharing datasets openly, satisfying individual and organisational desires and policies around open research, and compliance requirements from funders for example the UKRI open access policy.

We note that as suggested by the UNESCO Draft Principles on Open Science Monitoring (p4), indicators must be 'relevant for the specific monitoring task at hand' and careful consideration should be applied before reusing the indicators for other purposes.

To advance the indicators there was a desire to:

Find more automated methods to baseline levels of data availability statements which can be used to monitor uptake of good practice.

Encourage stakeholders to agree standards and methods to facilitate the inclusion and counting of good quality data statements in research outputs.

A specification was written for the work.

DAS Definition

The specification noted that DAS were expected to:

1. Describe where and how materials supporting the results reported in a research output can be accessed.
2. Include hyperlinks and persistent identifiers (e.g. Digital Object Identifier (DOI) or accession number where available) if materials are recorded or stored in a catalogue or repository.
3. State that materials are available in supplementary files if appropriate.
4. Explain why materials cannot be shared openly, for example to protect study participant privacy.
5. If a barrier to immediate access exists, provide clear reasoning (e.g. sensitivity or volume) and with a clear point of contact (e.g. access committee with the likelihood of a timely response).
6. Clearly separate raw data and any secondary data reuse that supports results and analyses.
7. If appropriate, clearly state that no materials were produced in the research.
8. Be positioned in a DAS Section or other separate and clearly marked part of the paper.

9. Use a DAS template.

These 9 criteria gave the pilot a point of reference from which further categorisation and assessment of quality could be gauged.

The categorisation mapping developed by the pilot was a significant outcome.[1] Some suppliers were able to provide categorisation that mapped to elements of the pilot specification, while others used to pilot as a chance to develop and test categorisation against the results of their analysis. The pilot categorisation mapping was an attempt to harmonise this activity and to produce a consensus of DAS categories that can usefully inform research.

The pilot started with a defined sample dataset of articles published in 2023 with full-text available in the public domain – please refer to supplementary data:

DAS pilot data from Glasgow and Leics.zip

The suppliers used DOIs as the input to access the articles which in majority were under CC-BY licence. This would indicate the majority of the DAS analysis reported back by suppliers was based on published versions (Version of Record - VoR).

The VoR analysis was at odds to some extent with institutional manual checking for DAS. Institutions tend to manually look for a DAS in the earliest available document they have when creating the repository record, which may be the Author Accepted Manuscript (AAM). The pilot results found that the AAM sometimes does not contain a DAS when the publisher VoR does.

The calendar year 2023 was chosen as it would provide a full year's worth of data and was recent enough to provide a complete as possible picture of current practice. Providing items that were free to read in the public domain it was hoped would give automated tools the best chance of being able to locate and analyse a document. This was considered a useful way to facilitate scrutiny and reproducibility of any finding to facilitate future discussion. The datasets provided by participating research organisations were uploaded and stored in a temporary shared FigShare Repository.

At minimum the dataset contained the following attributes: Local Identifier; Digital Object Identifier (DOI); Local Uniform Resource Locator (URL); Type; Year; OrgUnit (Organisation unit)

Partner institutions could provide additional attributes if they were interested in particular aspects they wanted to analyse. For example, if manual checking for the prevalence of DAS was recorded locally that value could be provided; or if the local institution was interested in running some analysis based on research funder, then that value could be provided.

DAS were counted and analysed initially using methods provided by suppliers. Suppliers were asked to:

1. Identify the percentage of items at an institutional level that contain a DAS (Prevalence).
2. Categorise findings according to the project specification and their own methods (Quality)
3. Explore further analysis by type of dataset, academic discipline if they wished.

Suppliers took different technical approaches to these tasks which consisted largely of a mix of pattern matching and the development of bespoke language models[2].

[1] UKRN Data Accessibility Statement Pilot Classification Discussion Document

[2] PLOS-DataSeer reused a simple open source DAS classifier tool available at https://zenodo.org/records/3470062/preview/alan-turing-institute/das-public-v1.2.zip?include_deleted=0#tree_item18

Each provider was able to deliver to some degree on the specification requirements. For example PLOS-DataSeer reported being able to fully address the 6 requirements listed below using currently available methods, with an additional 5 requirements proving more difficult to achieve. The PLOS-DataSeer mapping file is embedded below for reference.

UKRN pilots &
PLOS-DataSeer OSI re

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

The requirements that were largely met included:

- Describe where and how materials supporting the results reported in a research output can be accessed.
- Include hyperlinks and persistent identifiers (e.g. Digital Object Identifier (DOI) or accession number where available) if materials are recorded or stored in a catalogue or repository.
- State that materials are available in supplementary files if appropriate.
- Be positioned in a DAS Section or other separate and clearly marked part of the paper.
- Identify the percentage of items at an institutional level that contain a DAS
- DAS type: research data (digital or physical), code, software, supplementary materials

More difficult requirements to address were:

- Explain why materials cannot be shared openly, for example to protect study participant privacy.
- If a barrier to immediate access exists, provide clear reasoning (e.g. sensitivity or volume) and with a clear point of contact (e.g. access committee with the likelihood of a timely response).
- Clearly separate raw data and any secondary data reuse that supports results and analyses.
- If DAS text accurately describes the material it links to.
- If the link provided in the DAS resolves to an accessible resource.

While this was the PLOS-DataSeer reported outcome it was largely true across the supplier methodologies, and not surprising given the complexities inherent in developing automated tools to address the requirements.

Research Organisations (ROs) reviewed and analysed the datasets returned by the suppliers. The primary goal of the analysis was to:

and that was published with https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0230416

1. Make a comparison of the prevalence of DAS was found by suppliers through automated checking tools against manual checks provided by ROs.
2. Make an assessment of the categorisation provided by suppliers against the project specification and local RO methods for categorising DAS.

The method of analysis used was to:

1. Look in detail at random samples of data. The size of the random sample varied according to the time the RO had to look at it, but was usually around 10% of the returned datasets.
2. Highlight discrepancy between local RO checks and automated Supplier checks. For example where the RO said there was a DAS and the Supplier said there was not and vice versa.
3. Categorise the random sample 'blind' and then compare it to the category provided by the Supplier.

General note on analysis:

Detailed analysis of all pilot partner input data and supplier returns; and the provision of a harmonised, joined-up comparison of datasets proved to be beyond the scope of this small time-limited pilot. Our key findings relied on the analysis of small samples of data but we hope are sound enough to provide general comment and promote further work and discussion.

Establishing ways of working

From the onset, in the pilots we tried to adopt common working practices and templates where appropriate, to make it easier for partners who were involved in more than one pilot. For example we all used a communications file to record key actions and discussion points.

For the data availability statements pilot we agreed to have a fortnightly partners meeting even if there was little business and the meeting was short. This allowed partners who had missed a previous meeting or had a question to have a point of contact and reduce the risk of a long passage of time where any significant issues were not raised. This pilot logically got started before the other data pilots and so established the working pattern and did not suffer to the same extent as some others may have in terms of communication difficulties. The communications file, meetings, and emails remained the main communications methods. The main difficulty in communications arose in the wider programme where we needed to communicate between pilots at logical meeting points. More effort was invested in this because of the structure of the programme and in hindsight perhaps a project with work packages and a timeline would have been an alternative approach.

Evaluation and quality assurance

We evaluate ourselves against the original specification which was to explore co-creation of practical and agnostic good practice methods to:

- Monitor the prevalence of data availability statements in research articles
- Categorise the data statements in terms of usefulness

With the aim of helping research organisations better support good practice in open research.

We feel we made good inroads into how to count and categorise data availability statements. The work was not intended to be, nor is, at a mature enough state of development for some of the programme evaluation criteria suggested to be relevant. There were significant variations in findings because there is no agreed standard at present. Having information that illustrated this helped us to reinforce that standardisation is desirable.

We remind the reader that this was a pilot (rather than a formal research project) done with limited voluntary resources and as such further work is recommended. We were testing the feasibility of a process that can be considered as preliminary work.

Caveats

Just because a DAS was not listed on findings does not mean that there was none or that there was no data. The variety in wording, placement, and approaches to sharing data mean it can be difficult to capture and count all statements. Therefore the findings of different partners may not be directly comparable.

Key Findings and Recommendations

Topic	Finding	Recommendation
Discovery of DAS - positioning and wording	In general, suppliers were able to find more data availability statements, probably because local checks are necessarily cursory and manual. However, there is variation in what is found depending on the approach taken and positioning and wording of statements as formal subsections in publications or in supplementary material, footnotes or elsewhere in the publication. This made it difficult to fully validate checks. Some topic areas and some publishers had higher percentages of data statements.	Continue with manual checking short term. Continue to explore tools including commercial and open source options that can meet the specific needs rather than trying to fit the needs to an inflexible tool.
Discovery of DAS - document versions	Suppliers often looked at the published version which may have included a DAS which was not present at acceptance stage when research organisations typically do high level checks.	Encourage authors to add DAS as soon as possible when drafting papers. Noting that without careful moderation this can lead to a 'two DAS' issue where there is one in the text and another added in a publisher template as part of the publication process. Need to work with publishers to solve this.
Definitions of DAS - categorisation	Comparing categorisation across the results returned by suppliers proved to be difficult in some cases. For example some categorised DAS that stated no data shared or no data available as 'NA' while others counted them a valid DAS.	The methods we propose are to have a specification of definitions of what is an acceptable DAS and what falls short that can be applied regardless of whether a supplier solution or local resource is used to count DAS. We realise that suppliers may wish to retain confidentiality on algorithms used and yet we also need to be confident that

Topic	Finding	Recommendation
	Similarly - although a consistent input dataset was provided by institutions for suppliers to use it was not always possible for each supplier to access and analyse the same documents. This made it difficult to meaningfully compare a large set of results across the suppliers' returns.	numbers can be reliably compared. We also recognise some variation might be desirable in certain cases but that should always be clarified in the context. There was in the main commonality that we believe can be finessed into some standard high level categories that most stakeholders can agree are workable for reporting on quality of DAS to inform good research practice at research organisations. A copy of our working document is available here UKRN Data Accessibility Statement Pilot Classification Discussion Document Stakeholder reachout - publishers, infrastructure providers (e.g. Datacite), professional networks (e.g. Research Data Alliance) funders (e.g UKRI working on a new research data policy framework and The UK Research Excellence Framework People, Culture and Environment) research organisations, standards organisations e.g. National Information Standards Committee,
Sustainability - Maturity	We recognise that this indicator is at an early state in terms of maturity and some work is needed to move from what we can count at present to better ways to count levels of good practice in data availability statement usage.	Include this in stakeholder conversations.
Article Type Definitions	Standardised definitions and labelling of article types would help remove noise from future dataset. E.g. 'letter'	Include this in stakeholder conversations.

Topic	Finding	Recommendation
	is sometimes used for a full research article	
Research Field	There was a substantial variation by research field.	Not a specific output of this project. Some parties may wish to investigate further.
Definition of DAS - quality	A high percentage of DAS state that data are available 'on request' or in the 'supplementary' data. Future work could be to dig into this and come up with a 'standard 'way of articulating 'on request' e.g. NOT 'contact the author'	Encourage authors/stakeholders to avoid use of 'on request' Include best practice in all advocacy and engagement with stakeholders.
Standardisation	There are issues for manual checking and in using automated tools due to DAS placement, lack of consistent language, limitations on what text can be read, different (or lack) of templates.	Encourage more consistency in these areas to make text mining/pattern matching/language modelling more achievable Include a drive to standardisation in stakeholder conversations along with incentivisation. Note also that 'standardisation' should help with automated detection, it does not in itself solve the problem of quality or improved sharing. This outcome also requires advocacy, support, resources and training (e.g. findings in https://doi.org/10.6084/m9.figshare.13810220.v1) Worth noting that standardisation by using DAS templates can sometimes lead to 'box ticking' and a one size fits all approach is not desirable. Some publishers prefer that authors describe their data in their own words which when done correctly is more meaningful / useful in context. See above re. advocacy/training as a continued solution.

Topic	Finding	Recommendation
Sustainability	The maturity of open research practice differs among the pilot partners. This pilot provided an opportunity to foster a cross-stakeholder community of practice. UKRN and other organisations can be a more powerful lobby group for change than an individual research organisation or supplier.	Explore with pilot stakeholders the feasibility of establishing a community of practice where all interested stakeholders could come together to share ideas, experiences, and best practice with the aim of driving forward improvements and advocacy around both the prevalence and quality of DAS.

Lessons Learned

Experience	Positive/Negative	Impact and Recurrence	Lessons Learned	Best practice or problem?	Actions Required Implement Lessons Learned
Separate pilots meant that there was confusion and competition for resources.	Negative	HXH	Plan the project using project management methodology at the beginning to ensure best sequence of activity	Best practice	
Additional and disproportionate project management demands added during the project, often late in the day caused stress and confusion.	Negative	HXH			Agree key deliverables and any mandatory elements of these and related processes early in the project. E.g. Some users found inconsistent labelling of the output files made data analysis more difficult. They could have been

					more consistent with the specification if more time had been available to pursue this.
Regular reminders that the participants were volunteers and their contribution was gratefully received. Acknowledging the challenges placed by the structure of the programme,	Positive	MXH	Challenging projects can focus on more positive aspects with a bit of encouragement.		Ensure regular feedback and encouragement given to teams.

Conclusions and next steps

Standardisation is highly desirable if we wish to improve the prevalence and quality of DAS to support research organisations in good research practice.

We propose this classification that was broadly agreed as part of this pilot, as an input to stakeholder conversations.

Data availability statement classificatio

How can we sustain the momentum and foster stakeholder conversation going forward? We hope to discuss this at UKRN meetings as part of the UKRN action from this pilot.

In the meantime we recommend that research organisations:

- Continue to encourage uptake of good quality DAS and associated sharing / data management practices that enable them
- Provide guidance on writing the statements e.g. the Data Access Statements, Identifier and Citation (2022) used at the University of Glasgow..
- support initiatives to standardise the capture of DAS

Acknowledgements

Several partner suppliers and research organisations volunteered time and expertise to analyse, test, comment and explore the subject of DAS prevalence and quality and we are very grateful to those who have been involved.

Data Availability Statement

Two examples of datasets submitted by institutions (namely University of Glasgow and University of Leicester) for analysis by pilot suppliers can be accessed via University of Leicester institutional repository platform: https://doi.org/10.25392/leicester.data.28675934.

Data supporting the findings in this report are available from Zenodo at: https://doi.org/10.5281/zenodo.15599391

Author Contributions

Formal analysis: Valerie McCutcheon, Mick Eadie, and Laurian Williamson. Investigation: Valerie McCutcheon, Mick Eadie, Laurian Williamson, and Radoslaw Pajor. Methodology: Valerie McCutcheon and Mick Eadie. Project administration: Valerie McCutcheon, Mick Eadie, Laurian Williamson, and Radoslaw Pajor. Writing - original draft: Valerie McCutcheon. Writing - review & editing: Mick Eadie, Laurian Williamson, and Radoslaw Pajor.

References

Murphy, F. and Samors, R.J (2018) Belmont Forum Data Accessibility Statement Policy and Template - Endorsed 18 October 2018. https://zenodo.org/records/1476871

National Information Standards Committee ODI (Open Discovery Initiative) https://www.niso.org/standards-committees/odi

The SCOPE Framework for Research Evaluation SCOPE Framework for Research Evaluation | INORMS

UNESCO Draft Principles on Open Science Monitoring Support for UNESCO Draft Principles on Open Science Monitoring - Science Europe

UKRI open access policy

UKRI open access policy – UKRI

UKRI research data policy

https://www.ukri.org/news/ukri-developing-new-research-data-policy-framework/

Data Access Statements, Identifier and Citation

https://edshare.gla.ac.uk/1419/

UKRN Data Accessibility Statement Pilot Classification Discussion Document

https://doi.org/10.5281/zenodo.15223403

Appendix 3: Data Availability Statements. Annex 1: Specification for pilot: Data Availability Statements

SPECIFICATION for PILOT 4: VERSION 3.0

3rd July 2024

1. Introduction

Data Availability 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.

Underlying research data, code, analysis software and other materials can sometimes be separated out into different statements, and referenced with different names, e.g.: 'Code Availability Statements', 'Analysis Software', 'Availability of Data and Materials', 'Data Access Statements', and so on. For the purposes of this pilot, looking at the Prevalence and Quality of DAS, any statement referencing any type of underlying material will be counted as a DAS.

For the purpose of this document 'materials' is used as shorthand to describe research data (both digital and physical), software, code or supplementary materials.

2. Definition

Following discussion between project partners and suppliers and an overview of selected current literature and advice, the pilot will analyse and categorise data according to the following criteria. The pilot would expect a DAS to:

2.1 Describe where and how materials supporting the results reported in a research output can be accessed.

2.2 Include hyperlinks and persistent identifiers (e.g. Digital Object Identifier (DOI) or accession number where available) if materials are recorded or stored in a catalogue or repository.

2.3 State that materials are available in supplementary files if appropriate.

2.4 Explain why materials cannot be shared openly, for example to protect study participant privacy.

2.5 If a barrier to immediate access exists, provide clear reasoning (e.g. sensitivity or volume) and with a clear point of contact (e.g. access committee with the likelihood of a timely response).

2.6 Clearly separate raw data and any secondary data reuse that supports results and analyses.

2.7 If appropriate, clearly state that no materials were produced in the research.

2.8 Be positioned in a DAS Section or other separate and clearly marked part of the paper.

2.9 Use a DAS template.

3. Research Methodology

DAS' will be counted and analysed initially using methods provided by suppliers and applied to datasets provided by participating institutions.

Datasets will be uploaded and stored in the Pilot 4 FigShare Repository.

The pilot will:

3.1 Start with a small sample dataset.

3.2 Use dataset(s) provided by participating institutions from recently published items.

3.3 Identify the percentage of items at an institutional level that contain a DAS.

3.4 Categorise findings against the criteria outlined in Section 2 above.

3.5 Explore further analysis of DAS according to:

3.5.1 Item type

3.5.2 DAS type: research data (digital or physical), code, software, supplementary materials

3.5.3 Academic discipline

4. Scope

4.1 Dataset(s) will include articles only.

4.2 Dataset(s) will be from items published in 2023 to ensure more recent trends in publishing practice are accounted for.

4.3 Dataset(s) will include items where full text is in the public domain (either Author Accepted Manuscript (AAM) or publisher's Version of Record (VoR) only.

4.4 Dataset(s) will include the following fields as mandatory (See data template for more information: https://doi.org/10.6084/m9.figshare.26165794.v1):

4.4.1 Local Identifier

4.4.2 DOI

4.4.3 Local Uniform Resource Locator (URL)

4.4.4 Type

4.4.5 Year

4.4.6 OrgUnit (Organisation unit)

4.5 The template dataset will provide a baseline that will facilitate consistent analysis of data across participating institutions and across pilots as appropriate.

4.6 Some partner institutions may want to broaden the scope of the analysis locally - perhaps to look at slicing data by funder or by open access type or to broaden the range of data to include other years or item types.

4.7 Each supplier may have specific data requirements and/or be able to provide 'value add' data using their specific tools.

4.8 Extra data and analysis would generally be outside the scope of the pilot. But if useful and widely applicable it could form a part of final reporting and recommendations.

4.9 The data template linked to above includes some optional fields that could be included if an institution thought it was easy to provide and may help with further such analysis.

4.10 When assessing the quality of DAS some finer, possibly manual, analysis of a subset of the larger dataset will be required. For example to explore where appropriate:

4.10.1 If DAS text accurately describes the material it links to.

4.10.2 If the link provided in the DAS resolves to an accessible resource.

4.10.3 If a DAS template is helpful, or conversely, encourages a lack of detail and/or provision of accurate information

5. Issues

5.1 It is possible that not all full text documents will be licensed for this specific reuse by all pilot suppliers. This has been mitigated as far as possible by using only full text documents already in the public domain.

5.2 If required alternative means of access to AAMs could be explored. An optional dataset field that links to the repository or research information system document has been added to accommodate this. Any gaps will be accounted for in the analysis.

Appendix 3: Data Availability Statements. Annex 2: CWTS Review

Pilot 3: Data availability statements

Kathleen Gregory (KG), Thed van Leeuwen (TVL)

This pilot project on Data Availability Statements (DAS) represents a valuable initial contribution to investigate the prevalence and quality of DAS in the literature. By working with four data suppliers and analysing samples of open-access publications from participating institutions, the project team has taken steps toward mapping the prevalence, quality, and discoverability of DAS. The initiative also generated a promising classification scheme and a set of recommendations for measuring the existence and quality of DAS.

Our earlier conversations with the project team gave us confidence in the thoughtfulness and rigour of the approach. The exploratory nature of the pilot meant that some aims and details evolved during the process and are not necessarily explicitly listed in the report itself. It is encouraging to see how the team adapted their focus in response to methodological challenges and conceptual considerations. The report reflects this journey, and we see potential for the work to inform future efforts.

The executive summary and table showing the project timeline are particularly effective in orienting the reader. Bringing together the study's limitations in a single, clearly signposted section would further enhance accessibility for readers. We also believe the report would benefit from moving the contextual "Overview" and the DAS specification earlier, as these sections provide essential framing for the methods and findings. Readers could also benefit from having more details about how the DAS specification was initially developed. The nine point DAS definition is a helpful operationalisation, and acknowledging the multidimensional nature of DAS could enrich this section even further.-point DAS definition is a helpful operationalisation, and acknowledging the multidimensional nature of DAS could enrich this section even further.

Presenting the methodological steps in a more structured format—such as a table or bullet pointed-pointed workflow—would make the robustness of the approach more visible. Having dedicated sections for the methods and findings could also better structure the work. The team's reflections on openness, including the challenges posed by proprietary tools, are thoughtful and important. The detailed description of the PLOS-DataSeer workflow is especially useful; offering a similar high-level-level overview for the other providers, even if brief, would strengthen the transparency of the study. The inclusion of example institutional datasets is an excellent demonstration of openness. A concise description of all institutional datasets (sample sizes, sampling strategies, disciplinary considerations) would complement this nicely.

The project's classification of DAS stands out as a significant contribution with clear potential for wider application, including informing publisher templates and supporting future assessments of DAS quality. The team also raises the point that intellectual property rights / copyright impacted their ability to find and analyse DAS in the full-text of documents. This points to the broader relevance of open research information initiatives, such as the Barcelona Declaration. These observations could be developed further, as they speak to important systemic issues.

The findings are clearly presented in tabular form. A short narrative synthesis of the most important insights and recommendations would help readers navigate the material and provide further nuance. Clarifying the intended audience for each recommendation would also increase their practical value. The finding about differences in the presence of DAS in published versions of record and author accepted versions (which are often analysed by institutions) is also very interesting. This has repercussions for changing both publishing workflows and authors' practices.

The concluding reflections are candid and constructive, acknowledging the challenges of time, resources, and methodological constraints. We encourage the team to continue exploring how emerging DAS indicators might be used responsibly, including their potential benefits and unintended consequences. Counting DAS may be a useful starting point, but contextualisation will be essential for meaningful interpretation. It would be valuable to hear more about how the team envisions the future use of such indicators, particularly in light of frameworks like SCOPE and broader developments in open science.

Overall, this pilot lays nice foundations for future work, which is particularly noteworthy given the voluntary nature of the work. On that matter, we realize that the project was a pilot, and as such does not have the characteristics of a full research project. The outcomes were based on a blend of funded and voluntary time, under tight project timelines. This of course results in outcomes which are different than what would be expected from a fully funded, well-resourced project.

Appendix 3: Data Availability Statements. Annex 3: Team Response to CWTS Review

Thank you for sharing the review of the report UKRN provided summarising Open Research Indicators – Data Availability Statements.

We would like to clarify that, according to the reporting guidelines set out for all pilots, the report was not intended as a formal publication. Instead, it was meant to inform the parties and contribute content toward the development of the comprehensive summary covering the full set of pilots. In this context, our pilot was scoped specifically to explore co-creation of practical methods for monitoring the prevalence of Data Availability Statements (DAS) and assessing their quality within a small, selected dataset provided by partnering institutions. The pilot was not designed to provide a DAS classification framework nor establish a standardised protocol for DAS detection/extraction.

Given the experimental nature of the work, along with limitations in time, resources, varying repository configurations, legal considerations and different structure of data returns across partners, expanding the level of detail or supplementary material beyond what was included would have been challenging. We hope this context helps to clarify the rationale behind the scope of the report.