UKRN Open Research Indicators: Final Report

UKRN Working Paper: 09 / 12/02/2026 DOI: **10.5281/zenodo.18619284**

UKRN ROR ID: https://ror.org/049czkm07

Contributors are listed for the pilots to which they contributed, in each of the six appendices to this report. A full contributor list is provided below.

Final report author: Neil Jacobs

Funding acknowledgement: This work formed a part of the UKRN Open Research Programme, which is supported by a Research England Development fund grant. However, the work described in this report and its appendices went significantly beyond that originally proposed to Research England and has been done on a voluntary, unfunded basis by those involved.

Conflict of interest statement: All those involved have identified the organisations to which they are affiliated.

Full contributors list, Z-A:

Laurian Williamson, University of Leicester
Christopher Warren, University of Bristol
Tim Vines, DataSeer
Etienne Vignola-Gagné, Elsevier
Christopher Tibbs, University of Exeter
Angus Taggart, University of Sheffield
Nick Sheppard, University of Leeds
Etienne Roesch, University of Reading
Leonidas Pispiringas, OpenAIRE AMKE
Radoslaw Pajor, University of Leicester
Daryl O'Connor, University of Leeds
Jonathon T. Newton, King's College London
Serena Mitchell, King's College London
Kerry Miller, University of Edinburgh
Kirsty Merrett, University of Bristol
Valerie McCutcheon, University of Glasgow
Mark Kelson, University of Exeter
Neil Jacobs, UKRN / University of Bristol
Iain Hrynaszkiewicz, PLOS
Alice Howarth, University of Liverpool
Phillip Hall, Digital Science
Johanna Groothuizen, King's College London
Bill Greenhalf, University of Liverpool
Evangeline Gowie, University of Reading
Jade Godsall, University of Bristol
Lavinia Gambelli, University of Bristol
Mick Eadie, University of Glasgow
Mary Donaldson, University of Glasgow
Harry Dimitropoulos, OpenAIRE AMKE
Anita de Waard, Elsevier
Sally Dalton, University of Leeds
Gillian Currie, University of Edinburgh
Ann Campbell, Digital Science
Frederick Breese, University of Manchester
Amanda Boll, Newcastle University
Nicola Barnett, University of Leeds
Alastair Arthur, University of Glasgow
Euan Adie, Overton
Jenni Adams, University of Sheffield

Executive summary

This report presents the findings of six pilot projects undertaken by the UK Reproducibility Network (UKRN) to explore the feasibility of developing robust, automated indicators for monitoring open research practices across UK universities. The pilots form part of the UKRN Open Research Programme (ORP), which aims to strengthen institutional support for openness through training, incentives, and the development of shared evaluation tools. Conducted voluntarily by academic and professional staff, alongside international data service providers, the work is exploratory and intended to inform future infrastructure and policy development rather than to establish definitive sector benchmarks.

The pilots focused on six aspects of open research: FAIR data; open data; data availability statements (DAS); downstream effects of sharing data and other research outputs; pre-registration; and the use of the CRediT taxonomy. Across all areas, teams examined the suitability of institutional and third party datasets, assessed automated extraction methods, and compared provider outputs with human validation.

The work demonstrated clear potential for generating meaningful indicators in some areas, particularly pre-registration and selected components of FAIRness (e.g., detection of persistent identifiers, metadata completeness). Automated detection of CRediT statements is also feasible with further refinement. However, most pilots concluded that current tools, metadata standards, and sector workflows are not yet mature enough to support reliable, automated institutional monitoring. Manual checking therefore remains necessary both to ensure accuracy and to validate emerging automated approaches.

A consistent theme across all pilots was the lack of standardisation—of terminology, metadata structures, DAS formats, article type definitions, and repository practices. These inconsistencies hinder automation, reduce reliability, and complicate cross provider comparisons. The report highlights an urgent need for coordinated development of shared definitions, taxonomies, and data standards, ideally led by a trusted, international body such as the Open Science Monitoring Initiative (OSMI).

The pilots also emphasised the importance of improving system interoperability and metadata quality across the research ecosystem. Enhancing repository capabilities, embedding structured DAS into publication workflows, improving the use of persistent identifiers, and aligning journal submission systems with open research requirements would all strengthen monitoring and streamline researcher workflows.
Across the pilots, there was strong commitment to transparency, sector leadership, and the use of open data wherever possible. However, the voluntary nature of the work, varied timelines, and uneven adoption of project management approaches created challenges. Future initiatives of this scale should be fully resourced and centrally coordinated to maximise efficiency, support sequencing between interdependent work packages, and minimise the operational burden on institutions and providers.

Overall, the pilots demonstrate that automated monitoring of open research is achievable but currently limited. Progress will depend on collective action to develop shared standards, improve system infrastructure, and embed good practice in the creation, use, and evaluation of indicators. Continued manual validation, transparent methods, and responsible use of metrics will be essential as the sector moves towards more mature, coordinated monitoring of open research practices.

Introductory remarks

The work reported here and in the related and embedded material was undertaken on a voluntary basis by teams of experts with day jobs. The teams included both professional and academic staff in UK universities and staff in international services and companies. Their time was not bought out or the work funded. The benefit of this is that the work has been embedded in expert, professional practice. The drawback is that, because the work proved significantly more that anticipated, it often needed to fit around that practice, imposing both undue strain on those involved and delays to completion of the projects. Responsibility for this lies with UKRN management, which has learned lessons for future projects.

Readers should note that the work done was exploratory; it aimed to explore the potential of different methods of monitoring open research practices, foster collaboration between different stakeholders involved in that monitoring, and help them improve those methods. As the work proceeded it became clear that only a minority of the outputs would be firm recommendations on best practice, the rest being recommendations on further work needed.

One consequence of this is that things have moved on. This is true in terms of the expertise, professional practice and interests of those in institutions, the development of methods, data and tools by services and companies, and the capabilities of technologies in the wider world. The work reported here has often contributed to that progress for those involved. Therefore, all of the quantitative figures reported here are out of date. There are reported in the interest of transparency, and to contribute to the contextualisation of any future indicators in use, but they should not be taken as indicative of the current situation or of future potential of any particular system or dataset.

Background and aims

Overview

The UKRN leads a multi-year programme – the Open Research Programme (ORP) – to promote the uptake of open research practices, including training, revisions to institutional recognition and reward policies and procedures, and sharing practice among institutions. It monitors and evaluates progress and aims to provide dashboards and reporting tools to institutions to enable them to do the same. As a first step, it initiated a set of pilots involving both institutions and solution providers to explore and refine the indicators that might be used.

There is a long, international history of efforts to monitor aspects of open research for a range of purposes, including to check compliance with policies, to inform planning for interventions to promote openness, and to help to evaluate those interventions. There is a short summary and further background in the UKRN Working Paper #2, including an outline of the topics being covered by the pilots and the rationale for those. This working paper builds on, rather than repeating, that background.

The overall aim of the pilots was to establish good practice in institutional monitoring open research, for example in the design and use of dashboards and reporting tools.

For the purposes of this work, indicators are to enable institutions to better support open research. They are not, for example, to assess research or researchers or to rank institutions against each other. This stems from the evaluation requirements of the ORP, which were to enable the programme and partners to monitor changes in open research practices. The decision also arises from a concern to avoid – as far as possible[1] – some of the recognised challenges in using indicators for research assessment. As such, indicators are at an aggregate and anonymised level, are specific to the mission and aims of the institution and where possible draw partly on institutional data as well as third party data.

We understand an "indicator description" to be a clear articulation of what it means to measure[2] something well. That indicator description would include:

1. a reasonably precise definition of the phenomenon being measured that provides criteria for deciding what is included and what is not in almost all likely cases;
2. data needed to produce the indicator in terms of its quality (relevance, scope, extent, granularity) and its features (validity, reliability, transparency, etc), but not specification of a particular dataset, although examples might be given;
3. the processing steps needed to turn data from potentially several sources into an indicator, in terms of processing functions (inputs/outputs) and processing attributes (transparency, reproducibility, etc); this doesn't specify particular tools although examples might be given;
4. The features of the indicator itself (e.g., scope of application, limitations, sensitivity, susceptibility to gaming).

The intention was that, once these descriptions were set out, then each partner in the ORP could use the data and services available to it to underpin dashboards and reporting tools. Where many ORP partner institutions could use similar data and services, then these outputs might be developed collectively.

[1] We recognise that aggregate and anonymous indicators can be used in ways that impact on research assessment.

[2] 'Measure' here and elsewhere in this report does not imply exclusive use of either qualitative or quantitative approaches.

Openness and specificity

The ORP work to develop indicators aims to balance two different drivers. The first is openness and transparency in the creation and use of the indicators, in keeping with UKRN's mission and commitments. The UKRN and some ORP partner institutions have signed the Barcelona Declaration on Open Research Information, which commits us to using open data as a default to underpin indicators. The second is specificity, that is, the indicators should count things that the ORP partner institutions value rather than, for example, things that are easy to count or things that are easily countable using open data. The two drivers can therefore pull in different directions, and the approach taken reflects this.

Methods

Establishing the pilot teams

As described in the previous working paper, four priorities were identified as dimensions of open research that UK institutions wanted to monitor. They were:

1. open and/or FAIR data (their completeness, reusability, reproducibility, etc)
2. data accessibility statements (their existence, completeness, accuracy, etc.)
3. pre-registration (its existence, completeness, different kinds, etc.)
4. the use of CRediT (a binary yes/no, but also perhaps the breadth of contributions thereby recognised – i.e., how 'well' CRediT is being used)

Some 14 institutions and 13 'solution providers' (organisations specialising in data processing that can underpin indicators) initially expressed interest in being part of a set of pilot projects to explore how best to monitor these aspects of open research.

There followed two rounds of iteration, in which the institutions outlined their priorities in monitoring these aspects of open research (using the INFORMS SCOPE framework[3]), with some limited but increasing degree of specificity, and providers outlined how they might work with the institutions to address these priorities using their data and systems. Institutions were then able to select which providers they wished to work with, on which pilots, and those providers were offered the chance to work on those pilots. This process took several months and resulted in the following pilot projects.

Pilot	**Participating institutions (leads)**	**Participating providers**
FAIR data	Sheffield, Edinburgh, Surrey, Reading, Glasgow, Exeter	Digital Science, Elsevier, OpenAIRE, PLOS/DataSeer
Open data	Bristol, Surrey, Sheffield, Reading, Glasgow, Exeter	Digital Science, Elsevier, OpenAIRE, PLOS/DataSeer
Effects of sharing data	Reading, Liverpool, Glasgow	Digital Science, Elsevier, OpenAIRE, PLOS/DataSeer, and additional data sources used
Data availability	Glasgow, Leicester, Surrey, Sheffield, Reading, Newcastle,	Digital Science, Elsevier, OpenAIRE, PLOS/DataSeer

[3] The use of the SCOPE Framework was not consistent across the pilots. Most used it to enable participating institutions to reflect on their reasons for wanting to monitor open research and the uses to which any indicators might be put, which then informed the providers. See Annex 2.

statements (DAS)	Manchester, Leeds, Edinburgh, Bristol, Exeter	
Pre-registration	Exeter, Manchester, Leeds, Reading	Digital Science, OpenAIRE, PLOS/DataSeer, Center for Open Science
CRediT	King's College London, Leeds, Sussex, Newcastle, Surrey, Liverpool, Glasgow, Bristol, Exeter, Reading	Digital Science, Elsevier, OpenAIRE, PLOS/DataSeer

While the set of participating institutions remained close to that which originally expressed interest, the set of participating providers was much more limited. We published a blog post that reflected on this outcome and made commitments to address some of its implications.

It rapidly became clear that the pilots looking at aspects of data sharing (DAS, FAIR and Open data) shared many concerns, and those pilots discussed ways to coordinate their work (see Overview Report, Annex 1).

Establishing ways of working

Since each of the pilot teams would be comprised of several institutions and providers, not all of whom might normally collaborate with each other, it was important to agree some partnership principles before the pilots started. These were adapted from some principles developed in the Netherlands between Dutch funders, universities and Elsevier. All partners (institutions and providers) were expected to act according to the following principles:

1. Sector-led: The work will be driven by the priorities articulated by the UK research sector, based on the UKRN call for priorities. This means that potential indicators and related solutions will be preferred that best meet sector priorities, before any consideration of which systems, solutions, or providers might be relevant to them.
2. Interoperability and solution neutrality: All partners are free to use their own or third-party products and services. All partners will make all reasonable efforts to ensure optimal interoperability for their data and services.
3. Transparency, inclusion, and collaboration: The work aims to make science and research more transparent, efficient, inclusive, openly and freely accessible, and collaborative. The partners agree that the broadest possible audiences should have the opportunity to participate, to make use of and to contribute to the scientific process.
4. Access to research data and metadata: Preference will be given to solutions that both exploit and contribute to FAIR and open sector-wide / system-wide data about research. At the same time, preference will also be given to solutions that both exploit and contribute to institutional data about research, to render solutions that are – as far as possible – context-sensitive and unsuitable for rankings[4].
5. Data portability: All partners agree that any data they choose to share as part of this work can be freely transferred by others to their own or to a third-party host environment, in line with the terms of any sharing agreement.

All participation was, of course, voluntary. Participants (institutions or providers) could withdraw at any time but, if intending to do so, were requested to discuss this first with UKRN and with other participants in the relevant pilots. All parties retained exclusive rights to their background IP. Any new IP generated during the pilots was shared openly using permissive

[4] This was to avoid, as far as possible, the potential for one kind of misuse of indicators.

licences. All parties agreed to respect appropriate levels of confidentiality. UKRN drafted a template written agreement that pilots could use to frame participation, if necessary.

While there was common purpose among the pilots, each was led by a different set of people and had diverse teams, contexts and challenges, and so UKRN initially encouraged each to adopt a project management approach appropriate to their circumstances. It became clear in the second half of the pilots that these had led to some confusion and difficulties in communication, and so a more harmonised 'agile' approach was introduced late in the process using kanban boards, with only partial adoption.

Establishing requirements

Each institution undertook an exercise whereby it used the SCOPE framework to develop a moderately detailed outline of the aspects of open research that they value and the context in which indicators of those aspects would be used. The intention was that these would be a reference source later in the pilots, against which to compare the indicator descriptions as part of their evaluation. The SCOPE documents are provided in Overview Report Annex 2.

Identifying the data

As noted above, the aim was to combine both institutional and third-party data to underpin indicators. The reasons for this were that such indicators would then likely be specific to institutional mission and context, difficult to use to rank institutions, and richer than indicators based on either institutional or third-party data alone.

Each pilot compiled a seed dataset from a sample of the institutional systems available to it, comprising relevant lists of research works. Several of the pilots collaborated to produce a shared dataset. These datasets were held on a temporary Figshare instance made available for the purpose by Digital Science. Providers and, in some cases, institutional staff then used these seed datasets to query third-party / provider sources, to explore how best to monitor their particular aspect of open research. More detail on the methods adopted by each pilot are outlined in the Appendices corresponding to each pilot.

Transparency

As the pilots were writing up it became clear that the participation principles noted above were insufficiently clear to guide decisions on which outputs would be shared from the pilots. Following discussion among the pilots, a more specific set of principles was agreed, shown in Overview Report Annex 3. Some providers produced data reports covering several of the pilots, and these are included in Overview Report Annex 4.

Validation

Most of the pilots undertook a validation step comparing the outputs from the automated processes with the results of a human check following the same specification and definitions. The guide for this check was provided by UKRN (see Overview Report Annex 5), though this was not followed by all pilots because of the effects of timelines and contexts.

Evaluation and quality assurance

Two evaluation exercises were added late into the pilots' plans. These exercises were both supported by the CWTS team at Leiden University, world leaders in the development and use of research indicators. They were planned to be a formative evaluation and a summative evaluation. However, the different schedules of each pilot meant that, in terms of pilot progress, the formative exercise happened rather late for some pilots and earlier for others. The summative evaluation took the form of an expert review and commentary on this report, which can be found in Overview Report Annex 6.

Findings

The detailed findings from each of the pilots are given in the Appendices and are only summarised here.

FAIR data

The pilot project aimed to develop indicators of constituent aspects of FAIRness (Findable, Accessible, Interoperable, Reusable) for datasets underlying published research. We termed these 'sub-indicators'. However, due to the preliminary nature of the work, it did not reach the stage of producing these sub-indicators, and thus could not fulfil the objectives outlined in the INORMS-SCOPE review. In brief, this was because of the complexities in optimising both recall and precision where, for example, some aspects of FAIRness were not well defined, it was unclear what should be the correct denominator, and/or reproducible methods were not available or used. Despite this, the project showed promise, suggesting that with further development, it may be possible to create sub-indicators for several key dataset attributes: the presence of a persistent identifier (PID), complete metadata, provenance information, clearly stated access conditions, and licensing details.

The work conducted by solution providers demonstrated potential in developing these sub-indicators, but several challenges remain. These include improving the accuracy of the methodologies, addressing provider-specific limitations, and transforming raw data into usable indicators.

To achieve this, most providers would need to incorporate a step to evaluate whether specific criteria (e.g. presence of a persistent identifier, PID) are met. Additionally, most providers would need to develop a method to calculate an appropriate denominator—excluding articles without data or those where data sharing is ethically or legally restricted. Finally, a transparent, reproducible method for creating each sub-indicator would need to be developed and openly shared.

These steps are essential to go beyond the limitations of the pilot and to realise the original goals. The findings suggest that, with more time and resources, it is feasible to build some robust, meaningful indicators of specific aspects of FAIRness that are valid and reliable across different kinds of research, and that can support improved research data practices.

Open data

The pilot study aimed to assess the feasibility of creating a reliable, ethical, and transparent indicator for data openness based on Data Availability Statements (DAS). However, manual checks revealed that current tools and methodologies fall short of this goal. Key issues included inconsistent repository and PID recognition, reliance on free-text DAS rather than actual datasets, and discipline-specific disparities in data citation practices.

Most solution providers could not meet the pilot's full specifications and focused only on aspects compatible with their existing tools. Repository identification was incomplete, and many tools failed to extract PIDs unless they were explicitly formatted. Categorisation of openness was based solely on DAS content, which is often ambiguous and inconsistent. For example, terms like "on request" were interpreted variably, leading to misclassification. Tools also struggled with multiple datasets cited in a single DAS, assigning only one openness category per statement.

Terminology inconsistencies further hindered progress. Only one provider used the pilot's specified openness categories, while others applied simplified or proprietary terms. Data cleaning and integrity were also problematic, with missing or corrupt metadata affecting accuracy. Commercial constraints limited transparency, as some providers withheld methodological details due to proprietary concerns.

A small-scale analysis at the University of Bristol showed that standardised DAS language significantly improved machine readability, highlighting the need for controlled vocabulary. Recommendations include developing a DAS taxonomy, standardising openness terminology, and embedding DAS in publication metadata.

Operational challenges also impacted the pilot. Running all eight pilots concurrently limited cross-pilot learning and delayed data delivery. Compressed timelines and lack of real-time collaboration further constrained analysis. Overall, the pilot underscores the need for collective action among funders, institutions, and publishers to standardise practices and improve the reliability of openness indicators.

Data Availability Statements (DAS)

This pilot study explored the feasibility of assessing DAS across research publications, focusing on their discoverability, categorisation, and potential for standardisation. This pilot ran in parallel with those for the FAIRness and Openness of data. While the three pilots did inform each other and use a common seed dataset from institutions, it would have made sense for them to run sequentially so that identified DAS could have been more easily used as the basis for assessing data FAIRness and openness.

The pilot found that while solution providers were generally able to identify more DAS than local manual checks, results varied significantly depending on how and where the DAS appeared in the publication—whether as formal subsections, in supplementary materials, or footnotes. This inconsistency made validation and comparison challenging.

Soluton providers often analysed the final published versions of articles, which may include DAS not present at the acceptance stage—when institutions typically conduct their checks—leading to discrepancies. Categorisation of DAS also varied: some solution providers treated statements like “no data available” as valid DAS, while others marked them as “not applicable.” Even with a consistent dataset provided by institutions, not all solution providers could access or analyse the same documents, complicating cross-provider comparisons.

The pilot highlighted the early-stage maturity of this indicator and the need to evolve from simply counting DAS to evaluating their quality and alignment with good practices. For instance, many DAS stated that data were available “on request” or in supplementary files, but lacked clarity. A standardised way of expressing such access conditions—avoiding vague phrases like “contact the author”—is needed.

Other challenges included inconsistent article type labelling (e.g., “letter” used for full research articles), substantial variation across research fields, and the lack of standardised DAS language and templates. These issues affect both manual and automated analysis.

Despite these limitations, the pilot fostered collaboration across stakeholders and highlighted the potential for collective action, most importantly on standards and common definitions.

Downstream effects of sharing data and other research outputs

This pilot project addressed a key gap in research evaluation: understanding the broader, downstream impacts of research outputs—beyond citation counts—including reuse in new research, policy, and societal applications. Recognising that current metrics inadequately capture these effects, the project brought together universities, solution providers, and infrastructure experts to co-develop new indicators and methodologies, with a focus on shared datasets.

The pilot's main achievement was a proof-of-concept indicator that combined citation network analysis with qualitative narrative interpretation. This enabled institutions to trace the influence of research outputs over time, offering richer, more responsible narratives of impact—particularly useful for exercises like REF2029. Participating universities shaped the project around their open research goals, seeking actionable insights to support researchers and promote open research.

Key outputs included:

- A prototype graph exploration tool using OpenAlex data to trace citation pathways;
- A set of institutional use cases and indicator requirements;
- Explorations by providers into data reuse, revealing methodological and disciplinary challenges;
- Recommendations on infrastructure gaps, such as inconsistent use of persistent identifiers and limitations in repositories.

Findings showed that while data citation is growing, it remains rare and inconsistently reported. Many institutions and providers lack systems to monitor reuse or generate the metadata needed to assess impact. The pilot's scope expanded to all research outputs due to limited metadata on shared datasets.

Quantitative results were constrained by data availability, citation inconsistencies, and database biases. Moreover, if the approaches piloted were pursued further, inequities in institutional resources – access to tools, infrastructure, and expertise – may limit the ability to demonstrate impact, reinforcing structural disparities. The pilot calls for systemic change to ensure research evaluation is fair, inclusive, and representative, and encourages institutions to reflect on their practices using existing tools and resources.

Pre-registration

This pilot explored the feasibility of using automated methods to identify whether institutional research outputs were pre-registered. The three solution providers (OpenAIRE, Digital Science, and PLOS/DataSeer) successfully developed approaches to detect self-reported pre-registration, with relatively consistent performance across providers. Accuracy exceeded 80% in general, though this was influenced by the unusually high prevalence of pre-registration in the selected sample. Sensitivity (correctly identifying pre-registered outputs) and specificity (correctly identifying non-pre-registered outputs) were both close to or above 80%. False negatives were under 10%, while false positives varied more widely, ranging from 3.5% to 15%.

Several limitations were noted. First, the human-coded benchmark may itself contain errors. Second, pre-registration was defined solely as a self-declaration; the pilot did not verify whether the content of the pre-registration matched the way in which the research was ultimately carried out. Third, the pilot did not assess whether the registration occurred before the research began. Fourth, while sensitivity was strong, specificity was slightly lower, indicating that automated tools were better at detecting true positives than ruling out false ones.

Importantly, the pilot also demonstrated the value of collaborative learning. Providers refined their algorithms by learning from each other's methods, and the hand-coded dataset served as a useful benchmark for tuning performance.

The pilot concluded that:

1. Automated identification of self-reported pre-registration is feasible and reasonably accurate.
2. Further work is needed to assess the timing and quality of pre-registration, beyond simple detection.
3. Encouraging researchers to link the final outputs of their research to pre-registrations—even if imperfect—adds transparency and enables third-party evaluation of quality and compliance.

This pilot highlights the potential of automated tools in supporting open research practices, while also underscoring the need for clearer standards and more robust metadata to ensure meaningful evaluation of pre-registration practices.

The use of CRediT

This pilot assessed the feasibility of using automated methods to identify CRediT (Contributor Roles Taxonomy) statements in research outputs. While CRediT offers a controlled taxonomy for author contributions, its inconsistent use across publications presents challenges for automated detection. The pilot found significant variation in how authorship statements were written and how CRediT roles were applied, affecting both sensitivity and specificity of text-mining approaches.

Common issues included:

- Coincidental use of CRediT terms without intentional taxonomy use, leading to false positives.
- Narrative phrasing or tense variations (e.g., “XX conceptualized the study”) that deviated from standard CRediT terms.
- Spelling differences (e.g., “conceptualization” vs “conceptualisation”).
- Formatting issues such as line breaks or unusual headings that disrupted algorithmic extraction.

The two solution providers, OpenAIRE and Digital Science, returned smaller datasets than expected. Both achieved high accuracy (>94%) in identifying when no contributor statement was present. However, during the pilot they varied markedly in detecting actual CRediT statements, with recall rates between around a third and over 90%, and differences also in rates of false positives due to varied CRediT statement formats. Both providers are rapidly improving their approaches to better handle these variations while balancing sensitivity (recall) and specificity (precision).

The pilot also noted that accuracy figures were based on a small, strictly defined validation sample, which may not reflect broader capabilities. Overall, the findings underscore the need for both standardised use of CRediT roles and their documentation in metadata, and improved algorithmic approaches to reliably detect and interpret author contributions.

Recommendations

Adoption

While most of the pilots concluded that more work was needed before automated indicators might be a viable option, some immediate prospects for cautious adoption were identified. These were: some elements of FAIR, pre-registration and, perhaps with some further tuning of the approach, CRediT. With care, an awareness of the limitations of the indicators, and appropriate contextualisation, institutions may have some confidence in specifying a requirement for monitoring systems in these areas with some expectation that the requirement will be met. Institutions adopting indicators should continue to compare the results with those from manual checks and, as far as possible, share the results to help improve the indicators.

Common definitions and vocabularies

These was a consensus across all pilots that active and deliberate work was needed to develop common, clear and robust definitions of the dimensions of open research that should be a priority for monitoring. This will require one or more trusted organisations to take a leadership role with stakeholders across the research system. While the global community of Reproducibility Networks can play a role, the more obvious candidate lead organisation might be the Open Science Monitoring Initiative, OSMI. Guided by the INORMS SCOPE framework, this work should reference the purposes for which monitoring takes place, and therefore the full range of relevant information, the limits of applicability, and the relevant context into which the indicator should be placed, etc. For example, where that information is quantitative, then agreement is needed on appropriate numerators and denominators. While taking into account current and anticipated system capabilities, this work should be led by the purposes for which the monitoring is intended.

Even where some clear definitions are available, the same terminology is not always used. Work is needed to align vocabulary across the research system.

This work will inevitably mean that some actors will need to change their practices. For example, systems may need to be re-coded or organisations may need to change procedures or workflows. Guidance derived from abstractions such as those developed within the OSTrails project may help in complex situations.

Examples where this work is urgently needed include: DAS, article-type definitions, and data curation actions.

System capabilities and interoperability

The pilots noted specific examples where particular systems, such as repositories, journal submission systems and indicator data sources, should improve their capabilities to better capture, hold and share metadata and persistent identifiers. These improvements should be planned with a view to improving not only immediate or local operations but also to improve workflows between systems across the research information landscape. This will enable better monitoring through improved metadata and persistent identifiers, more efficient inter-working between stakeholders and therefore smoother workflows for researchers, and more opportunities to reward those contributing to research including researchers and research enabling staff, journals, repositories and institutions. It is likely that forums will need to be deliberately established or adapted, and stakeholder interests addressed, in order for this coordinated approach to gain traction. As a signatory to the Barcelona Declaration for Open Research Information, UKRN would argue that the benefits outlined above are most likely to be quickly and thoroughly realised if the metadata are open.

Better workflows

If workflows are improved, then those supporting researchers will have greater confidence and influence when they go to researchers and advocate for good practice, for example in the provision of high-quality metadata, use of DAS, use of persistent identifiers, linking to pre-registrations, etc. This guidance is needed for researchers, journal editors, and others but, given the pressures people are under, it is only likely to be adopted if measures are first taken to make that easier, via better workflows.

Manual checks

Several of the pilots recommended that manual monitoring of open research practices should continue, partly because the automated approaches are not yet sufficiently robust, and partly to enable that robustness to be checked as it improves. Within the pilots we developed a protocol for manual checking and there could be sector-wide benefits if those checks were done in a consistent way and the results shared, if that could be done without enabling spurious comparisons to be made.

Good practice in monitoring

Being led by institutional experts and partnering with specialist indicator developers, the pilot teams were extremely sensitive to the need to monitor carefully. This included considering the stages in the INORMS SCOPE framework but went beyond that to the development of agreements and principles on, for example, partnership and transparency. Further work is needed to build on this, and on related work in many other places, to develop, share and embed good practice in the design, delivery and use of indicators even when, as in these pilots, those indicators are at an aggregate level that does not identify individuals. As noted above, an obvious candidate lead organisation might be the Open Science Monitoring Initiative, OSMI.

Future projects

The pilot teams achieved a huge amount based on voluntary efforts and lacking most elements of overarching management. Any work similar to the pilots in the future should be properly resourced and managed, for example to identify sequencing and dependencies, and to avoid putting unreasonable demands on operational teams.

Most of the solution providers noted the potential of generative AI and specifically Large Language Models (LLMs) in future projects, though some participating institutions and UKRN more generally note concerns about, for example, the reproducibility of indicators derived using such technologies.

Conclusions

The six pilots have explored whether robust, automated monitoring of certain open research practices is feasible. The answer is nuanced and varies between different practices. In some cases, such as making data FAIR and pre-registering studies, particular aspects of those practices can be monitored with a reasonable level of robustness. There is some prospect that automated text processing methods such as LLMs will be able to address other aspects of open research practices in due course. For many other aspects of open research that institutions wish to monitor, institutional experts will need to continue to rely on resource-intensive manual checks for now. For many aspects of open research that institutions wish to monitor, the pilots have revealed a clear need for the development and adoption of standards, good practice and guidance, for this to be coordinated across the research system, and for automated indicators to be continually compared with evidence of ground truth, usually achieved via manual checks. Leadership and resources are badly needed to achieve this.

Overview Report Annexes

- Overview Report Annex 1. Considerations for a coordinated approach between the data pilots
- Overview Report Annex 2: SCOPE Reviews by institutions related to the various pilots
- Overview Report Annex 3: Transparency principles
- Overview Report Annex 4: Data reports from selected providers
- Overview Report Annex 5: Manual validation check protocol
- Overview Report Annex 6: Summative evaluation of the UKRN open research indicator pilots by CWTS

Overview Report. Annex 1: Considerations for a coordinated approach between the data pilots

(April 2024)

Pilot 1: The FAIRness of data	Pilot 4: The prevalence of adequate DAS
Pilot 2: The Openness of data	Pilot 5: Quality and reliability of DAS

Background

During the Pilot Leads subgroup meeting of 27 March, the pros and cons of a coordinated approach were discussed. The rationale for this is that a coordinated approach could bring efficiencies and enhance the final outputs due to commonalities between the areas examined by each pilot. We tried to identify the sources of discomfort and challenges faced by the pilots and identify opportunities for optimisation.

The main items discussed are listed below. We invite all the leads of these pilots to review and contribute to these and list new ones in order that a decision can be made on future collaborations between the pilots. Please enter new items using the tracked changes function and comment using the Word comment function.

Collaboration opportunity	Pros	Cons	Considerations
General information sharing between the pilots	Could bring efficiencies, such as identifying areas of overlap (work with providers, datasets)	There is a time burden to information sharing	An efficient mechanism would be needed, one which did not result in extra burden for the leads and additional emails; i.e agree a communication policy / ways of working Data storage considerations Provider access considerations
The FAIR and open pilots will take as a starting point published articles and the data availability indicators	Significant time efficiencies and avoiding unnecessary duplication of work.	Organisational challenges	Would require agreement that this is the best approach; DAS pilot perhaps benefits less than the others.

within them. This means there is a logical sequencing between the three pilots (DAS -> open -> FAIR) which, if optimised, could mean streamlining of workflows for all three projects (and definitely for FAIR and open). Discussions with external providers could also be streamlined and/or combined across the three projects.			Would require clear organisation structure and timetable agreed among the co-leads. Meeting has been arranged by pilot 4 for April 12th to discuss common dataset.
Two of the three **pilots currently have two leads which could lend itself well to a division of responsibilities for subgroups** within a consolidated pilot	This would allow each lead to focus on a specific set of actions	Potential for breakdown in communication	Is there a natural division of responsibilities which could create efficiencies? Who would take the lead for the consolidated pilot?
Initial scoping work such as acknowledging prior work in the field could be shared	Demonstrates a consistent approach, economies in sharing information	Whilst the 4 areas are connected they are not the same; there are a variety of technical aspects of each which need to be considered	Would need to demonstrate that initial scoping work is thorough and that no essential elements of each pilot have not been missed, and that we are not reinventing the wheel
Sharing potential indicators and those which have been	Demonstrates a consistent approach	Potential for indicators to be omitted or mis-interpreted	Efficient information sharing mechanism is required

discounted between the groups			
A common dataset could be produced which could be the basis for the 3 **pilots**	Would bring a degree of standardisation across the 4 pilots Initial time taken to establish the dataset should be offset by efficiencies in avoiding duplication of effort.	Could introduce bias into the pilots by limiting the focus to the same institutions Common dataset across such diverse institutions? This may not be reflective of other institutional interest in a specific pilot. Data sample must be representative (agreement needed, representative of what?).	Rationale for institutions included in dataset needs to be robustly documented Other groups such as Pilot 6 have expressed an interest in this. At the moment, Reading, Bristol and Glasgow (and Leicester?) committed to providing data, but the sample of data to be used should be more representative of the make of HEI. A meeting has been organised to discuss a common dataset for pilots 1,2 and 4.
Given that most pilots have the same participants, in replacement of individual pilot meetings, it could mean a shared recurring Sprint meeting (monthly) for all 3 pilots, and subgroups could self-organise when and if needs be.	Less meeting and more efficient communication	It is possible that we lose/prioritise a given set of questions, and do not address others which would have otherwise been addressed by the pilots. Not all pilot participants are members of multiple pilots (resourcing issues etc). So may have purposely chosen to participate actively in a pilot, but no capacity to be involved in all other pilot meetings etc. Some pilots can be very technical (FAIRness) and we	

		should not underestimate or lose that aspect. The Openness and FAIRness pilots look beyond the DAS to the nature of the data cited. Fewer, deeper meetings beneficial but must recognise that elements of difference are going to be crucial to individual pilots.	

Overview Report. Annex 2: SCOPE Reviews by institutions related to the various pilots

The following zip file contains the reviews that institutions undertook to document their reasons for wanting to use indicators of open research practices. The documents are rough, working documents.

SCOPE reviews.zip

Overview Report. Annex 3: Transparency principles

UKRN Open Research Indicator pilots: Transparency in reporting Notes from a call, 31 January 2025

Introduction

This is a note of a call among institutions and providers to agree principles to guide reporting from the pilots. The note is not a verbatim record of the call, but a condensed summary that in places makes inferences from what was said on the call. The note should be agreed by all pilot participants before the outputs are released.

Considerations

1. The pilots are pilots. Everyone involved has been working voluntarily and in good faith to understand better ways to monitor aspects of open research. This is hard, and all the outcomes are provisional.
2. The value from the pilots is chiefly in the **methods used**, and in **our assessment of those methods**. It is much less in the output data or in the levels of open research found in the pilots. Further, there are risks, for example of enabling spurious rankings or comparisons, in releasing the output data or levels of open research found.
3. However, some output data may have value should anyone wish to try to reproduce the approaches taken or compare the results with those from a different approach.

Principles

a. We operate according to ‘as open as possible, as closed as necessary’.
b. Status of the findings:
    i. The report will make it clear throughout that the methods that have been piloted are experimental, and many proposed definitions of terms and classifications still need to be agreed by the wider community.
    ii. Any attempt to compare institutions or providers based on any shared data (or absence of data) are illegitimate. All data created in the pilots are experimental, and all institutions and providers are constantly improving their practices.
    iii. The report will compare methods on the basis of their potential to be valid, reliable, transparent, feasible, etc – as set out in the final report template.
c. Input data from institutions:
    i. Input data should be made openly available with sufficient explanation to enable others to understand the nature, scope, extent and limitations of those data.
d. Methods:
    i. Very detailed descriptions of the methods and procedures used by the providers should be made openly available, including query strings where relevant, but these descriptions may exclude some specific software, code and algorithms.
e. Output data that would underpin potential indicators:
    i. Output data may be made openly available as agreed by each pilot (all institutions and providers involved).
    ii. Data may be made openly available at an aggregate and anonymous level, combining data from all participating institutions without identifying them (e.g. excluding names and perhaps DOIs). In some cases, illustrative excerpts from datasets or aggregate summary data might be all that can be made openly available. Data should be released with sufficient explanation to enable others to understand the nature, scope, extent and limitations of those data.
f. Manual checks:
    i. the detailed description of how these checks were done should be made openly available.
    ii. the results of the manual checks should be made openly available for each proposed method, as aggregate and anonymous results for all the institutions combined, giving the numbers of true / false positives and negatives and the balanced accuracy score or other relevant finding, plus commentary.
g. Resource:
    i. The above principles will be adopted in the report so far as resources allow.
    ii. UKRN will enable curation and review of the outputs prior to release, and others involved in the pilots may contribute to this.
h. All participants will approve the report and other outputs before release.

Overview Report. Annex 4: Data reports from selected providers

In some cases providers supplied a data report covering several of the pilots. Those are included here.

PLOS/DataSeer

UKRN_PLOS_Public-Release.xlsx

Overview Report. Annex 5: Manual validation check protocol

Introduction

Some of the pilots plan to undertake manual checks as one way to evaluate some aspects the quality of the proposed indicators. Where that is the case, this document offers guidance on how to do that. Its aim is to encourage consistent checks informed by good practice.

Background

The overall aim of the pilots is to establish good practice in institutional monitoring open research, for example in the design and use of dashboards and reporting tools. Where a pilot has been able to propose a candidate method, working with providers, then it may take various approaches to check its validity and reliability. For example:

1. A reflection on the extent to which the proposed method does, in principle, yield an indicator that is a good proxy for the phenomenon it purports to represent. This might be best done through a review of the literature, and discussions within the pilot including all the relevant providers, and with CWTS.
2. Assuming that (1) is reasonably positive, then a check on the extent to which the proposed method corresponds to checks that a human would make.

The remainder of this protocol relates to (2).

Aims

The aims of the manual check of a method are to assess its:

i. validity - does it measure what it claims to?
ii. reliability – does it give the same measure in the same circumstances?

This is operationalised in this test for the OR indicators as:

i. validity = accuracy, that is the proportion of all instances that the method correctly identifies as either to be counted or not to be counted. This can be derived from the rates of false positives and false negatives.
ii. reliability = reproducibility, that is the extent to which accuracy is consistent between tests using the same conditions

Protocol

1. Agree the set of indicators to be tested. Each pilot has identified several specific features of open research that contribute to their overall aspect of open research. For example, the FAIR pilot has identified several features that would indicate that data is ‘Findable’. Some or all of these features might be subject to the test, depending on whether the proposed method does claim to indicate those features.
2. Decide what counts as that feature being present:
    a. Discuss this with any relevant provider and with CWTS, so that both the method and the test are working to the same definition of ‘what counts’.
    b. Manually review the content and links from 10 DOIs from the set compiled by pilot institutions, looking for edge cases.
    c. Review how the method categorises these 10 DOIs, and discuss with provider
    d. Refine the description of ‘what counts’ to account for these cases.

At this point, there is a choice.

- If the feature is relatively well-defined and relatively common, then use steps 3-7.
- If the feature is relatively well-defined but relatively uncommon, then use steps 8-13
- If the feature is not relatively well-defined, then use steps 14-15

For features that are relatively well-defined and relatively common

3. Randomly select a sample of DOIs from the set compiled by pilot institutions, excluding the 10 reviewed in Step 2. Otherwise, the set should be random and should be of at least 100 DOIs.
4. Choose one of the following approaches to checking different aspects of reliability:
    a. Have the manual check done by at least two reviewers, who either compare notes to converge on a single set of scores *(thus improving this aspect of reliability)* or record different scores and a Kappa coefficient for inter-coder reliability *(thus recording this aspect of reliability)*.
    b. Repeat steps 3-6 with both a different reviewer and, if possible, a re-run of the data extraction using the method. The difference between this balanced accuracy score and that of the first test is a measure of 'reliability'.
5. Undertake the manual check.
    a. Review the content and links from the sample of 100 DOIs, looking for the agreed features. For each DOI, and each feature, score 1 if the feature is present and 0 if it is not.
    b. Review the outcomes of the method developed by the pilot with the data provider. For each DOI, and each feature, score 1 if the method indicates that the feature is present and 0 if it does not.
6. Calculate a balanced accuracy score. This is a score for 'validity'. For more detail on how to do this, see [[5]]. A spreadsheet will be provided to do this calculation.
7. Discuss the outcomes with providers involved in the method and agree how the outcomes of the checks will be reported.

For features that are relatively well-defined but relatively uncommon

8. Choose an approach to checking different aspects of reliability, as step 4.
9. From the set of DOIs compiled by pilot institutes, identify a random set of outputs sufficiently large to yield 30 that are deemed positive for the feature by the proposed method.
10. Manually check each of these 30 positives to identify false positives (i.e. outputs wrongly declared as having the feature). Score 1 for each false positive.
11. From the original set of outputs, randomly sample 30 that the proposed method does not deem positive for the feature. Manually search these 30 for false negatives. Score 1 for each false negative.
12. Calculate a balanced accuracy score, as step 6. A spreadsheet will be provided to do this calculation.
13. Discuss the outcomes with providers involved in the method and agree how the outcomes of the checks will be reported.

For features that are not relatively well-defined

14. Develop short case studies illustrating the issues of reliability, based on either the sample used in step 2, and/or manually identified extreme and edge cases, and/or well-known cases.
15. Discuss the outcomes with providers involved in the method and agree how the outcomes of the checks will be reported.

[5] If T = True, F= False, P = Positive and N = Negative, then accuracy = (TP+TN)/(TP+TN+FP+FN). Because our data will be imbalanced between P and N, then this will be normalised = (TP rate + TN rate) / 2. For further information, see https://en.wikipedia.org/wiki/Precision_and_recall

Overview Report. Annex 6: Review of the UKRN open research indicator pilots report by CWTS

OR Indicator pilots
- final report outline