Appendix 1

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

Authors

Jenni Adams 1†, Angus Taggart 1, Kerry Miller 2, Gillian Currie 2, Mary Donaldson 3, Valerie McCutcheon 3, Kirsty Merrett 4, Jade Godsall 4, Christopher Warren 4

1 University of Sheffield, 2 University of Edinburgh, 3 University of Glasgow, 4 University of Bristol

† Correspondence should be addressed to Jenni Adams; E-mail: j.adams@sheffield.ac.uk.

Executive summary

In this pilot, HEIs and external providers operating in the scholarly communications sphere worked collaboratively to explore the possibility of developing indicators to measure the FAIRness of data underpinning research publications.

Using research publications as a starting point, and with a focus on Data Availability Statements and other references in the publications to underlying datasets, the providers Digital Science, Elsevier, OpenAIRE and PLOS/Dataseer sought to compile data on a set of sub-indicators of FAIRness identified as being of specific interest to institutions. Between them, the providers were able to derive data on five of these sub-indicators: existence of an acceptable PID, presence of a complete metadata set for the data, presence of provenance information, clear statement of access level and conditions, and existence of a licence. Data relevant to a small number of the other selected sub-indicators was also returned, but due to additional complications surrounding the verification process and timescale limitations, it was not possible to evaluate these within the limitations of the study.

Institutions explored the accuracy of the data returned by providers using a manual checking process (Overview Report Annex 5) which resulted in balanced accuracy calculations, in addition to identifying a set of caveats and additional contextual factors within which these numerical assessments must be considered. Institutions also identified the additional steps that would be necessary to derive functional sub-indicators from the data that was returned.

We conclude that while the work conducted here is necessarily preliminary, with no finished and functioning indicators, it has enabled a mapping of the possibilities and some initial work towards developing methods to compile the data which might (with further development and finessing of methodologies) inform the development of sub-indicators for some aspects of data FAIRness.

The process has also enabled institutions to identify a number of areas where authors, publishers and repositories could improve their practice to better enable both good practice in data FAIRness, and the monitoring of this. We have also identified a number of learning points which might inform further work of this nature going forward, including the need for a clear alignment of stakeholder priorities and the necessity of appropriately sequencing work packages.

Background and aims

The FAIR data principles create a standard by which the discoverability and reuseability of research data may be judged. The acronym FAIRstands for Findable (how easy it is to discover the existence and location of a dataset); Accessible (level of openness, and where data is not open, clear documentation of the dataset and if/how it can be accessed; Interoperable (use of open or widely used formats and software and high quality metadata that conforms to accepted standards); and Reusable (provision of the necessary metadata, contextual information and licence to enable reuse).

FAIR data is a key aspect of open research in that it ensures that not only are the data underpinning research publications shared - making it possible in theory for findings to be verified or replicated - but that they are shared in a way that actually enables these processes of verification, replicability and accountability to take place. FAIRness of data thus enables the actualisation of many of the key practices and goals of open research; for this reason, this project sought to develop means of evaluating publications with regard to the FAIRness of their accompanying data. This pilot has close links to the pilot on the openness of data (see Appendix 2), indicating the desirability of some shared vocabulary and definitions, and has dependencies on the work conducted for the pilot on Data Availability Statements (DAS, see Appendix 3) for the reasons detailed below.

We considered two main ways to approach the assessment of the FAIRness of data. The first would be to look at data in trusted repositories and to evaluate the FAIRness of the deposited items. However, we felt that this would bypass the question of Findability and would also involve focusing mainly on positive results - instances where researchers had made their data available in a repository, thereby already covering many of the key criteria of FAIR - creating an unrepresentative picture. Instead, we decided to proceed from published articles, emulating the experience of a user of the research. This necessitates using Data Availability Statements and other references to underlying datasets in order to obtain the relevant information to evaluate the FAIRness of the published research. For this reason, there is a close relationship between this pilot and that on DASs, which might be considered logically prior to this pilot.

Previous monitoring of FAIR data

Various existing attempts have been made to evaluate the FAIRness of data. The FAIR-Aware tool developed by the FAIRsFAIR project supports self-assessment of FAIRness by researchers at the point of data acquisition or publication,[1] while the F-UJI tool, developed in the same project, allows automated assessment of datasets in trusted repositories to evaluate their readiness for reuse. An open source API machine learning tool, F-UJI uses both its own code and FAIR-enabling services from repositories and third-parties to validate metadata standards against authoritative resources. While not mapping exactly onto the needs of the current project, these tools offer valuable comparative approaches, and the FAIRsFAIR Data Object Assessment Metrics which were developed as a precursor to the tools articulate in a valuably clear manner the specific conditions which would need to be met for the various aspects of FAIRness to be fulfilled. As detailed below, we drew substantially on these metrics in our own approach. Other existing tools and resources which informed the project included the RDA-SHARC templates for FAIRness evaluation criteria,[2] the RDA FAIR Data Maturity

[1] Huber, R., Cepinskas, L., Davidson, J., Herterich, P., L'Hours, H., Mokrane, M., von Stein, I., & Verburg, M. (2021). D4.5 Report on FAIR Data Assessment Toolset and Badging Scheme (V1.0). Zenodo. https://doi.org/10.5281/zenodo.6656444

[2] David, R., Mabile, L., Specht, A., Stryeck, S., Thomsen, M., Yahia, M., Jonquet, C., Dollé, L., Jacob, D., Bailo, D., Bravo, H., Gachet, S., Gunderman, H., Hollebecq, J.-E., Ioannidis, V., Le Bras, Y., Lerigoleur, E., Cambon-Thomsen, A., & SHARC Community. (2020). Templates for FAIRness evaluation criteria - RDA-SHARC ig (1.1) [Data set]. Zenodo. https://doi.org/10.5281/zenodo.3922069

model,[3] the FORCE-11 Guiding Principles for FAIR Data Publishing,[4] and the Australian Research Data Commons' FAIR Data Self Assessment Tool.[5]

Methods

The universities of Sheffield and Edinburgh were the lead institutions for this pilot, core members of which included Jenni Adams, Angus Taggart, Gill Currie and Kerry Miller. Several other members of the project team were affiliated with other institutions. The team proposed working with industry partners ('partners' or 'providers') to establish the feasibility of the above sub-indicators and create (or adapt or apply existing) algorithms/workflows to produce these measures where feasible. Full details of participating organisations are tabulated below:

Table 1. Participating organisations.

Pilot	Participating institutions (leads)	Participating providers
1: FAIRness of data	Edinburgh, Sheffield, Bristol, Exeter, Glasgow, Liverpool, Reading, Surrey	Digital Science, Elsevier, OpenAIRE, PLOS/DataSeer

The team began by reflecting why it was important to be monitoring open research at all,[6] identifying a number of reasons:

- To enable the identification of areas of existing good practice and areas for improvement in uptake of open research practices.
- To provide data to faculties and departments that will assist them in targeting interventions in order to improve the transparency and replicability of research in specific subject areas, and also aid planning at faculty and departmental levels.
- To inform medium and long term planning around open research advocacy and infrastructure.
- To inform development of standard training in all aspects of OR for postgraduate students, ECRs, and PIs which will meet needs.
- To support policy development and identify gaps in support and services.

Closing in on the specific research question, FAIR data and software are key aspects of open research that are central to UK research institutions' Open Research strategies. Support personnel are providing training, tools and support to enable researchers to achieve FAIR data, but there is little formal monitoring of how widely the practices are followed. The team

[3] FAIR Data Maturity Model Working Group. (2020). FAIR Data Maturity Model. Specification and Guidelines (1.0). Zenodo. https://doi.org/10.15497/rda00050

[4] https://force11.org/info/guiding-principles-for-findable-accessible-interoperable-and-re-usable-data-publishing-version-b1-0/

[5] https://ardc.edu.au/resource/fair-data-self-assessment-tool/

[6] The INORMS SCOPE Framework provides a helpful structure for documenting reflections that are important in the UKRN work, so we adapted this method to suit the aims of the pilot.

was keen to investigate whether it is possible to identify areas of good practice and track progress in adoption behaviours. This in turn would have implications for Research Culture Strategy work and other future initiatives. Consequently, this pilot sought to establish the measurability of “FAIRness” in research datasets that underpin published journal articles.

The approach taken

FAIRness is not a unitary quality - the acronym itself breaks down to four categories which themselves each contain several sub-components. For this reason, we decided to identify a set of sub-indicators of FAIRness which would span all categories of F, A, I and R and provide reasonably indicatory data on the FAIRness of an output or body of outputs, while stopping short of an exhaustive attempt to cover all components, something that would be unachievable within the scope of a pilot study.

To inform this process and to ascertain the current state of knowledge in this field, the team performed a literature review and identified twenty items that, at first look, were potentially relevant in seeking useful context around FAIR indicators and sub-indicators, so could be used to refine a methodology for this project. In particular, we sought and reviewed detailed articulations of the FAIR principles and the evidence needed to establish whether each individual sub-principle had been met. These were then reviewed by members of the project team.

The most relevant initiatives focused on automated tools and machine-readable metadata for FAIR data, while those that provided manual assessment templates or guiding principles were discarded because they required human intervention rather than automation. The literature review identified the FAIRsFAIR Data Object Assessment Metrics,[7] the RDA-SHARC Interest Group templates for FAIR evaluation,[8] and the 17 core metrics identified in Devaraju and Huber (2021)[9] as especially useful, and the team drew primarily on the former in our selection and articulation of a subset of identifiers of FAIR as areas of interest for the pilot.

The selected sub-indicators, together with information about the source from which they were derived, are listed in the UKRN_FAIR_Indicators_preliminary_indicator_subset document embedded here:

UKRN_FAIR_Indicator
s_preliminary_indicat

They are summarised below:

2.1 Findable

2.1.1 Establish whether the data underpinning the publication is assigned one or more Persistent Identifier/s which resolves to a landing page from which the data can be accessed, or, if the data cannot be shared, to a metadata-only record.

2.1.2 – Establish whether the assigned PIDs are provided by an established, globally recognised provider.

[7] Devaraju, A., Huber, R., Mokrane, M., Herterich, P., Cepinskas, L., de Vries, J., L'Hours, H., Davidson, J., & Angus White. (2022). FAIRsFAIR Data Object Assessment Metrics (0.5). Zenodo. https://doi.org/10.5281/zenodo.6461229

[8] David et al, 2020.

[9] Devaraju, A., & Huber, R. (2021). An automated solution for measuring the progress toward FAIR research data. *Patterns (New York, N.Y.)*, *2*(11), 100370–100370. https://doi.org/10.1016/j.patter.2021.100370

2.1.3 – Establish whether the associated metadata includes the following core elements: creator, title, data identifier, publisher, publication date, summary and keywords.

2.2 Accessible

2.2.1 – Establish whether metadata contain access level and access conditions of the data.

2.3 Interoperable

2.3.1 Identify whether the metadata of the object are available in a formal knowledge representation language.

2.3.2 - Identify whether namespaces of known semantic resources (excluding common namespaces, e.g., RDF, RDFS, XSD, OWL) are present in the metadata of an object .

2.4 Reusable

2.4.1 Identify whether the metadata for the dataset includes provenance information about data creation or generation.

2.4.2 Identify whether the metadata for the dataset includes the reuse licence applied to each dataset cited: this might be **Creative Commons licences**, public domain, bespoke, or otherwise.

2.4.3. Identify whether the dataset uses **open or accessible file formats**.

a. If not open format(s), are data available in a file format recommended by the target research community.

For the purposes of this pilot, 'data' and 'dataset' could include all materials gathered and/or created to inform research findings, and some providers included references to code[10].

For each of these sub-indicators, we highlighted component questions which would need to be answered in order for the criterion to be considered fulfilled. These are detailed in the Pilot specification document (Annex 1) which was presented to providers as a starting point for the work. The phases of the project were as follows:

1) **Completion of INORMS-SCOPE template (see Overview Report Annex 2).**
2) **Literature review and identification of sub-indicators of particular interest; development of pilot specification**.
3) **Specification shared with providers.** The institutional project team met with each of the providers to present the specification, address queries and make clarifications where needed. We acknowledged that it may not be feasible to create workflows to measure all of the selected sub-indicators, but were interested in the provider's response to and investigation of which of these it may be feasible to measure.
4) **Dataset shared with providers.** As a test dataset to work with in developing the sub-indicators, institutions participating in this and other projects each shared with providers a dataset of CRIS data relating to articles published during the 2023 calendar year and including the following fields: LocalID; [article] DOI; Link; PageLink; [item] Type; Year [of publication]; School/Faculty/Department. CRIS data is in the public domain, and this project did not involve sharing commercially or personally sensitive information at any stage. This format was proposed by the Data Availability Statement

[10] In fact there was some ambiguity about this. For example PLOS/DataSeer did not understand this as a core part of the specification, and could have done more if they had. They did include some information on code generation and sharing in their data output, see below.

Pilot team and agreed as standard, and datasets were uploaded by institutions to a FigShare project accessible to all providers. (See Appendix 1 Annex 2 for the full description of the test dataset.)

5) **Providers undertook the sub-indicator development processes.**
6) **Providers returned to institutional partners datasets** which used the test data as a starting point to derive additional data relevant to the evaluation of one or more of the identified sub-indicators. Institutions met with providers to give initial feedback, seek clarification and in some instances request additional data fields be included. In the latter instance/s, providers then issued revised datasets. NOTE: these datasets were in some instances received quite late in the process, creating limitations regarding the extent of the work that could be performed on them in phase (7) below.
7) **Institutions undertook a manual checking process** to evaluate the accuracy of the data returned with reference to the specific sub-indicators under development and understand the qualities of the methods used by the provider[11]

It was noted in advance there would likely be some issues with gathering and collating all this information: for instance, not all full-text documents will be licensed for use by pilot partners. Accordingly, there was an agreement to focus on open access publications for the purposes of the study. Similarly, the ability of pilots to share their findings based on provider tools/algorithms needed to be addressed with each provider individually to check if there was a risk of breaching providers' intellectual property rights by sharing datasets produced using their proprietary algorithms, or risks related to institutional concerns about misunderstanding the findings.

Findings

Following the data returns from providers, the institutional project team created a mapping document (Annex 3) to map the data fields returned by providers against the sub-indicators of specified interest to the pilot. The outcomes of this, in terms of which sub-indicators were reported on by which provider, are summarised in Table 2.

[11] In some cases providers routinely provide accuracy assessments for their standard indicators following standards/ good practices including precision, accuracy, F1 score etc. PLOS and DataSeer publish accuracy assessment statistics and confusion matrices for the regular OSIs, in their public releases. They did not do a separate accuracy assessment specifically for the UKRN data analysis due to resource constraints. They note that, while an accuracy assessment by institutions has value for the pilots, the accuracy assessment is specific to that context and not a measure of accuracy of the providers' tools in general.

Table 2. Summary of mapped sub-indicators for Pilot 1

X indicates data we were able to include in the test

[x] indicates that some relevant data was returned, but we were not able / decided not to include this in the test for the reasons stated in the mapping document (Annex 3).

	2.1.1 PID	2.1.2 PID provided by globally recognised provider	2.1.3 Dataset has complete metadata set	2.2.1 Dataset metadata specify access level & conditions	2.3.1 Dataset metadata are available in a formal knowledge representation language	2.3.2 Namespaces of known semantic resources are present in dataset metadata	2.4.1 Dataset metadata includes provenance information	2.4.2 Dataset metadata includes licence	2.4.3 Data deposit uses open or accessible file formats
	F	F	F	A	I	I	R	R	R
Digital Science	X			[X]				[X]	
Elsevier	X		X	X			X	X	[X]
OpenAIRE	X		X	X	[X]		X	X	
PLOS/DataSeer	X								

For sub-indicator 2.1.2 (PID provided by a globally recognised provider), all of the identifiers accepted as PIDs fulfilled this condition, so a decision was made to exclude this as a separate sub-indicator.

PLOS/Dataseer also returned data on whether data was generated during the study in question, which (if found to be accurate) could be valuable as a denominator for the other measures. It was decided to include this data in the manual checking process.

The data that was returned, and how it was derived

Digital Science

The original data provided by institutions was returned with additional fields added, including, for each article: Dimensions id, DOI, the field 'DAS Repository URL identified', which listed links to datasets underpinning a given article, and from which the existence of a PID on the list of those considered acceptable could be derived.[12] While material relating to the classification of DASs themselves was also returned, this was not evaluated within Pilot 1 given that it is the main focus of Pilot 4 (DAS).

In order to create an indicator from this return, it would be necessary to include an additional algorithm to assess whether the DOIs, URLs or accessions returned in this field matched with the agreed list of acceptable PID types, generating a positive or negative response. The positive responses could then be presented as a proportion of (a) the records for which data was generated (using a method like that employed in the PLOS workflow) or (b) the total number of records minus those that indicate in a DAS that data was either not collected or could not be shared for ethical reasons. To create an indicator, it would also be necessary to have a clear and reproducible step-by-step method for the derivation of the data return.

Table 3. Methodological details for Digital Science data return.

Sub-indicator	Method by which the data return was produced
2.1.1 PID: Establish whether the data underpinning the publication is assigned one or more Persistent Identifier/s which resolves to a landing page from which the data can be accessed, or, if the data can [no longer] be shared, to a metadata-only record.	Details of dataset identifiers were returned in the field 'DAS Repository URL Identified', which was derived as follows: The Dimensions Research Integrity Dataset was used to identify whether a DAS was present. Digital Science note that 'Dimensions Research Integrity preprocessing begins by converting the PDF to text strings, in order to extract and isolate segments [...]. Each isolated text segment is then labelled and used for training, evaluation, and validation within machine learning models.'. According to discussion between the provider and institutional project team, this text mining of the DAS was the source for data in the field 'DAS Repository URL'. This

[12] A list of acceptable PIDs for the purposes of the pilot was agreed by institutional partners, taking into account those that were identified as acceptable by providers. The comprehensiveness of this list could be expanded for future work as required.

	means that references to the dataset that were contained elsewhere in the article but not within the DAS itself would not be captured.

Elsevier

The original data provided by institutions was returned with additional fields added, with one limitation: the Elsevier project work focused solely on article DOIs for content licensed by Elsevier, for reasons of operational ease. Elsevier noted that this does not represent a significant limitation in terms of the scope and possibility of future work.

Another key point to note is that the Elsevier return only contains records for which a dataset record has been identified. This means that only positive (in a very general sense) results have been reported, making it impossible to check (for example) for false negatives - articles for which no underpinning dataset was identified, but which do refer to a dataset - as this data is not present.

The methodology behind the creation of the data is detailed in the report, and is summarised in brief below.

For all of the below sub-criteria, Elsevier extracted the text of DASs from Elsevier-licensed content listed in the dataset provided by institutions. This was achieved by:

1. Matching the article DOIs with ScienceDirect content in XML format.
2. Applying a regular expression to find sections with headers that included phrases commonly used to preface a Data Availability Statement; extracting the header and following section of text where this was the case.
3. Using a regular expression to extract DOIs and URLs from the DAS.
4. An LLM (GPT-4 via Azure OpenAI) was also used on the DAS texts to identify and resolve additional links. The tool was prompted to find DOIs, other identifiers, repository names, publication dates, repository URLs, information about openness, and entity types (data or code).

Links in the DASs were resolved to access underlying datasets. Access to the data links provided in DASs was supported by Elsevier's analytical and assessment tools, especially Research Data Monitor. A key steps here was:

5. Matching the data DOIs with Elsevier's internal Research Data Monitor dataset in Elsevier's internal data lake, or if unavailable for the DOI, making an API call, in order to find more metadata about the datasets. (Elsevier's Research Data Monitor serves as an aggregator for metadata from various data repositories, including DataCite, Zenodo, Digital Commons, and others.)

Colleagues from Elsevier indicated that no proprietary code was used, but repository-specific code had to be generated for each repository to extract the data, making the process time-consuming.

In order to create indicators from the data return (if/where data is sufficiently accurate), the following steps would be needed:

1. Obtain a data return based on *all* article DOIs published in the given period, rather than solely Elsevier-licensed material, and including negative results (no dataset identified) as well as positive results (dataset identified)

2. For each of the sub-indicators, create an additional algorithm to assess whether the data returned by the provider (e.g. '["cc-zero"]' or '[null,null]' in the 'reuse_licence' field) indicates that the feature is present or absent.
3. For each of the sub-indicators, calculate the percentage of records for which the feature is present, as a proportion of (a) the records for which data was generated (using a method like that employed in the PLOS workflow) or (b) the total number of records minus those that indicate in a DAS that data was not collected, or could not be shared for ethical reasons.

In order to create indicators, it would also be necessary to have clear and reproducible step-by-step methods for the derivation of the data return regarding the individual sub-indicators.

Table 4. Methodological details for Elsevier data return.

Sub-indicator	Method by which the data return was produced
2.1.1 PID: Establish whether the data underpinning the publication is assigned one or more Persistent Identifier/s which resolves to a landing page from which the data can be accessed, or, if the data can [no longer] be shared, to a metadata-only record.	See above
2.1.3 Complete dataset metadata set: Establish whether the associated metadata includes the following core elements: creator, title, data identifier, publisher, publication date, summary and keywords.	See above
2.2.1 Access conditions for dataset: Establish whether metadata contain access level and access conditions of the data.	See above
2.4.1 Provenance of dataset: Identify whether the metadata for the dataset includes provenance information about data creation or generation.	See above
2.4.2 Licence of dataset: Identify whether the metadata for the dataset includes the reuse licence applied to each dataset cited: this might be Creative Commons licences, public domain, bespoke, or otherwise.	See above

OpenAIRE

Note: Subsequent to the main analysis, the OpenAIRE team identified a bug in the algorithm used. Readers should note both updates to the methods and revised results in files embedded below. Unfortunately there was insufficient time to manually check these revised results. Readers should therefore be aware that the findings outlined in the main report below for OpenAIRE are likely to be inaccurate.

Annex X – OpenAIRE Contribution_ Update

Pilot 1 Results revised.xlsx

In contrast to the other providers, rather than returning details of the dataset, licence type, access level, etc for each article DOI, the OpenAIRE return provides a numerical count per article of the underpinning datasets for which each of the criteria are met.

The OpenAIRE methodology is set out in the OpenAIRE Pilot 1 methodology document:

OpenAIRE Pilot 1 methodology.docx

It is summarised for the relevant indicators below. Overall, the following techniques were used:

1. **Level 1**: Utilize enriched metadata records available in the OpenAIRE Graph.
2. **Level 2**: Use methods to query Persistent Identifiers (PIDs) and retrieve additional metadata and information.
3. **Level 3**: Extract insights from the mining of Data Availability Statements (DAS).

In comparison to the data returns from the other providers, there would be fewer additional steps required to create indicators from this return provided the data is accurate – this would largely involve ensuring that the overall numbers of articles with a positive result was presented as a proportion of (a) the records for which data was generated (using a method like that employed in the PLOS workflow) or (b) the total number of records minus those that indicate in a DAS that data was either not collected, or could not be shared for ethical reasons, rather than using as a denominator the total number of articles, which we assume to be the case in the aggregated data returned by OpenAIRE. In order to create indicators, it would also be necessary to have clear and reproducible step-by-step methods for the derivation of the data return regarding the individual sub-indicators.

Table 5. Methodological details for OpenAIRE data return.

Sub-indicator	**Method by which the data return was produced**
2.1.1 PID: Establish whether the data underpinning the publication is assigned one or more Persistent Identifier/s which resolves to a landing page from which the data can be accessed, or, if the data can [no longer] be shared, to a metadata-only record.	Derived from the deduplicated article record in the OpenAIRE graph in combination with DAS mining. The OpenAIRE graph identifies relationships between PIDs and the nature of the relationship. DAS-mining: an algorithm uses regular expressions and context-aware mining rules to identify the presence of Data Availability Statements (DASs) and extract relevant text snippets related to these statements.
2.1.3 Complete dataset metadata set: Establish whether the associated metadata includes the following core elements: creator, title, data identifier, publisher, publication date, summary and keywords.	Extracted from the dataset metadata in the OpenAIRE graph. The provider then also performed dedicated calls to the DOIs to check the well-known parsable, formal metadata representations
2.2.1 Access conditions for dataset: Establish whether metadata contain access level and access conditions of the data.	From the OpenAIRE graph, the provider extracted Access Rights information (Open Access/Embargoed/Restricted/Closed Access/Not Available) and also licence metadata information, where applicable, in order to understand and classify whether access levels and conditions are stipulated. Secondly, the information was extracted where available from the DAS-mining process (technically open data, legitimate reasons for controlled, closed, or no longer available, author-controlled).
2.4.1 Provenance of dataset: Identify whether the metadata for the dataset includes provenance information about data creation or generation.	Extracted from the dataset metadata in the OpenAIRE Graph.
2.4.2 Licence of dataset: Identify whether the metadata for the dataset includes the reuse licence applied to each dataset cited: this might be Creative Commons licences, public domain, bespoke, or otherwise.	Extracted from the dataset metadata in the OpenAIRE Graph. Secondly, the information was extracted where available from the DAS-mining process.

PLOS/DataSeer

The original data provided by institutions was returned with additional fields added, including, for each article DOI, ‘Data_DOIs’ (a list of DOIs provided in the article that were detected by the DataSeer algorithm as linking to a dataset); ‘URL_Data’ (A list of any URL/DOIs provided in the article that were detected by DataSeer algorithm as hosting data or code online; ‘Accessions’ (a list of Accession IDs from repositories that host data or code online that were detected by the DataSeer algorithm) and ‘Data_Generated’ (was data generated in this text as assessed by DataSeer algorithm). See Overview Report Annex 4 for the anonymised PLOS/DataSeer dataset.

PIDs for datasets and code were reported separately by the provider, and for this reason were evaluated separately during the manual checking procedure. This is an exception to our agreed practice of using a broad definition of ‘data’ to include both data and code. For this pilot, PLOS/DataSeer did not create new algorithms or processes but applied their existing method for measuring data and code sharing in publications, that is produced regularly as part of PLOS Open Science Indicators. The existing indicators and outputs from this process were mapped to the pilot specifications in producing the data output for the pilot. The PLOS-DataSeer mapping file is embedded below.

UKRN pilots &
PLOS-DataSeer OSI re

More detail on the methods, accuracy, and open data from their application in a non-UKRN pilot context are available from: Public Library of Science (2022). PLOS Open Science Indicators. Public Library of Science. Dataset.

https://doi.org/10.6084/m9.figshare.21687686.v9

In order to create an indicator from this return, it would be necessary to create an additional algorithm to assess whether the DOIs, URLs or accession IDs returned in these fields matched with the agreed list of acceptable PID types, generating a positive or negative response. The positive responses could then be presented as a proportion of the records for which data/code was generated. In order to create an indicator, it would also be necessary to have a clear and reproducible step-by-step method for the derivation of the data return.

Table 6. Methodological details for PLOS/ Dataseer data return.

Sub-indicator	**Method by which the data return was produced.** *Methods described fully in PLOS (2022)*[13] *and specifically for pilots see [*[14]*]*

[13] Within this publication there is a file called ‘OSI-Methods-Statement_v9_Dec2024’ and section of the text called ‘Data and Code Generation’ and ‘Data and Code Sharing’.

[14] We used the lists of DOIs for articles published in 2023 that were provided by pilot institutions in the shared project folder. These DOIs were combined into one spreadsheet and appended with licence information from Unpaywall by Digital Science. For simplicity we focused only on the 19138 articles with a CC BY licence.

2.1.1 PID: Establish whether the data underpinning the publication is assigned one or more Persistent Identifier/s which resolves to a landing page from which the data can be accessed, or, if the data can [no longer] be shared, to a metadata-only record.	Gazetteer recognises data DOIs - starts with DAS then also searches full text of articles for data DOIs; converts URLs containing DOIs to DOIs; lists these in the data return. Separate detailing of identifiers for data and code.
Data generated: was data generated in the study? - this was not itself one of the stated sub-indicators, but could be useful in deriving the sub-indicators, providing a denominator for the other measures	Extracting phrases which indicate data generation using BERT model and returning a TRUE/FALSE response.

Assessing the accuracy of the data returned

In evaluating the data returns from providers, we opted to use the manual checking process (see Overview Report Annex 4) shared by UKRN. For each of the indicators we deemed it feasible to test (see the mapping document in Annex 3 and summary table 1.1 above), we carried out a sample check of 10 records and subsequently deemed each sub-indicator 'relatively well-defined and relatively common', meaning that we then proceeded to complete a check of 100 randomly selected records, recording for each sub-indicator in each of the provider's data returns:

- Whether the provider deemed the feature to be present
- Whether a manual check showed the feature to be present. Two different members of the institutional project team completed separate checks; the few instances where there was any disagreement were discussed and resolved during project team meetings.

A README was created for each manual check to clarify the process and ensure consistency between checkers. 'Notes' fields were used to record any particular observations during the checking process. Once the manual checks were complete, a balanced accuracy template spreadsheet was used to calculate balanced accuracy scores for each provider and sub-indicator. Results are summarised below.

We found 6437 of these articles on PubMedCentral (PMC) and successfully downloaded the xml version of the manuscript for all of them. Another 9421 articles were located on OpenAlex or via biblioglutton; of these we were able to download 6496 articles in pdf format.
We processed the PMC xml files through DataSeer's GenShare pipeline, obtaining results for 5901 articles. The pdf files from OpenAlex were converted to TEI (a form of xml) with Grobid and then processed with the TEI-specific version of GenShare. After removing duplicate DOIs and discarding articles that could not be processed by GenShare, we obtained results for 6011 of the pdf articles. These 11912 articles were analysed for data and code generation, and for data and code sharing.
The articles were additionally analysed using the open source DAS classifier tool from Colavizza et al (2020).
The articles were additionally analysed using the preliminary study registration (preregistration) indicator developed by PLOS and DataSeer in consultation with Pilot 6 institutions.

Summary of results from manual checking process - tabulated for each provider

[Note that the OpenAIRE figures are known to be inaccurate; see explanation above]

Digital Science

Table 7. Balanced accuracy calculations for Digital Science data return.

Sub-indicator	Balanced accuracy score (to 3 decimal places)
PID	0.881

See Discussion section below for caveats and contextual notes

Elsevier

Table 8. Balanced accuracy calculations for Elsevier data return.

Sub-indicator	**Balanced accuracy score (to 3 decimal places)**
PID	0.988
Complete metadata set	0.662
Provenance information	0.667
Access level / conditions	0.307
Licence	0.842

See Discussion section below for caveats and contextual notes

OpenAIRE

Table 9. Balanced accuracy calculations for OpenAIRE data return.

Sub-indicator	**Balanced accuracy score (to 3 decimal places)**
PID	0.608
Complete metadata set	0.618
Provenance information	0.596
Access level / conditions	0.613
Licence	0.590

See Discussion section below for caveats and contextual notes

PLOS/DataSeer

Table 10. Balanced accuracy calculations for PLOS/DataSeer data return.
(See also PLOS 2022)

Sub-indicator	Balanced accuracy score (to 3 decimal places)
Data generated	0.974
PID (data)	0.907
PID (code)	0.810

See Discussion section below for caveats and contextual notes

Discussion

Limitations to the results, and additional observations

Digital Science

Caveats and contextual observations

The process involves noting whether the provider had indicated the feature is present, then noting whether the feature is in fact present. This leaves the possibility that the provider identified a DOI which is actually not for a dataset - it may be for the article itself, for example - and the manual check also identifies a (different) DOI that *is* for a dataset - in this instance, it appears that the provider's response has been verified by the check, when in fact only one of the two positive responses is accurate. This should be noted as a limitation of the overall methodology of the manual checking process.

In some instances, there were issues with the DOIs extracted by the provider - they were garbled or incomplete, indicating that this process would benefit from further finessing. In some instances the model failed to pull out a PID from the DAS, while in others the DAS existed but was not identified.

There is not a simple 1:1 relationship between an article and related data as presumed by the existing models for indicating FAIRness. We found a diverse set of relationships like "source material used in the creation of", "is an output of this research", "Software Tooling used" or "data used in an alternative study". Partial documentation in the DAS was common. Without a formal method of indicating the related material and its relationship to the paper, confirming that a full set of related material had been documented in the pilots was very difficult.

Elsevier

Caveats and contextual observations

As noted above, the Elsevier project work focused solely on article DOIs that represent content licensed by Elsevier, for reasons of operational ease. The articles included in the data return are hence only a subset of those provided by institutions, resulting in a smaller sample, and one that may also be unrepresentative given potential disciplinary skews and any particularities of practice regarding DASs in Elsevier-licensed journals.

As noted above, another key caveat is that the Elsevier return only contains records for which a dataset record has been identified. This means that only positive (in a very general sense) results have been reported, making it impossible for the team to examine (for example) false negatives - articles for which no underpinning dataset was identified, but which do refer to a dataset - as this data is not present. The accuracy scores tabulated in the results section above must be considered in this context.

Specific limitations to the sub-indicator for access level and conditions were also noted. The data return was not able to specify (1) reasons that may have been given for a lack of open availability of data; and (2) the point of contact listed for accessing author-controlled data; though in discussion, Elsevier colleagues noted that it may be possible to report on these details in future through the development of the LLM approach. Currently the 'access_conditions' field is binary, meaning that this only tells us that the access conditions for data that are not openly available are stated, not what they are.

During the manual checking process, the presence of a CC licence was taken as an acceptable indication of the openly available nature of a dataset. On the topic of licences, it should also be noted that there is not a 1:1 relationship between data deposits and licences, and the existence of a licence does not mean there is a licence for every component (e.g. all files; data and code).

In a number of instances, checkers observed that while more than one dataset was listed in a DAS, the provider had only identified one of these (or fewer than the total amount). Occasionally the PID or URL given by the provider was not a correct parsing of the information in the DAS and did not resolve.

For the complete metadata and provenance information checks, we observed that in some instances the publication date extracted by the provider was not actually the publication date but another date from the dataset metadata, such as the technical metadata creation date, a 'last updated' date, or a date prior to the record's publication (possibly, the date the submission was drafted). In a few instances, the provider captured metadata from the article itself (e.g. with regard to title or creators) rather than the dataset.

OpenAIRE

Caveats and contextual observations

Due to the numerical presentation of the data return (e.g. 'Number of datasets with an acceptable PID: 2'), it has not been possible/practicable to ensure we are talking about the same dataset/s as those the provider believes they have identified. This is a variation of the caveat noted for PLOS/DataSeer and Digital Science, i.e. that a positive response from the provider and a positive response from the manual check would appear to indicate the veracity of the provider return, but the provider and check may refer to the features of different potential datasets (not all of which, from the provider perspective, are necessarily correctly identified as underpinning datasets of the article). The numerical presentation meant that in some ways the data return was closer to the form of an indicator (rather than data that could be developed into an indicator), though one drawback to this is that the criteria for a positive response were not always transparent.

An additional caveat relating to these numerical quantities is that for reasons of practicality, coupled with the required binary output of the checking process (Is the feature present or not), we did not verify the exact numbers of datasets identified as possessing a given feature (e.g. 'Number of datasets with an acceptable PID: 2'), only that AT LEAST ONE of the dataset/s underpinning the article possesses the feature. Our anecdotal evidence was that these numerical values contained a large number of false positives.

Where an article DAS (rather than the provider return, the figures in which were often larger than the actual number) indicated that there were more than six datasets, we excluded the record from the random check given the time commitment that would be involved in checking a large number of datasets for one individual record. This may have introduced a minor disciplinary skew away from subject areas likely to draw on a large number of available datasets.

During the manual checking process, the presence of a CC licence was taken as an acceptable indication of the openly available nature of a dataset. On the topic of licences, it should also be noted that there is not a 1:1 relationship between data deposits and licences, and the existence of a licence does not mean there is a licence for every component (e.g. all files; data and code).

PLOS/Dataseer

Caveats and contextual observations

On the separation of data and code (which were reported separately by PLOS/DataSeer with reference to presence of PIDs), it should be noted that this is not in practice clear-cut. For example, in many cases a deposit identified in the DAS as a data deposit will also contain code without stipulating this fact. This makes the process of reporting on data and code PIDs separately potentially challenging, and/or places limits on the degree of accuracy that is obtainable.

It should also be noted that there is not a 1:1 relationship between an article and a dataset, meaning that a positive response re one dataset does not mean that all datasets underpinning the article necessarily align with the given sub-indicator of FAIRness.

Additional observations from the manual checking process as a whole (not provider-specific)

A range of problematic author practices[15] were observed, including:

- Stating only ‘Data available on request’ with no further information
- Giving confusing or contradictory information in the DAS
- Mentioning that data is available in a named repository but giving no further information, or only a generic link to the repository rather than a specific identifier
- Claiming that data will be or is publicly available but giving no details or identifiers
- Giving an incorrect PID in the DAS that does not resolve
- Giving a URL (e.g. a UK Data Service or Open Science Framework URL) in the DAS for a dataset that has a PID, instead of giving the PID itself

Problematic publisher behaviours observed included

- Presenting the "full text" of an article without including the content of the DAS and additional metadata, making the information less easily findable and accessible.
- There was also a lack of standardisation in where in the article and its metadata statements of data availability could be found (if at all).

In terms of repository practice, we noted that in many cases, while the data was openly available, there was no explicit statement or clarification of this in the visible metadata of the

[15] PLOS noted that their analysis also highlighted the prevalence of ‘private for peer review’ links instead of DOIs in DASs, observing it quite often with Figshare and Dryad datasets for example, linked to a timing issue of when the DOI rather than a private link is available for a dataset linked in a DAS.

repository record. We also noted that Zenodo does not appear to have options for software licences, meaning that authors had to select 'other-open' for code archived from GitHub that did have a specific licence, e.g. GPL-3.

We observed that it was common for dataset metadata to contain a complete metadata set apart from keywords. This raised the point that excluding keywords from the definition of a complete set might give a more informative indication as to the appropriate documentation of a dataset in the metadata.

Summary of the findings

The necessarily preliminary nature of the work conducted during the pilot means that it has stopped short of producing the desired sub-indicators of FAIRness; consequently, it is not currently possible to achieve the aims identified in the project's INORMS-SCOPE review. Nevertheless, and despite the early maturity stage of the current work, the project has indicated that with additional efforts, it may be possible to develop sub-indicators to measure at least the following attributes of datasets underlying published research:

- Existence of an acceptable PID
- Presence of a complete metadata set
- Presence of provenance information
- Clear articulation of access level/conditions
- Provision of licence information

This is not to say that with more time and an opportunity to build on prior findings, it may not also be possible to address other sub-indicators among those identified as being of particular interest to this pilot.

The findings from the work conducted by providers show potential in developing the above-listed sub-indicators, but more work would be needed in order to:

- Optimise the accuracy of the methodologies developed.
- Address the caveats detailed in the Discussion section of this report, especially (for those providers that were able to explore the most complete range of sub-indicators), the need for a more complete data return to fully assess provider methodologies in the case of Elsevier, and, in the case of OpenAire, the prevalence of false positives in terms of the numbers of identified datasets (which consequently affect subsequent measures).
- Develop functioning indicators from the data returned - this would involve:
    - For the majority of the providers (aside from OpenAire, whose methods already achieve this), incorporating into the process a stage by which data is evaluated to assess whether a criteria (e.g. presence of an acceptable PID) has been met.
    - For the majority of the providers (aside from PLOS/Dataseer, whose methods already include this), developing and testing a calculation of an appropriate denominator which does not include all published articles, but only (a) those for which data was generated or (b) the total number of publications minus those that indicate in a DAS that data was not collected or could not be shared for ethical reasons.
    - Producing and making openly available a clear and reproducible step-by-step method for the creation of the indicator.

Together with some of the reflections highlighted below, these are the challenges that would need to be addressed in order to meet the goals it was not possible to meet within the pilot itself.

Suggested next steps

Additional work is needed to proceed from the current preliminary position to a place of having openly mapped and reproducible methods towards sub-indicators for FAIR data. It is beyond the scope of the pilot to specify the form such work should assume. Some more immediate next steps which may be considered by UKRN and pilot institutions include the following:

- Perhaps a prerequisite to other steps would be dialogue among institutions and publishers and funders on what they want to measure, to align on more common definitions and approaches. That could eventually be captured as metadata in articles.
- Dialogue with all relevant stakeholders on how best to enable better quality data sharing approaches that would result in and be evidenced by better quality DAS.
- Dialogue with journal/publishing stakeholders to facilitate the level of good quality article metadata that would make efforts towards indicators for FAIRness more straightforward and standardisable - for example, through standardised approaches to the inclusion of, and metadata surrounding, DASs, as also recommended by the DAS pilot team.
- Dialogue with repositories to advocate for inclusion of a full range of metadata fields (and where appropriate, to mandate completion of these during submission) and a full range of options in drop-down menus (for example, regarding software licences).
- Engagement with researchers to provide further guidance and instruction on:
    - The importance of including a clear DAS in an article, regardless of whether the journal has a field for or otherwise requests this; using the Acknowledgements section to provide a DAS where no other option is available; advocating with journal editors and publishers where required for the inclusion of a DAS.
    - When referring to available data in a DAS, giving a PID where one is available rather than simply stating that the data is available, giving a generic link to a repository, or giving a repository page URL rather than the PID.
    - Checking that PIDs provided in DASs are correct and resolve to a landing page where the data or code can be found.
    - Avoiding ‘Data is available on request’ as a DAS where possible, or if this does need to be used, explaining why and specifying the process for making such a request, ideally in a metadata-only dataset record that is linked to by DOI from the DAS in the article.

General reflections on the project

Some useful learning points were identified by the institutional project team; these are detailed below.

Workload and sequencing challenges

- The project as a whole (comprising several pilots) was extremely ambitious, and was made more so by the fact that:
    - Work was undertaken on all pilots simultaneously, so that staff from a given institution may be participating in a number of pilots at the same time. In the context of this pilot, this placed limits on what we were able to achieve, in addition to the general limits created by the fact that this was voluntary work for both institutions and providers, to be fitted around existing work.

  - ○ Some of the pilot projects (e.g. the pilot on DASs) were logically prior to others (e.g. this pilot). For this reason, it was challenging for institutions to conduct the pilots simultaneously rather than sequentially, which would have enabled logical progression, allowed learning points to be taken forward into subsequent pilots, and may have avoided duplication of work and reduced workload challenges.
- Our recommendation would be for similar work in future to be sequenced as a series of work packages or to assume a more limited and realistic scope.

Differing stakeholder priorities

- A collaboration between HEIs and external, in some instances commercial, stakeholders with understandable investments in their own intellectual property, and interests in developing proprietary and/or organisation-branded tools, creates some obvious challenges.
- From institutions' perspectives, a desirable outcome would be to develop open, reproducible, repository-agnostic workflows that could be adapted by any institution or organisation into a tool to produce open research indicators. However, this is unlikely to align perfectly with priorities of the external partners as detailed above. Appropriate funding of future work may enable developers / software engineers to be hired explicitly to create open source tools and/or workflows.

Conclusions

While not resulting in finished sub-indicators for the measurement of FAIRness of data, this pilot has conducted valuable preliminary work which has enabled the identification of several areas in which further progress may be possible. In a more substantial pilot that was sequenced to allow prior work (e.g. in the DAS pilot) to be more fully built upon, it may also be possible to derive data and in time potentially indicators on more of the sub-indicators identified as of particular relevance, and it may also have been possible for institutions to further explore some of the additional data returned by providers which was less feasible to assess within the timescale of the project.

In the context of a pilot study in which all participants were volunteering their time alongside existing workloads, we feel a significant amount was achieved and some productive ground laid for ongoing work towards the development of sub-indicators for the FAIRness of data.

References

Australian Research Data Commons (n.d.). FAIR Data Self Assessment Tool. Online resource. https://ardc.edu.au/resource/fair-data-self-assessment-tool/

David, R., Mabile, L., Specht, A., Stryeck, S., Thomsen, M., Yahia, M., Jonquet, C., Dollé, L., Jacob, D., Bailo, D., Bravo, H., Gachet, S., Gunderman, H., Hollebecq, J.-E., Ioannidis, V., Le Bras, Y., Lerigoleur, E., Cambon-Thomsen, A., & SHARC Community. (2020). Templates for FAIRness evaluation criteria - RDA-SHARC ig (1.1) [Data set]. Zenodo. https://doi.org/10.5281/zenodo.3922069

Devaraju, A., Huber, R., Mokrane, M., Herterich, P., Cepinskas, L., de Vries, J., L'Hours, H., Davidson, J., & Angus White. (2022). FAIRsFAIR Data Object Assessment Metrics (0.5). Zenodo. https://doi.org/10.5281/zenodo.6461229

Devaraju, A., & Huber, R. (2021). An automated solution for measuring the progress toward FAIR research data. *Patterns (New York, N.Y.)*, *2*(11), 100370–100370. https://doi.org/10.1016/j.patter.2021.100370

FAIR Data Maturity Model Working Group. (2020). FAIR Data Maturity Model. Specification and Guidelines (1.0). Zenodo. https://doi.org/10.15497/rda00050

FORCE11 (n.d). Guiding principles for Findable, Accessible, Interoperable and Reusable Data Publishing, v. b1.0. Online resource. https://force11.org/info/guiding-principles-for-findable-accessible-interoperable-and-re-usable-data-publishing-version-b1-0/

Huber, R., Cepinskas, L., Davidson, J., Herterich, P., L'Hours, H., Mokrane, M., von Stein, I., & Verburg, M. (2021). D4.5 Report on FAIR Data Assessment Toolset and Badging Scheme (V1.0). Zenodo. https://doi.org/10.5281/zenodo.6656444

PLOS (2022) PLOS Open Science Indicators. Public Library of Science. Dataset. https://doi.org/10.6084/m9.figshare.21687686.v9

Annexes

Appendix 1: FAIR Data. Annex 1: Specification for Pilot 1: The FAIRness of Data

V1.1 – last updated 11/07/2024

Authored by Kerry Miller, Jenni Adams, Gillian Currie

1. Headline

Pilot 1 (‘the pilot’) will build on the work of Pilot 4 looking at the prevalence and quality of **Data Access/Availability Statements (DASs)**, as a necessary precursor to producing indicators of the FAIRness of the data underpinning the published articles incorporating those DASs. The pilot will also work in conjunction with Pilot 2 in establishing Scope, Definition Criteria, Methodology, and a Glossary which will reduce duplication of effort. As a result, much of this specification is drawn directly from the excellent work already done in Pilots 2 & 4, particularly their own specifications.

The **FAIR** data principles create a standard by which the transparency and reproducibility of research data may be judged. Each of the four distinct facets of FAIR can be considered independently, with some being easier to assess, and achieve, than others, but ultimately it is only when they are brought together that we get a full picture.

Findable – This is a measure of how easy it is to discover the existence and location of any given dataset. Without this, it is impossible to judge the other three facets.

Accessible – There are degrees of accessibility to be considered here. Whilst fully **Open Data** (See Pilot 2, and definition of Open Data in glossary) may be considered the gold standard, it is not always possible to achieve this for ethical, legal, or technical reasons. However, just because data is not open does not mean that it is not accessible: clear statements of where the data is stored, the reasons for restrictions, and route to applying for access to the data can still make this data accessible.

Interoperable – Interoperability requires both the use of open or widely used file formats and software and the creation of high quality metadata that conforms to accepted standards and vocabularies within the relevant discipline. Without these, it may be impossible for new users to open the dataset or to understand how it may be mapped onto other existing data.

Reusable – In order to be reusable, data must be assigned a licence that permits broad reuse of these both for validation of existing outputs and new research. As with accessibility, there are degrees of reusability, from an entirely open licence such as CC0 to highly restrictive licences which can prevent any reuse of data without entering into a contract with the data creator. Data with no licence cannot be considered reusable.

2. Definition criteria

As a baseline, the pilot will seek to establish the measurability of the following criteria. Of necessity, some of these criteria will build upon the work done in producing a reliable DAS (UKRN Pilot 4). The pilot leads and partners acknowledge that it may not be feasible to create workflows for measuring all of the below - in part, the aim of the pilot is to pinpoint which of these can be feasibly measured, and the input of solution providers will be crucial in determining this.

The pilot will focus on the following sub-identifiers of FAIR:

2.1 Findable

2.1.1 Establish whether the data underpinning the publication is assigned one or more Persistent Identifier/s which resolves to a landing page from which the data can be accessed, or, if the data can [no longer] be shared, to a metadata-only record.

2.1.2 – Establish whether the assigned PIDs are provided by an established, globally recognised provider.

2.1.3 – Establish whether the associated metadata includes the following core elements: creator, title, data identifier, publisher, publication date, summary and keywords.

2.2 Accessible - As the "Openness of a dataset" is the focus of Pilot 2, we will not examine that here.

2.2.1 – Establish whether metadata contain access level and access conditions of the data.

2.3 Interoperable

2.3.1 Identify whether the metadata of the object are available in a formal knowledge representation language.

2.3.2 - Identify whether namespaces of known semantic resources (excluding common namespaces, e.g., RDF, RDFS, XSD, OWL) are present in the metadata of an object .

2.4 Reuseable

2.4.1 Identify whether the metadata for the dataset includes provenance information about data creation or generation.

2.4.2 Identify whether the metadata for the dataset includes the reuse licence applied to each dataset cited: this might be **Creative Commons licences**, public domain, bespoke, or otherwise.

2.4.3. Identify whether the dataset uses **open or accessible file formats**.

a. If not open format(s), are data available in a file format recommended by the target research community.

3. Methodology

Institutions will work with industry partners ('partners' or 'providers') to establish the feasibility of the above sub-indicators and create algorithms/workflows to produce these measures where feasible.

As part of this process, institutions will provide industry partners with a test dataset of published articles (see below), the **DAS**s of which will provide a starting point for partners' identification of feasible sub-indicators and creation and testing of the algorithms/workflows necessary to produce these.

The pilot will:

3.1 Start with a small sample database provided by the lead or co-lead institutions based on recently published items 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 an acceptable DAS.

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

Categorise findings according to:

Findability

3.3.1 Identify the % of publications for which the underpinning data (where this exists and can be shared in full/part or as a metadata only record) is assigned one or more Persistent Identifiers (PID) which resolves to a landing page from which the data can be accessed.

Component questions:

- Is there a data identifier?
- Is the data identifier specified based on a commonly accepted persistent identifier scheme and syntax?
- Does it resolve to a landing page with metadata containing further information on how (if this is possible) to access the data object?

3.3.2 – Identify the % of publications for which the underpinning data (where this exists and can be shared in full/part or as a metadata only record) has a PID that is provided by an established, globally recognised provider of such.

Component questions:

- Where a PID for meta/data is included in the article, what is the provider of the PID?
- Is this an established, globally-recognised PID-provider?

3.3.3 – Identify the % of publications for which the underpinning data (where this exists and can be shared in full/part or as a metadata only record) is accompanied by associated metadata that includes the following core elements: creator, title, data identifier, publisher, publication date, summary and keywords.

Component questions:

- Does the metadata include all of the core elements listed above?

Accessibility

3.3.4 – Identify the % of publications for which the metadata of the underpinning data (where this exists and can be shared in full/part or as a metadata only record) contains the access level and access conditions of the data.

Component questions:

- Does the metadata include details of whether the data is open, technically open, embargoed, controlled, or closed?
- If the data is controlled, closed, or no longer available, is a reason provided?
    - Is this a legitimate reason (i.e. due to sensitivity of data, stipulations of data sharing agreement or contract, reasons of intellectual property)?
- If the data is **author-controlled** (available on request), does the metadata provide a clear point of contact for the process of requesting access?

Interoperability

3.3.5 - Identify the % of publications for which the metadata of the underpinning data (where this exists and can be shared in full/part or as a metadata only record) are represented using a formal knowledge representation language.

Component questions:

- Are parsable, structured data embedded in the landing page? OR:
- Are parsable, formal metadata (e.g., RDF, JSON-LD) accessible through content negotiation, typed links, or SPARQL endpoint?

3.3.6 - Identify the % of publications for which the metadata of the underpinning data (where this exists and can be shared in full/part or as a metadata only record) use known semantic resources.

Component questions:

- Are namespaces of known semantic resources (excluding common namespaces, e.g., RDF, RDFS, XSD, OWL) present in the metadata?

Reusability

3.3.7 - Identify the % of publications for which the metadata of the underpinning data (where this exists and can be shared in full/part or as a metadata only record) includes provenance information about data creation or generation.

Component questions:

- Are properties representing data creation, i.e identifier, creator, title, publisher, publication year and resource type present in the metadata? (see DataCite mandatory properties)

3.3.8 - Identify the % of publications for which the metadata of the underpinning data (where this exists and can be shared in full/part or as a metadata only record) includes a reuse licence for the data.

Component questions:

- Does the data have a standard licence e.g. a Creative Commons, Open Data Commons, Open Database License licence, or a bespoke licence or rights statement? OR:
- If software, does it have a MIT or GPL licence?
- If the data is controlled, closed, or no longer available, is a reason provided?
  - Is this a legitimate reason (i.e. due to sensitivity of data, stipulations of data sharing agreement or contract, reasons of intellectual property)?

3.3.9 Identify the % of publications for which the underpinning data (where this exists and can be shared in full/part or as a metadata only record) uses open or accessible file formats.

Component questions:

- Are data available in a format categorised as an open format? (e.g. see Library of Congress Recommended Formats Statement)
- Are data available in a recognised scientific file format (e.g., Library of Congress dataset formats, Wolfram Alpha supported file formats), OR
- Are data available in a long-term file format as defined in ISO/TR 2229935?

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 – see Annex 2 below:

4.4.1 LocalID (local CRIS repository ID number)
4.4.2 DOI (publisher DOI)
4.4.3 Link (publication link)
4.4.4 PageLink (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 School/Faculty/Department (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 1, and enable cross-comparison with other pilots following the same basic template, i.e. Pilots 2 and 4. See Overview Report: Annex 1.

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 funder, research output type 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.

4.8 For the purposes of this pilot, 'data' will include all materials gathered and/or created to inform research findings, including code.

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

Glossary

- **Creative Commons licence**: a licence which allows for re-use and re-distribution of the underlying file(s). There are a range of these licences according to the level of restriction applied (https://creativecommons.org/share-your-work/cclicenses/)
- **Data Access/availability statements**: 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.
- **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 licence.
- 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 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.
- **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 licence to make the publisher's version freely available), or Diamond (similar to Gold, but with no payment involved for the licence).

Appendix 1: FAIR Data. Annex 2: full description of the test dataset

This Annex provides the scope of the dataset requirements requested from the industry partners.

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

   a. LocalID (local CRIS repository ID number)
   b. DOI (publisher DOI)
   c. Link (publication link)
   d. PageLink (link to repository splash page/similar)
   e. Type (item type, i.e. journal article)
   f. Year (year of publication i.e. 2023)
   g. School/Faculty/Department (the managing organisation unit record in CRIS)

4. 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 1, and enable cross-comparison with other pilots following the same basic template, i.e. Pilots 2 and 4. See Overview Report Annex 1.
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 funder, research output type 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.
6. These parameters 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.
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.
8. For the purposes of this pilot, ‘data’ will include all materials gathered and/or created to inform research findings, including code.

Appendix 1: FAIR Data. Annex 3: Mapping document

UKRN Pilot 1
mapping - notes.xlsx

Appendix 1: FAIRness of data. Annex 4 - CWTS Review

Kathleen Gregory

Overall summary

This pilot project aimed to explore developing and testing possible sub/indicators for the FAIRness of a sample dataset of open access articles with data availability statements (DAS). The project team gave a dataset to four different data providers, along with a list of specifications for FAIR sub-indicators. Each provider tested their own method to identify and assess the FAIRness of datasets in the DAS, according to the provided specification, and returned their results to the project team. The project team then conducted a manual check of the accuracy of the providers' approaches.

The methodology used by the team was strong, as is the analysis of the limitations of the data returned by providers. While the team did not create a finalized set of indicators, they successfully uncovered and discussed limitations in the data and various approaches, which can be taken into account in the future development of FAIR indicators. They were also able (at a high level) to provide some insight into the different approaches of data providers, while balancing a tension in working with partners with proprietary interests. Overall, this is solid and impressive work (which was done in a short time) and which has findings that can inform future efforts for data preparation, DAS, and ultimate indicator development.

I am writing this review with the knowledge that the project team's report can't be changed. Even though that is the case, I provide further comments and questions on the various aspects of the report and invite the authors to respond in a reply or to think about taking some of these ideas forward in their future work.

Framing and positioning of the report (Background and aims)

The team provides a strong rationale to use DAS as the starting point in their analysis, as this enabled them to assess findability and look at data sharing practices 'in the wild' rather than by starting with repositories. I wonder if the team considered looking at the reference lists of papers or footnotes of the papers in order to capture other data citations or mentions of shared data? While I understand that DAS were used for practical limitations here, if data citation (e.g. in reference lists) continues to be encouraged by various stakeholders, does this somehow stand at odds with the focus on DAS used in this (and other) pilots? It would be interesting to hear the team's perspective on this.

The background section is well situated in previous efforts to evaluate FAIRness, such as the FAIRsFAIR project. I would be curious to learn more in this section about why these existing tools did not map "exactly onto the needs of the current project." Just a sentence more here would make the case stronger. A small point: rather than linking to the FAIR data principles on the GOFAIR webpage, a formal citation to the paper originally proposing the principles (Wilkinson et al., 2016) would be more appropriate.

Rigour of approach (Methods and data)

I appreciate that the methodology is situated in the SCOPE Framework - this is great. I wonder if the team could come back and reflect more specifically on the aims they identify on page 3. There is some implicit reflection on these throughout the report, but explicitly doing so would be an interesting case to see the use of SCOPE throughout the process.

The starting point for drafting the sub-indicators was a literature review. I am missing details about how the literature search was conducted: in which databases? At which time? With which criteria? (In an ideal world, it would also be interesting to see a list of the reviewed sources). The sub-indicators themselves are sensible, but raise a few questions. For indicator

2.1.2, did the team develop a list of "established, globally recognized" PID providers? This would be helpful for others in the future and to contextualize any indicator. I also wonder how the core elements listed in 2.1.3 were developed. (In my experience – and as the results suggest later – keywords or information describing the subject of datasets are not always present).

The methodological steps are really clearly laid out; this is very nicely done. The only thing I would want more detail on is a description of the manual checking process. There is a bit more later on, but I think it is important to know already about sampling for the manual checks, for example. There is a description of the dataset shared with providers in Annex 2, specifically for one institution's dataset, but I am missing a high level overview the dataset, e.g. any information about disciplines associated with the articles or data, which are important in considering data sharing and referencing practices more generally. Such differences could impact how data are described in DAS. The report also mentions dependencies on pilot 4; could this be spelled out a bit more? Were the same data being used as in pilot 4, or just the same method for developing the sample? It is a bit unclear.

Relevance of findings and results

Table 2 suggests that the presence of a PID is something that is easily measured. Other sub-indicators, such as those related to interoperability are barely there. This matches the challenge in operationalizing the "I" in FAIR more broadly. The section analysing "the data and how it was returned", by indicator and provider, is a strong contribution of this report, as is the analysis of what would need to be done to move to the indicator stage.

There were significant limitations and challenges which the authors encountered with the data that were returned, perhaps most significantly the last-minute error/bug in the OpenAIRE approach. I wonder if this data should even be presented in the body of the report itself, given that the analysis is likely not correct. Maybe it could have been moved to an appendix and not presented alongside the other approaches? All in all, though, the team does a commendable job at pointing out and labelling limitations in the data retrieved and methodology (e.g. the fact that Elsevier limited their approach to Elseiver articles).

I could use more details about how the accuracy scores were calculated in this section, though. These scores are doing a lot of comparative work (even with the caveats provided); transparency about the calculation would provide more context.

Points for reflection (Discussion)

The biggest strength of this report is calling attention to limitations, which is continued in the Discussion. They also highlight the diverse relationships which are present in how authors refer to data in DAS, e.g. as a source or an output. Again, this resonates with other work which demonstrates variability in how researchers reference data (in text, in data citations, in data mentions or footnotes) and similar relationships to data sources and articles. This raises a further point which could be investigated in future work (more empirically perhaps, but that would also inform indicator development): to compare data references in DAS to references in other parts of an article.

Overall, this report highlights that there is still much to be done at the level of the data/DAS before indicators are developed.

Appendix 1: FAIRness of data. Annex 5 – Team response to CWTS Review

Response to Review from CWTS of the FAIRness of Data Indicator Pilot report.

This was a short-term pilot project based on the work of volunteers from many different institutions and all teams gave as much time and effort as they could but coordinating a project on this scale was challenging therefore pragmatic decisions had to be made based on what could be achieved in the time available.

Requested changes to the report:

- Some of the footnotes (e.g. 14) look to have been added by providers - it would be good to make this explicit, otherwise the use of 'we' in these notes could be confusing, as in the rest of the article 'we' refers to the authors. This could be achieved by prefacing provider-added notes with 'PLOS clarify:' or similar.

Requested changes to the CWTS review:

- To correct the opening statement 'This pilot project aimed to explore developing and testing possible sub/indicators for the FAIRness of a sample dataset of open access articles with data availability statements (DAS)', to clarify that the dataset included articles both with and without DASs.

Response to reviewer

Reviewer Comment 1:

Framing and positioning of the report (Background and aims)

The team provides a strong rationale to use DAS as the starting point in their analysis, as this enabled them to assess findability and look at data sharing practices ‘in the wild’ rather than by starting with repositories. I wonder if the team considered looking at the reference lists of papers or footnotes of the papers in order to capture other data citations or mentions of shared data? While I understand that DAS were used for practical limitations here, if data citation (e.g. in reference lists) continues to be encouraged by various stakeholders, does this somehow stand at odds with the focus on DAS used in this (and other) pilots? It would be interesting to hear the team’s perspective on this.

Our response:

In this pilot we started from DAS for pragmatic and content reasons: within the project’s limited time frame, systematically mining citations was not feasible given the difficulties of identifying citations within reference lists, as opposed to publications and other outputs; and DAS typically contain FAIR-relevant details that references rarely contain (e.g. access level/conditions, reasons for restrictions and sometimes licence). We don’t view our approach as at odds with community efforts to strengthen formal citations- rather, it is complementary. We agree that it would be interesting to explore structured data citations in reference lists (and mentions in footnotes/supplementary materials) and would propose incorporating this assessment alongside DAS as part of a robust indicator pipeline. In addition, when looking for DAS type text in footnotes, references, citation lists etc, it proved very difficult for providers to identify these accurately. Pilot 2 reports that the tools often confuse cited datasets with datasets which are published alongside the article, and this problem is compounded by multiple datasets and how these are weighted (it seems to be first past the post dataset). The tools aren't sophisticated enough.

Reviewer Comment 2:

The background section is well situated in previous efforts to evaluate FAIRness, such as the FAIRsFAIR project. I would be curious to learn more in this section about why these existing tools did not map "exactly onto the needs of the current project." Just a sentence more here would make the case stronger. A small point: rather than linking to the FAIR data principles on the GOFAIR webpage, a formal citation to the paper originally proposing the principles (Wilkinson et al., 2016) would be more appropriate.

Our response:

As far as we are aware, existing tools focus on dataset-level assessment within repositories and assume a known PID and machine-actionable metadata therefore using these existing tools would only help assess data within a repository. Our pilot begins from articles and DASs to evaluate findability and reporting in practice.

We agree that this citation would be more appropriate: Wilkinson, M., Dumontier, M., Aalbersberg, I. et al. The FAIR Guiding Principles for scientific data management and stewardship. Sci Data 3, 160018 (2016). https://doi.org/10.1038/sdata.2016.18

Reviewer Comment 3:

Rigour of approach (Methods and data)

I appreciate that the methodology is situated in the SCOPE Framework - this is great. I wonder if the team could come back and reflect more specifically on the aims they identify on page 3. There is some implicit reflection on these throughout the report, but explicitly doing so would be an interesting case to see the use of SCOPE throughout the process.

Our response:

We agree that it may have been valuable and insightful to return to the SCOPE framework in the later stages of the report.

Reviewer Comment 4:

The starting point for drafting the sub-indicators was a literature review. I am missing details about how the literature search was conducted: in which databases? At which time? With which criteria? (In an ideal world, it would also be interesting to see a list of the reviewed sources). The sub-indicators themselves are sensible, but raise a few questions. For indicator 2.1.2, did the team develop a list of "established, globally recognized" PID providers? This would be helpful for others in the future and to contextualize any indicator. I also wonder how the core elements listed in 2.1.3 were developed. (In my experience – and as the results suggest later – keywords or information describing the subject of datasets are not always present).

Our response:

In March 2024, relevant literature was identified in a non-systematic way; members of the project group identified literature that contained potentially useful information based on their own knowledge and research, and this list of tools (https://fairassist.org/). In this way, a list of existing literature on FAIR indicators was compiled. Reading from the list was assigned to members of the project group to provide a brief summary of the tool or publication and note any indicators that were relevant to our pilot (Findable, Accessible, Interoperable, Reusable). The reviewed resources are listed here.

A list of recognised and well used PIDs was chosen based on participant discussions during the agreement of the pilot specification; it is not designed to be an exhaustive list but a test of how well the providers could identify DAS and PIDs from these providers within published articles. The list used was:

- Digital Object Identifier (Datacite)
- Handle

- Protein Data Bank Identifier (PDB)
- PubMed Central ID
- PubMed ID
- uniprot
- Ena
- Cambridge Crystallographic DataBase (CCDB)

The list of core elements in 2.1.3 is derived from the Datacite Mandatory Properties (https://datacite-metadata-schema.readthedocs.io/en/4.5/properties/overview/#table-1-datacite-mandatory-properties). The expectation was not that all of these should appear in the DAS, but that they would be found in the metadata available on the Landing Page that each PID would direct readers to.

Reviewer Comment 5:

The methodological steps are really clearly laid out; this is very nicely done. The only thing I would want more detail on is a description of the manual checking process. There is a bit more later on, but I think it is important to know already about sampling for the manual checks, for example. There is a description of the dataset shared with providers in Annex 2, specifically for one institution's dataset, but I am missing a high level overview the dataset, e.g. any information about disciplines associated with the articles or data, which are important in considering data sharing and referencing practices more generally. Such differences could impact how data are described in DAS. The report also mentions dependencies on pilot 4; could this be spelled out a bit more? Were the same data being used as in pilot 4, or just the same method for developing the sample? It is a bit unclear.

Our response:

The manual checking process is detailed in this document:

https://docs.google.com/document/d/12rUkkuY5lWde-Nlt4QEaByNKObjhsAVC/edit. This was provided to all pilots by UKRN in the form of a template accompanied by guidance.

As the project had already overrun by this point and there was little capacity within the institutions for the manual checking this was done according to the recorded process as far as possible. However, each checker was responsible for selecting the records they would check at random.

Although it would have been interesting to look at different disciplines, we did not capture the discipline of individual papers as this presents its own challenges which were outside the project scope. We did, however, collect School or Dept from the institutions' CRIS systems where possible, but differences between how these are organised in each institution meant that it was not possible to use them as a reliable proxy for discipline.

All data sets are available in project "UKRN Pilots 1,2, and 4" on Figshare but as they are in a Private project area access will need to be requested. The same institutional datasets were used as the input for pilots 1, 2, and 4.

Pilot 4 identified whether or not DAS existed within those records and assessed the quality of the DAS within those records (e.g. did it include a PID). Pilot 1 then used these identified DASs to assess the FAIRness of the datasets.

Reviewer Comment 6:

There were significant limitations and challenges which the authors encountered with the data that were returned, perhaps most significantly the last-minute error/bug in the OpenAIRE approach. I wonder if this data should even be presented in the body of the report itself, given that the analysis is likely not correct. Maybe it could have been moved to an appendix and not presented alongside the other approaches? All in all, though, the team does a commendable job at pointing out and labelling limitations in the data retrieved and methodology (e.g. the fact that Elsevier limited their approach to Elseiver articles).

Our response: We were only notified of the OpenAire error/bug after completing the analysis and report, and after the project timescale had ended. The approach to reporting the error/bug was negotiated between UKRN and OpenAire.

Reviewer Comment 7:

I could use more details about how the accuracy scores were calculated in this section, though. These scores are doing a lot of comparative work (even with the caveats provided); transparency about the calculation would provide more context.

Our response: Accuracy scores were calculated using formulae provided by UKRN as part of the standardised manual checking process (see link to this document above). We suggest that clarification of this process is provided by UKRN at the level of the master report for the pilots, as this is common to all pilots that used the manual checking process.

Reviewer Comment 8:

Points for reflection (Discussion)

The biggest strength of this report is calling attention to limitations, which is continued in the Discussion. They also highlight the diverse relationships which are present in how authors refer to data in DAS, e.g. as a source or an output. Again, this resonates with other work which demonstrates variability in how researchers reference data (in text, in data citations, in data mentions or footnotes) and similar relationships to data sources and articles. This raises a further point which could be investigated in future work (more empirically perhaps, but that would also inform indicator development): to compare data references in DAS to references in other parts of an article.

Our response: Yes, we agree. Future work should look for data citations / references throughout a publication, at least until a standardised DAS format and location in publications is established across publishers.