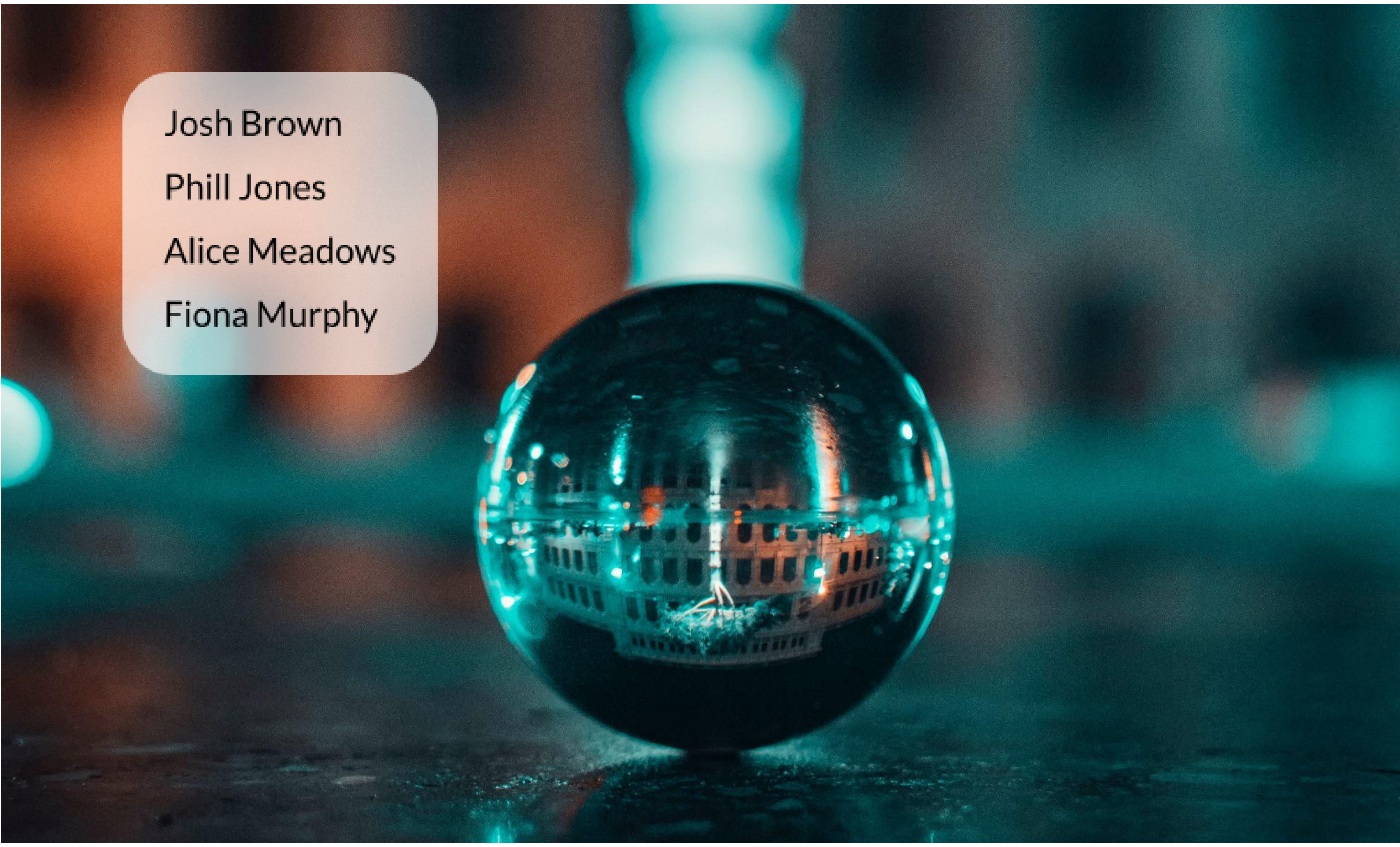
UK Research and Innovation

MORE BRAINS

Metadata to support the UKRI Open Access Policy:

Landscape and community readiness analysis



Josh Brown

Phill Jones

Alice Meadows

Fiona Murphy

29 October 2021

Revised November 28 2022

DOI: 10.5281/zenodo.7386901

Table of Contents

The authors would like to gratefully acknowledge the contributions of Petr Knoth and Matteo Cancellieri of CORE (https://core.ac.uk/) to this report. They were commissioned by UKRI to provide the data and graphs relating to the UK repository landscape, which are included in chapter 6 of this report. In addition, they provided many helpful explanations of the nature, significance, and limitations of the available data and reviewed our descriptions of their work. We thank them for their contributions and many helpful exchanges, and note that responsibility for any errors or omissions in the way that data has been presented or interpreted lies solely with the authors.

1 Executive summary

In August 2021, UKRI commissioned MoreBrains Cooperative to evaluate the current state of the landscape of descriptive information (metadata) for journal articles and academic long-form content like books, book chapters, and monographs. UKRI's new open access policy[1] sets out a number of requirements for publications resulting from the research they have funded. Meeting these requirements can often only be achieved or evidenced by the use of descriptive metadata associated with these publications, such as licensing information, persistent identifiers for items and contributors, links to funding grants, and relationships between versions of articles wherever they might be hosted.

These requirements have been shaped by a wide community consultation. They represent a broad agreement on what is needed for access to UK publicly-funded research to be as open as possible.

The policy sets out two routes to compliant open access: route 1 being to "publish the research article open access in a journal or publishing platform which makes the Version of Record immediately open access via its website"; and route 2 being to "publish the research article in a subscription journal and deposit the Author's Accepted Manuscript (or Version of Record, where the publisher permits) in an institutional or subject repository at the time of final publication".

MoreBrains analysed the policy text to identify specific pieces of metadata that are required by each of these routes, and then identified common metadata standards and frameworks (or 'schema') used in the publishing community (for route 1) and the repository community (for route 2). We examined each candidate schema and established which of the metadata required by the policy are currently present, or missing, from each.

For the metadata that is currently present in these schema, we then assessed the availability of information. Not every piece of metadata that could be shared is always present. We endeavoured to gauge the scale and extent of gaps in metadata coming from publishers and repositories, by analysing how often the relevant metadata was present for UK research in two major collections of aggregated information: the Crossref Registry[2] for journal publications; and the CORE[3] database of information from repositories.

Our findings and recommendations are summarised in tables 1 and 2 (below) in detail, but several stand out and are worthy of particular emphasis:

- **For publisher metadata, the most accessible pathway to filling gaps in current coverage is to work within the major standard framework for describing content, the Journal Article Tag Suite (JATS)**[4]. This is widely adopted, and its dedicated community group (JATS for Reuse - JATS4R) is ready to assist in developing guidance and best practice which could deliver improvements in coverage
- Crossref provides unique identifiers for individual articles (among many other things, including funding grants) and gathers a considerable volume of metadata alongside each

[1] https://www.ukri.org/wp-content/uploads/2021/08/UKRI-180821-UKRIOpenAccessPolicy-2.pdf
[2] https://www.crossref.org/
[3] https://core.ac.uk/
[4] https://jats.nlm.nih.gov/

identifier. Crossref's members include both publishers and funders, making the organisation a natural place for collaboration across both sectors. **Crossref should be used as the primary registry for gathering JATS-compliant metadata from more than 14,000 publisher members around the world**

- Publishers display wide variations in the accuracy and completeness of the metadata they provide. We were unable to investigate the diverse possible causes for this, and **we strongly recommend that further work is undertaken — in partnership with publishing community bodies — to identify common issues that currently reduce metadata quality and completeness**
- DataCite's remit is analogous to Crossref's, providing a range of services built around persistent identifiers for outputs and metadata. **If UK repositories (of which approximately 120 are DataCite members) consistently registered identifiers with DataCite for content, and provided policy-required metadata as part of this process, then DataCite could be used as the primary registry for gathering metadata for repository content, with the additional advantage of international coverage**
- Our analysis of repository metadata was limited by variations in practice across the many institutional repositories in the UK. Only about half of those examined use a dedicated standard intended to support the interoperation and analysis of metadata (RIOXX); the others do not, so we were unable to compare their data. **If RIOXX were updated to support the requirements of the UKRI open access policy, AND reliably adopted across the repository landscape, it would be a valuable tool in enabling UKRI to monitor progress towards its policy goals**
- **We recommend that UKRI should support the use of DataCite and RIOXX as a pathway to meeting policy requirements for UK repositories.** Individual repositories should be given the freedom to choose between these approaches, based on their own local requirements with an assertive support programme instituted to ensure that coverage issues are comprehensively addressed
- **Specific policy requirements will present challenges to the community, and may call for additional interventions on part of UKRI to drive improvements.** Examples include funding acknowledgements, which are often missing , and licensing information, which is often both missing and inconsistent. There is a very long tail of small publishers for whom providing machine-readable licensing information in particular will be a challenge. We set out practical steps that could be taken to ameliorate these, and other potentially problematic metadata in our summary of recommendations in section 2
- The UKRI policy calls for ORCID IDs to be supported for authors and, given the policy applies to all authors that are in scope of the UKRI research articles open access policy, it is reasonable to assume that the desirable position is for ORCID IDs to be provided for all authors. Repositories often do not interact with authors, making this data nigh on impossible to collect. Repository experts we spoke to assumed that they would be able to gather such information from publishers, but publishers that collect ORCID IDs typically do so only for the corresponding author — they have little to no contact with other authors. ORCID itself is not yet technically prepared to support the process of gathering these additional IDs, and there are serious workflow and administrative burden challenges. This policy requirement is ambitious. It may take several years before a scalable, practical solution is available, and even longer for this to be rolled out across the publishing and repository systems. **We advise a**

longer-term, phased approach to implementing the requirement for ORCID IDs for UKRI funded authors

The UKRI open access policy has generally been well-received, and is seen as a positive step. However, specific challenges exist across the publishing and repository communities for the creation, sharing, and analysis of complete, accurate metadata. Pathways to improvement exist in almost every case. With appropriate clarification and guidance from UKRI, practical support for changes to standards etc., and leadership by example in leveraging identifiers and the sharing of metadata by UKRI itself, there is an opportunity to further the core goals of the open access policy, and improve the depth and accuracy of our understanding of the UK research landscape.

2 Recommendations

There are two routes to compliance laid out in the UKRI open access policy. As described in section 4.1 and section 5, these are:

Route 1: *The version of record is made available on the primary publisher's platform, provided the platform and article meet the technical, licensing and metadata requirements. The metadata source for this route would therefore be publishers.*

Route 2: *The work is deposited in a repository in addition to being published in a subscription journal. The repository and article in the repository must meet the technical, licensing and metadata requirements and there must be no embargo period. The metadata source for this route would be repositories.*

Below, we summarise our recommendations for each of the two routes to compliance based on the origin for the metadata required to demonstrate policy compliance.

2.1 To support route 1 requirements

Metadata to support compliant open access via route 1 originates with the publisher, as explained in section 4. Below is a table that summarises each of the requirements we discovered for publisher metadata, as described in section 4, along with the current status of readiness accounting for the various levels of adoption we document in section 6. For a glanceable view of the current status of each requirement, we have applied a BRAG classification:

Blue: Metadata fields are available and well-described with adequate adoption
Green: Metadata fields are available and well-described, but adoption needs to improve
Amber: Metadata fields are available but are poorly used or misused
Red: There are no metadata fields and no mechanism to monitor levels of compliance

Req #	Requirement	Content type	Mandatory or encouraged?	Current status	Recommendations
1	Acknowledge UKRI funding	Articles Long-form	Mandatory	🟩	• Work with Crossref and publishers to promote the use of Crossref grant identifiers (DOIs) and best practice • Provide specific instructions on what should be included in metadata and how it should be formatted • Register DOIs for UKRI grants and require their use in funding acknowledgements
2	Date of deposit	Articles	NA		
3	Data Access Statement	Articles	Mandatory	🟧	• Coordinate with Research Data Alliance, STM, JATS, Crossref, Force11, NISO, and Scholix to define good practice and reporting pipelines, and develop standards in machine readable DAS • Work with other funders and publishing associations to encourage uptake of DAS statements in publishing workflows
4	ISSN	Articles	Mandatory	🟦	*No action needed*
5	Article level PID (DOI, URN, Handle)	Articles	Mandatory	🟦	*No action needed*

6	ORCID ID (all UKRI authors)[5]	Articles	Mandatory	🟧	- Address workflow and data protection issues in collaboration with ORCID, publishers, and publishing system vendors - Investigate levels of ORCID coverage among corresponding authors and co-authors and the feasibility of pathways to increasing coverage, in collaboration with ORCID and publishers
7	Is ORCID ID authenticated?	Articles	Encouraged	🟥	- Conduct a follow-up project to investigate actual levels of ORCID authentication and feasibility of increasing levels of authentication, working with publishers, publishing associations and ORCID - Provide support in communicating best practice in detail, describing exactly what tags need to be used to enable verification
8	Licence (non-proprietary format)	Articles Long-form	Mandatory	🟩	- Promote NISO ALI:license_ref tags[6] as best practice for publishers - Work with publishing associations and Crossref to encourage adoption of best practice by publishers. - In collaboration with the community, encourage the development of best practice and specific instructions on what metadata fields need to be included and how they should be formatted
9	Preservation location (Portico etc)	Articles	Mandatory	🟥	- Work with CLOCKSS, Portico, etc. to develop mechanisms of cross-checking preservation locations that don't involve encoding at the article level
10	Self-archiving policy (registered in Sherpa-Romeo)	Articles	Mandatory	🟥	- Promote publisher registration of self-archiving policy with Sherpa-Romeo as best practice

[5] The policy requires that ORCID is to be supported for all UKRI-funded authors and is regarded as desirable for all authors. We have therefore treated this requirement as mandatory for the purposes of our analysis.

[6] http://www.niso.org/schemas/ali/1.0

				🟥	• Work with publishers, publishing associations, and Sherpa-Romeo, and leverage Crossref communications channels to encourage and promote best practice • Work with the community to encourage the development and documentation of best practice for reporting
11	Citation data in I4OC (http://opencitations.net/)	Articles	Mandatory	🟥	• Work with Open Citations to develop mechanisms for cross-checking whether a publication's citations have been included in their databases
12	Repository registered in OpenDOAR	Articles	NA		
13	PID for funders	Articles	Encouraged	🟩	• Work with funders to facilitate transition from Open Funder Registry to ROR IDs • Work with Crossref and publishers to promote best practice • Provide specific instructions on what metadata fields need to be included and how they should be formatted
14	PID for research performing organisations	Articles	Encouraged	🟧	• Work with Crossref, ROR, and publishers to promote best practice • Select ROR as the identifier of choice and strongly recommend or mandate its use • Provide specific instructions on what metadata fields need to be included and how they should be formatted
15	PID for grant	Articles	Encouraged	🟧	• Work with Crossref and publishers to promote the use of Crossref grant identifiers (DOIs) and to promote best practice • Provide specific instructions on what should be included in metadata and how it should be formatted • Register DOIs for UKRI grants and require their use in funding acknowledgements
16	PID for project	Articles	Encouraged	🟥	• Work with RAiD, Crossref, and publishers to support RAiD as it matures into a project identifier that is fit for purpose

				🟥	• Engage in communications to help the publishing, institutional, and researcher communities understand what a project is and how projects differ from grants • Provide specific instructions on what metadata fields need to be included and how they should be formatted once project IDs are sufficiently mature

Table 1: Findings of our examination of current metadata coverage to support route 1 to compliant open access, and recommendations for actions to improve coverage and completeness of that metadata

2.2 To support route 2 requirements

Metadata to support compliant open access via route 2 originates with institutional repositories, as explained in section 4. Below is a table that summarises each of the requirements we discovered for repository metadata, as described in section 4, along with the current status of readiness accounting for the various levels of adoption we document in section 6. For a glanceable view of the current status of each requirement, we have applied a BRAG classification:

Blue:	Metadata fields are available and well-described with adequate adoption
Green:	Metadata fields are available and well-described, but adoption needs to improve
Amber:	Metadata fields are available but are poorly used or misused
Red:	There are no metadata fields and no mechanism to monitor levels of compliance

Req #	Requirement	Content type	for Repositories	Current status	Relevant recommendation
1	Acknowledge UKRI funding	Articles	Mandatory	🟧	• Work with Crossref and publishers to promote the use of Crossref grant identifiers (DOIs) and best practice • Provide specific instructions on what should be included in metadata and how it should be formatted • Register DOIs for UKRI grants and require their use in funding

		Long-form		🟧	acknowledgements ● Drive investment in technical integrations between repositories and other metadata systems, including CRIS systems and the Publications Router
2	Date of deposit	Articles	Mandatory	🟧	● Work with leaders in the repository community to develop best practices around deposition dates
3	Data Access Statement	Articles	Mandatory	🟥	● Work with RIOXX and Datacite to develop the schema to include data access statements ● Promote integrations into other metadata systems like CRIS, Publications Router, and Crossref
4	ISSN	Articles	Mandatory	🟧	● Encourage the repository community to work on mechanisms and workflows to improve levels of ISSN coverage ● Promote the use of RIOXX or the DataCite schema, giving repositories freedom to choose the approach that best works for them ● Provide specific instructions on what metadata fields need to be included and how they should be formatted ● Work with Crossref and Publications Router to develop ways to supplement repository metadata
5	Article level PID (DOI, URN, Handle)	Articles	Mandatory	🟧	● Work with the repository community to promote best practices in identifier reporting ● Provide specific instructions on what metadata fields need to be included and how they should be formatted ● Work with Crossref and Publications Router to develop ways to supplement repository metadata
6	ORCID ID (all UKRI-funded authors)[7]	Articles	Mandatory	🟧	● Work with Crossref and Publications Router to develop ways to supplement repository metadata

[7] The policy requires that ORCID is to be supported for all UKRI-funded authors and is regarded as desirable for all authors. We have therefore treated this requirement as mandatory for the purposes of our analysis.

7	Is ORCID ID authenticated?	Articles	Mandatory	🟧	● Conduct a follow-up project to investigate actual levels of ORCID authentication, and work with the community to investigate feasibility of pathways to increased authentication levels ● Provide support in communicating best practice in detail, describing exactly what tags need to be used to enable verification
8	Licence (non-proprietary format)	Articles Long-form	Mandatory	🟧	● Work with repositories to encourage adoption of NISO ALI:license_ref tags[8] ● Provide specific instructions on what metadata fields need to be included and how they should be formatted
9	Preservation location (Portico etc)	Articles	NA		
10	Self-archiving policy (registered in Sherpa-Romeo)	Articles	NA		
11	Citation data in I4OC (http://opencitations.net/)	Articles	NA		
12	Repository registered in OpenDOAR	Articles	Mandatory	🟥	● Repository registration should not be an output-level metadata field. Work with OpenDOAR to develop reporting mechanisms and workflows for cross-checking registration compliance
13	PID for funders	Articles	Encouraged	🟧	● Work with funders to facilitate transition from Open Funder Registry to ROR IDs ● Work with repositories to promote best practice ● Provide specific instructions on what metadata fields need to be included and how they should be formatted
14	PID for research performing organisations	Articles	Encouraged	🟥	● Work with repositories and ROR to promote best practice ● Select ROR as the identifier of choice and strongly recommend or mandate its use ● Provide specific instructions on what metadata fields need to be included and how they should be formatted

[8] http://www.niso.org/schemas/ali/1.0

15	PID for grant	Articles	Encouraged	🟥	• Work with Crossref and repositories to promote the use of Crossref grant identifiers (DOIs) and best practice • Provide specific instructions on what should be included in metadata and how it should be formatted • Register DOIs for UKRI grants and require their use in funding acknowledgements
16	PID for project	Articles	Encouraged	🟥	• Work with RAiD and repositories to develop RAiD into a project identifier that is fit for purpose • Engage in communications to help the publishing, institutional, and researcher communities understand what a project is and how projects differ from grants • Provide specific instructions on what metadata fields need to be included and how they should be formatted once project IDs are sufficiently mature
17	Final version or AAM	Long-form	Mandatory	🟧	• Work with the repository community to promote best practices in version recording • Promote registration of DOIs or other PIDs for content • Promote the use of RIOXX or the DataCite schema, giving repositories freedom to choose the approach that best works for them • Provide specific instructions on what metadata fields need to be included and how they should be formatted • Work with Crossref and Publications Router to try to develop ways to supplement repository metadata

Table 2: Findings of our examination of current metadata coverage to support route 2 to compliant open access, and recommendations for actions to improve coverage and completeness of that metadata

3 Introduction and Background

UK Research and Innovation (UKRI) began a review of its open access policy in 2018. After informal consultation, analysis, and evidence-gathering, a revised policy was released for formal consultation in 2019. That consultation closed in May 2020, and the final policy was announced in August 2021.

UKRI's goals for the policy are ambitious, and reflect the important role UKRI is playing in advancing open access to publications arising from publicly-funded research, not least in its active participation in cOAlition S. However, the transition to open access involves many stakeholders: funders, institutions, publishers, libraries, institutional and subject repositories, and, of course, researchers who play major roles throughout the scholarly communications process as authors, reviewers, editors, and readers. Communicating the practical requirements of the UKRI OA policy to each of these groups will be a challenge. The policy requirements need to be clear, achievable, and, crucially, measurable.

The fundamental goal of the policy is for UK research to be openly accessible to all without paywalls or limits on the re-use of the ideas expressed in research outputs. However, there are variations in practice across the publishing and repository landscape, which has complicated our understanding of what progress is being made towards this goal. Much of the technical guidance in the policy therefore relates to the descriptive information, or metadata, which is associated with research outputs. These metadata elements can take the form of information about the article itself (e.g. the title or authors), the publication venue (e.g. the journal name, or volume/issue numbers), the location of the article (e.g. a link to a digital copy), or conditions of access to the article (e.g. the licensing terms under which it is made available to the reader).

Some metadata are vital for open access. The licence, which sets out what authors, readers, and other stakeholders like institutions, funders, and aggregators, are allowed to do with the article is key; it can grant wide permissions that maximise access and re-use or, conversely, it can block certain uses. Other metadata are crucial to help potential readers discover articles. Both these factors are important to open access, particularly that of broader dissemination, since accessing an article requires learning that it exists and knowing how to find it. Still other metadata can be classified as more administrative, including employment affiliations for authors and funding acknowledgements for the research described in the article. These are essential pieces of information for the UKRI OA policy. They show which articles are subject to the policy requirements, as well as supporting relevant processes, such as paying article processing charges (APCs) for OA articles and reporting on outcomes from funded projects.

There are many potential publication venues for an article, and many services or platforms that gather together, or aggregate, metadata. This can result in inconsistencies in metadata formats and structures, which can then make synthesis and analysis difficult or impossible. To address this, standards have been developed to ensure that particular pieces of information are expressed in a consistent way, no matter the source. This can work at the level of an individual metadata element (e.g. defining a fixed way to express the licence for an article) or for groups of metadata elements.

In this study, we have looked at some common rules for grouping metadata elements together: schemas and application profiles. A schema is a logical plan showing the relationships between metadata elements, normally through establishing rules for the use and management of metadata[9], while an application profile is a way to bring together elements from one or more schemas to support a particular application or re-use of the metadata. At the very least, all these standards and profiles need to be consistently structured and well documented for the community to be able to use them. Further, communities of practice need to be developed where best practice can be developed, agreed, and socialised through outreach and education.

The UKRI policy relies on every participant in the research communication process creating and sharing consistent, reliable, standardised metadata. Without this, it becomes difficult or impossible to bring together information from many sources for analysis and, therefore, difficult or impossible to understand and track the policy's impact on the world of research communication.

With this last point in mind, UKRI commissioned a project to evaluate the current state of the descriptive metadata landscape for (primarily) journal articles and academic long-form content like books, book chapters, and monographs. The OA policy sets out metadata-related requirements, such as licensing information, persistent identifiers for items and contributors, links to funding grants, and relationships between versions of articles wherever they might be hosted. These requirements are the result of the evolution of the landscape since the original RCUK policy was adopted, and have been shaped by a wide community consultation. They represent a broad agreement on what is needed for access to UK publicly-funded research to be as open as possible.

Critically, much of the metadata that is required by the policy originates with either publishers or repositories, depending on the route to compliance that is taken by researchers. If journals, repositories, and other open content platforms are not able to provide these metadata, they will not be able to support the researchers and research-performing organisations that are covered by the policy. Therefore, publishers and other content platforms will be under pressure from their users to reliably and consistently make this information available in a way that will support automatic harvesting, aggregation, and analysis. Without this, it will be laborious and impractical to report on open access and to demonstrate how the landscape is changing as a result of UKRI and other funders' policy decisions[10].

In commissioning this project, UKRI specified three objectives (as stated in the request for proposals):

1. *To provide UKRI with an understanding of the current landscape of metadata schema, standards, and profiles*
2. *To provide UKRI with an understanding of the capability of the current metadata landscape to support the UKRI Open Access Policy, including developing illustrative use cases*
3. *To enable UKRI with the ability to make informed decisions on how to support metadata to enable the UKRI Open Access Policy*

These objectives required a detailed assessment of the explicit and implicit metadata requirements of the UKRI OA policy, an evaluation of the current ability of open access content and metadata

[9] As defined in ISO 23081 https://www.iso.org/obp/ui/#iso:std:iso:23081:-1:ed-2:v1:en

[10] See, for example, the UK government's recent policy paper outlining the scale of bureaucratic burden in HE and research: Department for Business, Energy, and Industrial Strategy/ Department for Education. (2020) Reducing bureaucratic burden in research, innovation and higher education. Available at: https://www.gov.uk/government/publications/reducing-bureaucratic-burdens-higher-education/reducing-bureaucratic-burdens-on-research-innovation-and-higher-education

providers to meet those requirements, and the identification of barriers and opportunities on the pathway towards open access, as well as an indication of actions that may be needed to overcome or take advantage of those.

The MoreBrains team surveyed the most commonly adopted metadata standards, schemas, and profiles (such as RIOXX[11], Dublin Core[12], and JATS[13]) in use across metadata sources and aggregation points. We also looked at relevant international persistent identifier (PID) systems (Crossref[14] and DataCite[15]). PIDs are globally unique alphanumeric strings that act as long-lasting digital references to objects, which, when maintained, automatically redirect users and machines to the resource even when the primary URL might change. We included PID providers because creating unique identifiers for publications via these services includes sharing metadata about the item, and both services use structured schemas that are broadly compatible. A key characteristic of PIDs is persistence. Importantly, persistence is a function of organisational systems, not technologies, so PID providers need appropriate community governance and continuation plans that are not tied to support from a single organisation. As such, they represent valuable and timely sources of information about publications and other research outputs. Crossref counts some ~14,000 publishers amongst its members, and DataCite has ~120 UK institutions in its membership, meaning that both have potentially excellent UK coverage.

To bring together our assessment of these, we undertook a mapping exercise to match the specific requirements of the policy to elements with the various schemas, and used this to evaluate the current capabilities of each to address the metadata requirements of the UKRI OA policy.

For metadata not currently adequately supported by existing schemas and profiles, we have set out an indication of the process for amending or extending the relevant community standards. This includes likely timelines for the public release of any revisions. In some cases, we recommend more efficient ways to implement workflows to obtain the same information that does not involve changes to schema. We also provide an assessment of any barriers or challenges that might cause a time lag from schema release to widespread adoption and implementation of the revised elements.

For metadata supported by existing schemas and profiles, we conducted an assessment of the current levels of completeness.

Bringing these analyses together, we conducted targeted interviews with relevant experts to understand the reasons for gaps or omissions in metadata coverage and/or standards adoption, and identify any incentives that create inconsistencies in coverage or openness of metadata from particular sources or community segments. Our interviews also helped us to surface particular challenges that may emerge in seeking to comply with the policy, and highlighted opportunities for efficiency gains or reductions in the administrative burden of managing the publication, funding, analysis, and reporting of outputs and metadata.

[11] https://rioxx.net/repositories
[12] https://dublincore.org/
[13] https://jats.nlm.nih.gov/
[14] https://www.crossref.org/
[15] https://datacite.org/

4 Methods

4.1 Policy analysis

To understand the degree to which it is possible to monitor compliance with the UKRI open access policy through metadata monitoring and analysis, our first step was to analyse the policy itself. The text of the policy was reviewed line-by-line, with requirements categorised as either mandatory or recommended as well as applying to either articles or long-form as defined by the policy itself.

There are two routes to compliance with the policy. As we explain in section 4, an important difference from a technical perspective is that the metadata to be monitored will originate in different parts of the scholarly ecosystem. Requirements were therefore categorised based on where the relevant metadata would originate (either from publishers or repositories).

The next step was to compare the requirements of the policy to relevant application profiles and schema. Again, the two routes to compliance correspond to metadata originating from two different places. For historical, technical, and cultural reasons, publishers and repositories take different approaches to metadata, using different standards and schemas.

We selected the most relevant schemas to compare with the requirements for each metadata origin: (1) JATS-NLM, and the Crossref metadata schema for publishers; and (2) DataCite, RIOXX3, and Dublin Core using OpenAIRE guidelines for repositories. By inspecting the documentation for each of the schemas and application profiles, we identified the specific fields that would be populated by metadata for each requirement of the policy. Importantly, we noted when no specific metadata field was available that could be used to satisfy requirements, indicating requirements that cannot be monitored through metadata without further development of the schemas and application profiles.

4.2 Metadata landscape review

Once we had established which requirements could technically be monitored through metadata for each of the routes to compliance and for each of the content types specified in the policy, we developed estimates for the level of metadata completeness for fields we could identify through the schema analysis.

Given the time and resource constraints of this project, we streamlined this process by assessing the data available at major aggregation points. Specifically, we used CORE[16] for institutional repository content, and Crossref, via a collaboration with Curtin Open Knowledge Initiative (COKI), for journal articles, to identify gaps or common omissions in coverage. The findings from this technical review are presented in section 5; the complete schema review including the specific field tags for each schema and application profile are presented in Appendix A.

Working with CORE and COKI, we asked each metadata aggregator to query their databases for how many records contain metadata for each field. Attempts to assess metadata completeness levels for UK repositories that do not use RIOXX proved impossible within the scope of this project. Despite the existence of the OpenAIRE guidelines, our partner, CORE, advised us that metadata structure for

[16] https://core.ac.uk/

these repositories was too variable to allow for programmatic assessment of metadata completeness.

Not every requirement that we identified could be directly queried through this approach. For example, whether every author on a publication has an ORCID record is not directly queryable. However, we were able to make some observations by comparing numbers of authors to numbers of ORCID IDs. Other requirements, such as embargos and deposition dates in repositories, are associated with more complex challenges.

Some requirements, for example, whether citations are available in the open citations database or whether journal specific archiving policies are recorded in SHERPA-ROMEO, should not be considered to be article-level metadata. We have suggested alternative approaches, like cross-referencing, as an alternative means for assessing these requirements. These issues are explored in greater depth in section 6.

4.3 Interviews and qualitative analysis

In tandem with the quantitative analysis, we conducted a qualitative investigation through a series of semi-structured interviews. Calls were scheduled between 26 July to 24 September 2021, with the first two being conducted under NDA restrictions. Interviewees were selected on the basis of both individual expertise, such as experience with policy, and their ability to speak from a specific stakeholder viewpoint. Thus there were conversations with service providers, repository representatives, and publishers. Interviewees were asked for permission to record the sessions so that all their insights could be recorded. A sample interview outline is given in Appendix C.

When the interviews were all completed, we synthesised the key findings along several axes, including 'general: stakeholder type', 'technical: social', and 'direct policy: additional findings'. We have grouped the main points in section 7, and have leveraged many of the insights arising throughout this report as indicators for its recommendations and remedies.

5 Policy analysis

We began by exploring the degree to which it is technically possible to monitor policy compliance levels by analysis of publicly available metadata. The policy sets out a series of requirements for research articles and long-form content:

Research articles: *Peer-reviewed research articles, including reviews and conference papers, that are accepted for final publication in either a journal, conference proceeding with an International Standards Serial Number (ISSN), or publishing platform.*

Long-form content: *Academic monographs, book chapters, edited collections and trade books (when the sole output of UKRI-funded research). Scholarly editions, exhibition catalogues, scholarly illustrated catalogues, text books, and works of fiction are out of scope. Each of these categories is further defined in Annex 1 of the policy itself.*

Requirements are further broken down by the route that an author takes to ensure compliance with the policy. Each route to compliance will dictate where the metadata originates that would need to be checked.

Route 1: *The version of record is made available on the primary publisher's platform, provided the platform and article meet the technical, licensing, and metadata requirements. The metadata source for this route would therefore be publishers.*

Route 2: *The work is deposited in a repository in addition to being published in a subscription journal. The repository and article in the repository must meet the technical, licensing, and metadata requirements and there must be no embargo period. The metadata source for this route would be repositories.*

Table 3 (below) shows the results of our analysis of the UKRI policy metadata requirements for articles and long-form outputs, categorised by requirements for content platforms (publishers) for Route 1 compliance, and institutional repositories for Route 2 compliance. We also indicate whether the metadata requirement is mandatory or encouraged. Requirements are numbered because they are referenced throughout the report.

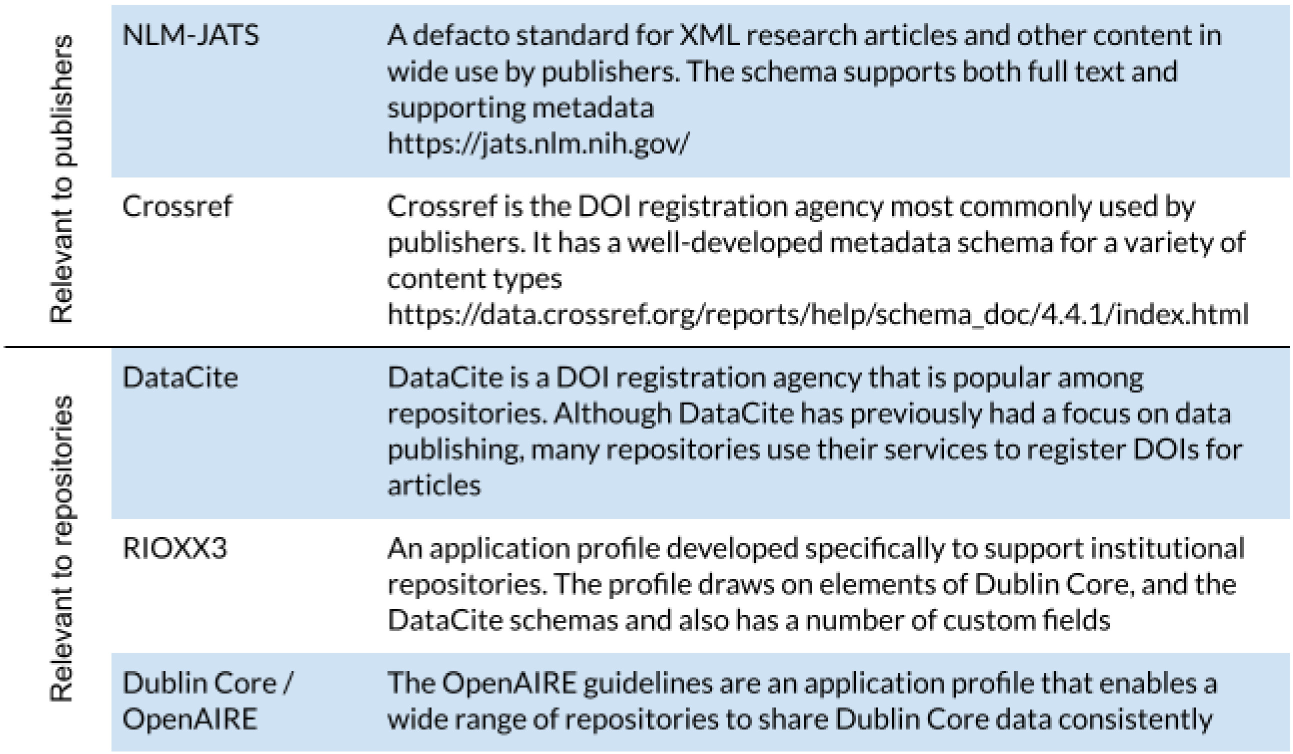
Req #	Requirement	Content type	for Publishers	for Repositories
1	Acknowledge UKRI funding	Articles	Mandatory	Mandatory
2	Date of deposit	Articles	NA	Mandatory
3	Data Access Statement	Articles	Mandatory	Mandatory
4	ISSN	Articles	Mandatory	Mandatory
5	Article level PID (DOI, URN, Handle)	Articles	Mandatory	Mandatory
6	ORCID ID (all UKRI-funded authors)	Articles	Mandatory	Mandatory
7	Is ORCID ID authenticated	Articles	Encouraged	Mandatory
8	Licence (non-proprietary format)	Articles	Mandatory	Mandatory
9	Preservation location (Portico etc)	Articles	Mandatory	NA
10	Self-archiving policy (registered in Sherpa-Romeo)	Articles	Mandatory	NA
11	Citation data in I4OC (http://opencitations.net/)	Articles	Mandatory	NA
12	Repository registered in OpenDOAR	Articles	NA	Mandatory
13	PID for funders	Articles	Encouraged	Encouraged
14	PID for research performing orgs	Articles	Encouraged	Encouraged
15	PID for grant	Articles	Encouraged	Encouraged
16	PID for project	Articles	Encouraged	Encouraged
1	Acknowledge UKRI funding	Long-form	Mandatory	Mandatory
8	Licence (non-proprietary format)	Long-form	Mandatory	Mandatory[17]
17	Final version or AAM	Long-form	NA	Mandatory

[17] The policy text clearly states that it is a requirement for research articles in both repositories and publisher platforms to have one of the permitted licences in a machine-readable, non-proprietary format. For long-form outputs, the policy merely states that a CC-BY or other Creative Commons Licence, or an Open Government Licence, is acceptable. While Annex 2 to the policy sets out 'technical requirements for research articles', there is not such an annex currently available for long-form outputs and so the policy is technically silent on this point. We have however, for the purposes of this study, assumed a similar approach was intended for both articles and long-form outputs.

Table 3: policy metadata requirements with the content type they apply to and whether they are mandatory or encouraged for metadata originating from publishers (in the case of route 1 to compliance), or from repositories (in the case of route 2 to compliance).

As metadata for route 1 and route 2 originates from publishers and repositories respectively, we conducted a technical review of schemas and application profiles that are used by those stakeholders to express metadata. This analysis is critical to assess the technical feasibility of monitoring progress towards full open access because unless a field is present within a schema or application profile, it is impossible to monitor whether it has been populated.

For historical, cultural, and other reasons related to operating conditions, different schema and application profiles are used by publishers and repositories. A series of relevant metadata schema and application profiles were selected for assessment. These were:



Relevant to publishers	NLM-JATS	A defacto standard for XML research articles and other content in wide use by publishers. The schema supports both full text and supporting metadata https://jats.nlm.nih.gov/
	Crossref	Crossref is the DOI registration agency most commonly used by publishers. It has a well-developed metadata schema for a variety of content types https://data.crossref.org/reports/help/schema_doc/4.4.1/index.html
Relevant to repositories	DataCite	DataCite is a DOI registration agency that is popular among repositories. Although DataCite has previously had a focus on data publishing, many repositories use their services to register DOIs for articles
	RIOXX3	An application profile developed specifically to support institutional repositories. The profile draws on elements of Dublin Core, and the DataCite schemas and also has a number of custom fields
	Dublin Core / OpenAIRE	The OpenAIRE guidelines are an application profile that enables a wide range of repositories to share Dublin Core data consistently

We were unable to practically assess how well Dublin Core would support the policy because its schema is deliberately flexible in design, in order to support as wide a range of applications as possible. As such, without the use of an application profile like RIOXX or OpenAIRE to prescribe how the fields are used, consistency between repositories is effectively impossible. In the UK context, according to our data partner, CORE, adherence to the OpenAIRE guidelines[18] on how to implement Dublin Core are so variable that meaningful computational assessment of metadata completeness is not possible. Further, according to data scientists working in this field, variations in metadata structure between repositories that use Dublin Core without consistent implementations of application profiles, are seriously limiting the capabilities of services that can be developed on top of the aggregated data.

[18] https://guidelines.openaire.eu/en/latest/

We compared the technical metadata requirements that we synthesised from the policy (table 3) with the capabilities of the individual schemas and application profiles to support those pieces of information as documented metadata fields. Our assessment for each field was based on whether the field was described in the documentation for that schema. For example, if we look at the NLM-JATS schema documentation page on licences[19], we see that the use of NISO and access licence indicators (ALI)[20] machine-readable licences are supported. Greater adoption of the ALI standard by the publishing industry as best practice, supported by industry associations, is needed to increase the availability of this metadata. The resulting metadata tag structure would look something like this:

```
<permissions>
      <license>
          <ali:license_ref xmlns:ali="http://www.niso.org/schemas/ali/1.0/">
          http://creativecommons.org/licenses/by/3.0/
          </ali:license_ref>
      </license>
</permissions>
```

Tables 4 and 5 (below) show at a glance which of the fields required for policy compliance are currently supported by the common schemas we assessed, which were selected based on their applicability to each route to compliance.

Req #	Requirement	Content type	for Publishers	JATS	Crossref
1	Acknowledge UKRI funding	Articles / Long-form	Mandatory	Yes	Yes
3	Data Access Statement	Articles	Mandatory	Yes	Yes
4	ISSN	Articles	Mandatory	Yes	Yes
5	Article level PID (DOI, URN, Handle)	Articles	Mandatory	Yes	Yes
6	ORCID ID (all UKRI-funded authors)	Articles	Mandatory	Yes	Yes
7	Is ORCID ID authenticated	Articles	Encouraged	Yes	Yes
8	Licence (non-proprietary format)	Articles / long-form	Mandatory	Yes	Yes
9	Preservation location (Portico etc)	Articles	Mandatory	No	Yes
10	Self-archiving policy (registered in Sherpa-Romeo)	Articles	Mandatory	No	No
11	Citation data in I4OC (http://opencitations.net/)	Articles	Mandatory	No	No
13	PID for funders	Articles	Encouraged	Yes	Yes
14	PID for research performing orgs	Articles	Encouraged	Yes	Yes
15	PID for grant	Articles	Encouraged	Yes	Yes

[19] https://jats.nlm.nih.gov/publishing/tag-library/1.3/element/ali-license_ref.html
[20] http://www.niso.org/schemas/ali/1.0

16	PID for project	Articles	Encouraged	No	No

Table 4: Publisher-relevant fields in the policy mapped to JATS and the Crossref schema

For route 1 compliance, where metadata originates with publishers, we examined the readiness of the JATS standard, as this will be largely what is used to transmit metadata onwards from publisher platforms, and the Crossref schema, as this represents the largest aggregation of metadata sourced directly from publishers.

For route 2, where metadata originates with repositories, we examined the RIOXX and OpenAIRE application profiles, as these are designed to impose a uniform structure on repository data and are both intended to facilitate reporting to funders. We also reviewed the DataCite schema, as 120 repositories in the UK are DataCite members, which could enable them to obtain DOIs for AAMs and other content. DataCite could therefore serve an analogous function for repositories to that of Crossref in the publisher community, making it a viable pathway to gathering data about repository content alongside RIOXX and OpenAIRE.

Req #	Requirement	Content type	for Repositories	RIOXX	OpenAIRE	DataCite
1	Acknowledge UKRI funding	Articles / long-form	Mandatory	Yes	Yes	Yes
2	Date of deposit	Articles	Mandatory	Yes	Yes	Yes
3	Data Access Statement	Articles	Mandatory	No	No	No
4	ISSN	Articles	Mandatory	Yes	Yes	Yes
5	Article level PID (DOI, URN, Handle)	Articles	Mandatory	Yes	Yes	Yes
6	ORCID ID (all UKRI-funded authors)	Articles	Mandatory	Yes	Yes	Yes
7	Is ORCID ID authenticated	Articles	Mandatory	No	No	Yes
8	Licence (non-proprietary format)	Articles / long-form	Mandatory	Yes	Yes	Yes
9	Preservation location (Portico etc)	Articles	Mandatory	No	No	Yes
10	Registered in OpenDOAR	Articles / long-form	Mandatory	No	No	No
13	PID for funders	Articles	Encouraged	Yes	Yes	Yes
14	PID for research performing orgs	Articles	Encouraged	No	No	Yes
15	PID for grant	Articles	Encouraged	No	Yes	No
16	PID for project	Articles	Encouraged	No	No	No
17	Final version or AAM	long-form	Mandatory	Yes	Yes	No

Table 5: Repository-relevant fields in the policy mapped to RIOXX version 3, OpenAIRE, and the DataCite schema

Based on our understanding of the current policy, many of its requirements can be assessed by inspection of metadata, particularly for fields that are currently mandatory. There are gaps in the application profiles for repositories for data access statements and authentication flags for ORCIDs.

Preservation location is a mandatory field in the policy for both publishers and repositories, but is not present in most schemas. In general, preservation location should not be considered an article-level metadata item. Similarly, there are no fields for recording the self-archiving policy associated with a publication, OpenDOAR registration status for repositories, or whether citations have been made open, for the same reasons. In each case, there are more efficient ways to create workflows to obtain the information than expanding the schemas. The full results of our analysis, complete with the specific metadata tags for each schema or application profile that correspond to each policy requirement, can be seen in Appendix A.

The PIDs for funders, research performing organisations, and projects are included in this analysis because PIDs for research management are encouraged in the policy. These PIDs map to the five priority PIDs identified as part of the Jisc PID roadmap[21]. They have varying levels of adoption in the community and there is a need to encourage their development and adoption over time. At present, the PID most commonly used for funders is a DOI, provided by Crossref as part of the Open Funder Registry. Over time, these will be replaced with ROR identifiers for funders, as Crossref has indicated that it will retire the Open Funder Registry, which only identifies funders, with ROR, which identifies a range of research-performing and research-supporting organisations, including funders.

We note that funding information is currently often contained in project information in RIOXX. This is problematic because grants and projects are distinct entities (for instance, a project can have one, many, or no grants, and a grant can be used to fund things that are not projects), and there are distinct PID systems available to support them. Examples in the RIOXX documentation show the project field containing internal funder grant identifiers rather than a persistent identifier.

The example below is taken from RIOXX documentation [22]:

```
<rioxxterms:project
funder_name="Engineering and Physical Sciences Research Council"
funder_id="https://doi.org/10.13039/501100000266">
        EP/K023195/1
</rioxxterms:project>
```

None of the entities identified in this example is a project.

6 Current metadata landscape

Having assessed the ability of current schemas and application profiles to support UKRI OA policy requirements, we then examined the actual coverage of certain key metadata fields. Given the time constraints of this investigation, we did not undertake an exhaustive analysis of the availability of every single field. Instead, we sampled data from key aggregation points for a limited number of fields. Here, we cross reference to the requirement numbers in tables 3, 4, and 5. For both publishers and repositories, we looked at ORCID coverage for authors (req. 6,7), licences (req. 8), and funding

[21] Brown, Josh (2020) *Developing a persistent identifier roadmap for open access to UK research.* https://repository.jisc.ac.uk/7840/

[22] https://dev.rioxx.net/profiles/v3-0-beta-1/#rioxxterms:project

acknowledgements (req. 1,13). Practically speaking, these are highly pertinent to showing compliance with the policy.

Additionally, for repositories, we explored the coverage of additional PIDs, vital for linking AAMs to published works (namely ISSNs (req. 4) and DOIs (req. 5)). Analysis of coverage of these additional PIDs was not included in the scope of this exercise.

6.1 Date of deposit / licensing / embargo (req. 2)

The policy specifies that there should be no embargo on the availability of publications being run through route 2. There are several ways that compliance with the embargo clause may be partially monitored using existing metadata schemas. For instance, there are fields in the ALI schema for the start dates and end dates of licensing conditions. However, adoption of these components is currently extremely low. It is also possible, in principle, to compare deposit dates in repositories with publication dates of the version of record. Our partner, CORE, have assessed the feasibility of this latter approach with some success[23] although a number of data quality challenges exist.

According to CORE, deposit dates in repositories are unreliable as they are often reset, for example, when the repository software is updated or records are migrated between systems, or when metadata itself is updated. This is a critically important finding with relation to route 2 to compliance. If deposition rates are unreliable, it is not possible to assess whether an article was deposited at the same time as the version of record was made available. So there is currently no reliable way to assess whether content was deposited with zero embargo after publication. Resolving this issue is not straightforward, as it will involve the development and adoption of best practice approaches that ensure accurate recording of the date of deposition. This may require workflow changes at institutional repositories, technical fixes to various repository softwares and other systems that repositories integrate with, and community agreement on best practice.

6.2 Licensing data (req. 8)

Licensing data is of particular importance for the policy, as it sets out the conditions of access and re-use for the content covered by it.. Publications (or the repository version) must be licensed CC-BY, OGL, or CC-BY-ND with permission. Consistency in licensing metadata is a challenge that has received a lot of attention, with the National Information Standards Organization (NISO) releasing version 1 of the Access License Indicators (ALI) Schemas in 2015[24] (note that a new version of the ALI schema is being prepared and should be published in the near future). This schema component is a working solution for machine-readable licences. It is compatible with both CC and OGL licences, as well as many others, and with the schemas we reviewed here.

[23] https://scholarlycommunications.jiscinvolve.org/wp/2020/02/24/core-raises-repository-data-quality-by-consolidating-information-from-external-datasets/

[24] http://www.niso.org/schemas/ali/1.0

6.3 Self-archiving policy and preservation location (req. 10)

Documentation of preservation locations is not supported by Crossref and DataCite schemas and is not currently accepted best practice. This aspect of the policy could not be monitored through metadata without further schema development or a different approach. For repositories, similarly, it is not currently supported in any relevant schema.

We recommend a more efficient approach, relying on existing practices and workflows, in that publisher self-archiving policies be cross-checked through SHERPA RoMEO, as stated in the UKRI OA policy itself.

6.4 Open Citations (req. 11)

There is no field in any schema to indicate whether publications have their citations uploaded to the Open Citations database. Inclusion of such a field is not recommended as this information is not considered article-level metadata. In addition, changes to any given schema to incorporate this information would have to go through the usual process for expanding each schema. We recommend the more efficient approach would be to check the OpenCitations Corpus[25] directly to see if the citations from each publication have been included.

6.5 OpenDOAR registration status (req. 12)

OpenDOAR registrations status is not an included field in any of the schemas we looked at. Ultimately, this is not metadata about the outputs so much as the venue for their archiving or dissemination. We recommend creating a list of authorised repositories and performing a cross check using the OpenDOAR API.

6.6 Metadata originating from publishers (for route 1)

For route 1, the primary source of metadata is scholarly publishers. To assess the ability to use metadata originating from publishers to monitor compliance, we looked at Crossref metadata, with the assistance of COKI.

The dataset from COKI contained metadata from over 193,000 Crossref records that COKI were able to connect to UK institutions based on publisher-supplied affiliation data, over the period from 2011 to 2021. There are just over 14,000 records from 2011, rising to nearly 23,000 in 2020. There are fewer records associated with 2021, as the year is not complete. Analyses involving trends over time were, therefore, restricted to the period between 2011-2020 to avoid biases caused by journal publication schedules.

Figure 1 below shows the percentage population of relevant fields based on the Crossref data warehoused by COKI for selected fields. We analysed funder data (req. 1, 13), licences (req. 8), and ORCID (req. 6, 7). We have not included DOI (req. 5) or ISSN (req. 4) in this graph because in the dataset all records have both DOIs and ISSNs. This result may seem surprising until we consider that

[25] https://opencitations.net/corpus

DOIs are the primary key field of the Crossref dataset and that the source of this data are publisher journal workflows, so both identifiers are always present by default.

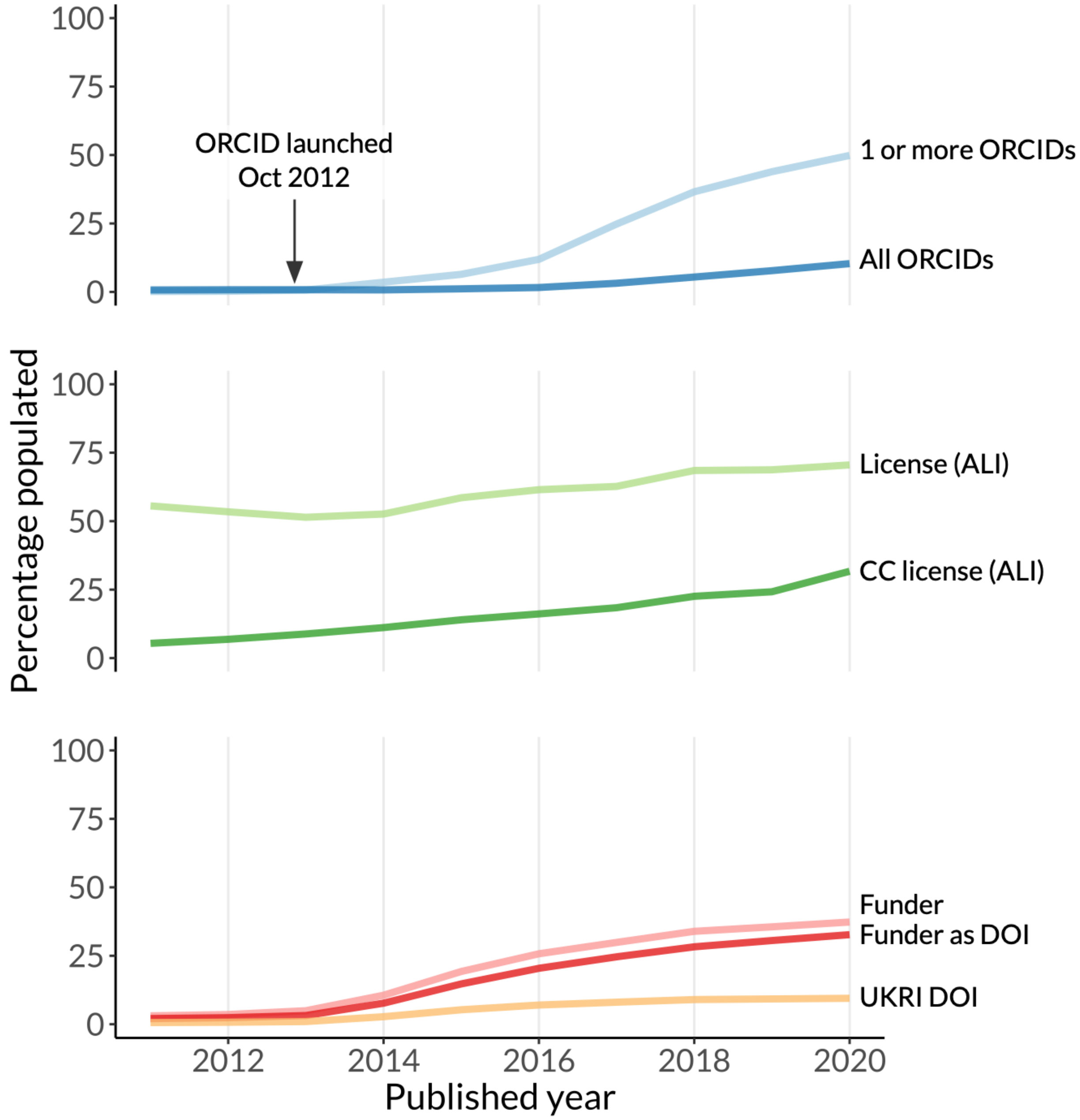


Figure 1: the percentage of Crossref records containing data for relevant fields that are defined in the schema. Raw data came from the Crossref API and was aggregated and processed by COKI. Summarisation performed by MoreBrains.

6.6.1 ORCID records (req. 6, 7)

The blue line shows ORCID record completion rates. The number of records with at least one ORCID has been rising steadily since the launch of the ORCID registry in October 2012, and has now reached 50%. The percentage of records where all authors have an ORCID (req. 6) is much lower, at around 10%, which is in line with anecdotal estimates of the number of single author articles. It is possible that most content containing all ORCIDs are single-author publications, although more work would be needed to confirm that finding. The number of authenticated ORCIDs (req. 7) is not plotted on this graph as there is a perfect equivalence between the number of ORCIDs and the number of authenticated ORCIDs — publishers do not submit ORCIDs in article metadata unless they state that they are authenticated. However, there are known issues with the veracity of claims about authentication of ORCIDs. An ORCID ID is considered to be authenticated if it is gathered through the single sign-on system that ORCID provides[26], however, not all publishers use the authentication pathway, allowing users', and particularly co-authors', ORCIDs to be manually entered. Work is needed to educate relevant stakeholders on the process of authentication and its importance. Publishers should be encouraged to improve their metadata workflows to accurately and consistently communicate the authentication status of ORCID IDs.

Our conversations with publishers suggest that workflows that integrate the collection of ORCIDs for the corresponding author are increasingly considered to be best practice, but workflows to include other authors have generally not been implemented. There are good reasons for this. Inclusion of all authors' ORCIDs, particularly through the authenticated pathway, creates extra administrative steps for both publishers and authors. Anecdotally, publishers report that corresponding authors act as the primary source of contact, with co-author engagement in the submission and publication process being low to non-existent. There are legitimate concerns that adding the extra step of requiring multiple co-authors to authenticate would represent an overly onerous workflow, particularly given that some authors may have changed jobs, or may not already have an ORCID. For outputs with multiple UKRI-funded authors, the time and costs involved in corresponding with each one to obtain their authenticated ORCID ID would be prohibitive.

6.6.2 Licensing data (req. 8)

The green lines show the percentage of licence fields that are correctly populated. According to COKI, these counts include only well-formed URLs. The dark green line shows the number of licence URLs that correspond to creative commons licences.

The number of records with well-formed URL-based licences is over 60%, however it is rising quite slowly. On the other hand, the number of CC licences is climbing at a faster rate, in keeping with the shift towards open access.

Positively, meta-analysis of the data by COKI shows that poorly-formed URLs are a negligible percentage of records, and that there are no instances of free-text in the licence URL field of the Crossref data. COKI filters out licence records that don't use the *<ali:license_ref>* tag as described in section 5, so it is possible that some publishers include free text in the licence metadata field without using this tag. That noted, it seems that quality control mechanisms for machine readable licences are working well.

6.6.3 Funding information (req. 1, 13)

The pink, red, and gold lines represent the percentage of records with funder information. In 2020, over 40% of records had some kind of funder information, with the vast majority being in the form of

[26] J. Brown, 'What's So Special About Signing In?', ORCID, Feb. 16, 2017. https://info.orcid.org/whats-so-special-about-signing-in/ (accessed Nov. 21, 2022).

a funder DOI. The proportion of records in which metadata is included for the specific award (req. 15) is almost the same as for the funder DOI, so are not shown in the plot. The reason why levels of population for these two metadata components are almost identical is unclear based on the data available, and would require analysis of greater depth than is possible here.

The number of entries with a UKRI funder DOI is approximately one quarter of the total number with a funder DOI, which is broadly in line with the proportion of UK research funded by UKRI.

As noted above, the Open Funder Registry is on course to be folded into the Research Organization Registry (ROR), so funder DOIs will be replaced by RORs. Adoption of DOIs for grants will also change how funding acknowledgements work, as a grant DOI will be globally unique (unlike many current grant numbers used) but will also be linked to the funder that registered it. This will mean that metadata that includes a grant DOI will link both the funder and the grant to the output.

6.6.4 Metadata completeness distribution across publishers

The publishing industry is diverse in terms of organisation size. Crossref has over ~14,000 organisational members, nearly 4,500 of which submitted at least one record in 2020, based on data from COKI. One the other hand, the distribution of numbers of publications is very concentrated. The four largest publishers (Elsevier, Springer Nature, Wiley, and Taylor & Francis) account for just over 50% of all records submitted to Crossref that were affiliated with UK institutions in 2020. The top 10 most prolific publishers, which account for over 70% of all records, include two institutional presses and a learned society, as described in Table 6.

Publisher	Metadata records submitted to Crossref in 2020 (% of total)
Elsevier	19.6%
Springer Nature	14.7%
Wiley	9.8%
Taylor and Francis	6.3%
Oxford University Press	4.9%
Institute of Electrical and Electronic Engineering	4.3%
Sage	3.3%
Cold Spring Harbour Laboratory	2.8%
British Medical Journal Group	2.8%
Multidisciplinary Digital Publishing Institute	2.3%
Total of top 10	**70.9%**

Table 6: The most prolific 10 publishing companies account for over 70% of all publications connected to a UK-based affiliation. Of these top 10, the four largest publishers are commercial organisations, followed by a diverse group of learned societies, institutional presses, and other commercial publishers.

Figure 2 shows the aggregate percentage of populated metadata fields for the 10 largest publishers compared to the 'long tail', which published the remaining 29% of UK content in 2020.

Despite the difference in scale between the largest publishers and the rest, the levels of metadata compliance are broadly comparable for most of the relevant fields, with the large publishers performing slightly better on most counts. The notable exception is licensing data, for which the largest publishers perform significantly better than the long tail. It is possible that this is due to the perceived strong commercial incentives among the large publishers to get licensing correct.

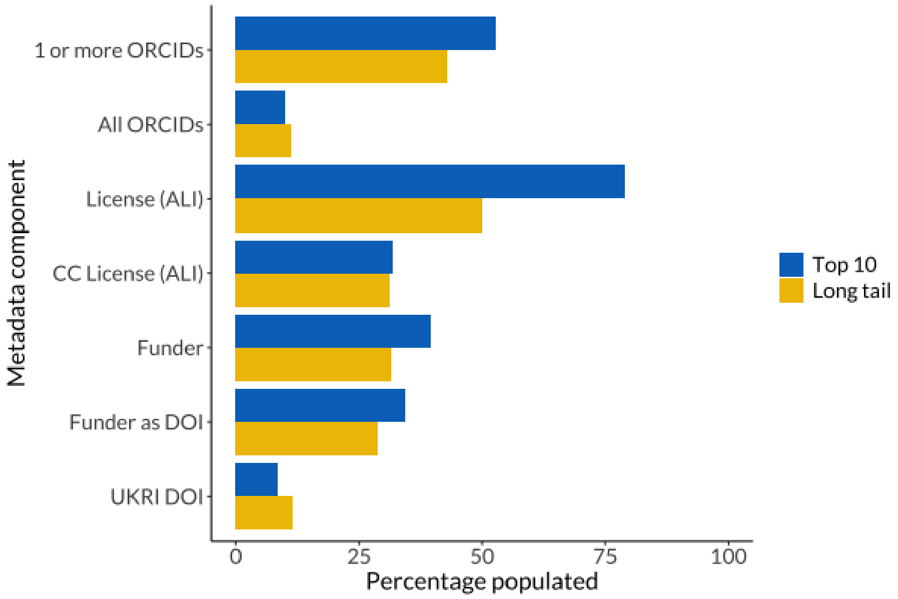


Figure 2: The percentage of Crossref records containing data for relevant fields that are defined in the schema for articles published in 2020 for all publishers. Separate bars are shown for the top 10 largest publishers by records submitted to Crossref in 2020 and remaining 2,282 metadata providers that we call the 'long tail'.

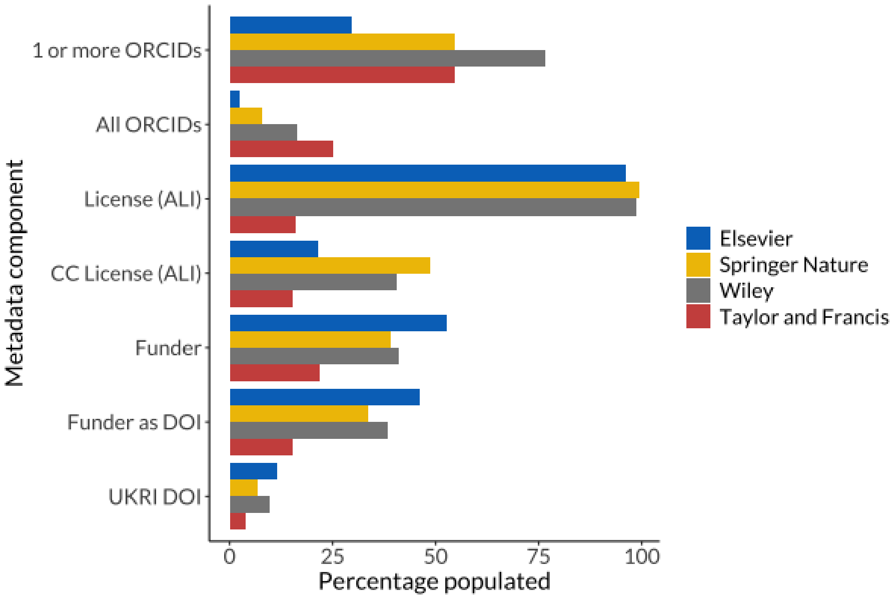


Figure 3: The proportion of completed relevant Crossref metadata fields for the largest four commercial publishers for articles published in 2020.

Metadata practice varies across publishing organisations for a variety of reasons: historical, disciplinary, financial, compliance, and so forth. Given the importance of the four largest publishers in terms of volume of UK output, we looked at the relative levels of metadata compliance for each of them. Figure 3 illustrates that across these four, metadata completeness varies. This could be due to a mix of factors, including historical custom, pace of adoption of digital workflows, acquisitions and mergers, and the presence of an analytics product within the same family of companies. Notably, Wiley performs particularly well on ORCIDs for at least one author. Springer Nature has a larger proportion of CC licences. Elsevier does well on funding information, which perhaps is a reflection of their interest in research management and analytics services like Scopus and Sci-Val.

Since the four largest publishers have similar levels of metadata completeness, and there is comparatively little variance between them and the long tail, it would be tempting to think that there is little variance in metadata completeness between publishers in general. This is not the case. Figure 4 breaks down metadata compliance rates for the second group of publishers: those ranked from five to 10 in terms of the number of UK articles in 2020.

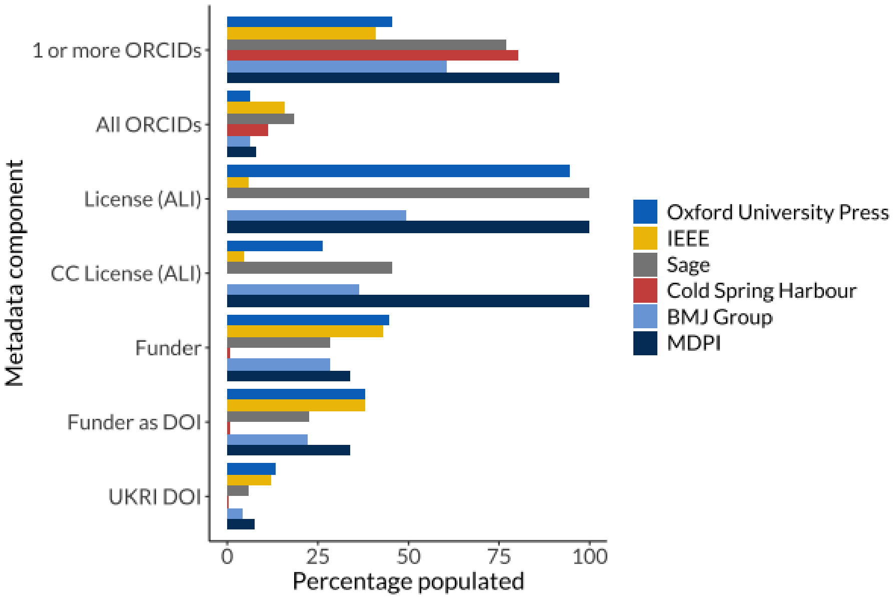


Figure 4: The same data but for publishers ranked six to ten by the total number of records in Crossref for 2020.

We see from figure 4 that there is significant variation in metadata completeness rates between the second group of publishers. In licensing data, for example, Sage and OUP have high rates of completeness, but the others provide very little licensing metadata.

These data, taken together, suggest that the levels of metadata completeness are variable across publishers. While larger commercial publishers seem to be doing better than most, our findings suggest that there is significant variation in levels of completeness and coverage within and between publishing organisations. Moreover, there is still plenty of room for improvement in metadata coverage to support efficient monitoring of policies like the UKRI open access policy, as exemplified by variations in funder acknowledgement, low levels of ORCID coverage, and the need for machine-readable licences for the 'long tail'.

The characteristics of the publishers that are least compliant are unclear. There may be correlations with geography, governance form, organisational structure, business model, or technical considerations like choice of platform or content management system. Finding correlations and segmenting publishers to identify characteristics of those most in need of support is beyond the scope of this investigation, but may be a good candidate for further investigation. Based on our current knowledge, a multi-pronged approach may be necessary to maximise metadata availability. Outreach and clarity may help the largest publishers increase their compliance rates, while for the long tail, a more in-depth analysis is needed to establish approaches and channels to identify the greatest opportunities for improvement.

6.7 Metadata originating from repositories

For route 2, the primary origin of metadata will be repositories. To assess the metadata landscape for repositories, we worked with CORE, which aggregates metadata from repositories globally and also acts as the UK national aggregator. CORE works closely with repositories to help them develop best practices, as well as with funders such as Research England, to help them assess compliance with the REF2021 open access policy[27]. CORE also aggregates metadata and full text for cross-repository discovery of open access content and to provide a machine-readable search interface.

6.7.1 Dublin Core (incl. OpenAIRE guidelines adherence), RIOXX

The repository metadata landscape is not as standardised as that for publisher metadata. As mentioned in section 5, the Dublin Core schema establishes basic building blocks but does not have specific guidelines for implementation for repositories. The OpenAIRE guidelines for Dublin Core and RIOXX (which draws from Dublin Core as well as a number of other schemas) are application profiles designed to create that structure as a basis for standards and best practices.

Of the 132 UK repositories that CORE analysed, 65 (49%) used the RIOXX application profile and 67 (51%) used Dublin Core. For those using Dublin Core and not RIOXX, use of the OpenAIRE guidelines is considered good practice, although there is no way to directly tell if a repository is doing so . Analysis of the metadata that CORE holds for non-RIOXX repositories proved challenging. Almost 2M such records were identified, with just over 162,000 being recognised as research outputs including both articles and long-form content. For 33 of the non-RIOXX repositories, the metadata provided to CORE were not sufficiently descriptive to indicate which records corresponded to research outputs.

All the non-RIOXX records had a field in their schema for related URLs, which could be used for funder (req. 1, 13), funding award (req. 15), or other identifier (req. 5), however, type tagging was not consistent. CORE's attempts to parse the metadata using the OpenAIRE guidelines yielded no results. Only 191 records had an identifiable ISSN (req. 4) and no ORCID IDs (req. 6, 7) could be extracted. On the other hand, 34% of records had an identifiable licence record.

Based on this first pass analysis by CORE, it appears that repositories in the UK that do not use the RIOXX application profile are currently not providing sufficiently rich information for straightforward metadata monitoring of UKRI OA policy requirements.

For repositories that use the RIOXX application profile, the situation is more positive. While, as section 5 makes clear, the RIOXX standard does not cover as broad a range of relevant pieces of information as the Crossref schema, CORE were able to perform a number of metadata completeness analyses.

[27] https://www.ref.ac.uk/media/1228/open_access_summary__v1_0.pdf

6.7.2 Distribution of records by repository

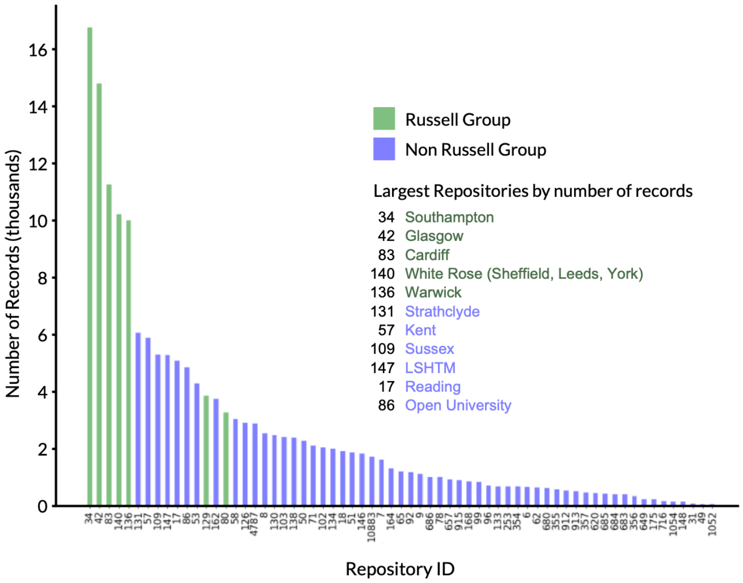


Figure 5: Distribution of records from the largest to the smallest repositories. Those coloured green are Russell Group Universities.

The level of metadata quality provided by repositories varies considerably across HEIs. An in-depth analysis of variations within peer or mission groups is outside the scope of this study but, for illustrative purposes, we highlight Russell Group members as an example of the challenges of identifying patterns in repository coverage. A more detailed study could explore other demographic clusters such as TRAC peer groups, although institutional culture, policy, and systems-adoption factors will likely render comparisons problematic. Figure 5 shows the number of records found in the CORE database associated with each institution that uses the RIOXX application profile. The five repositories with the most records are all Russell Group institutions; Southampton and Glasgow each have more than 14,000 records.

However, of the 24 Russell Group universities, only seven have adopted RIOXX for their repository application profile. It appears that there is no correlation between levels of research intensiveness and institutional strategic investments in repository standards, etc.

6.7.3 DOIs (req 5)

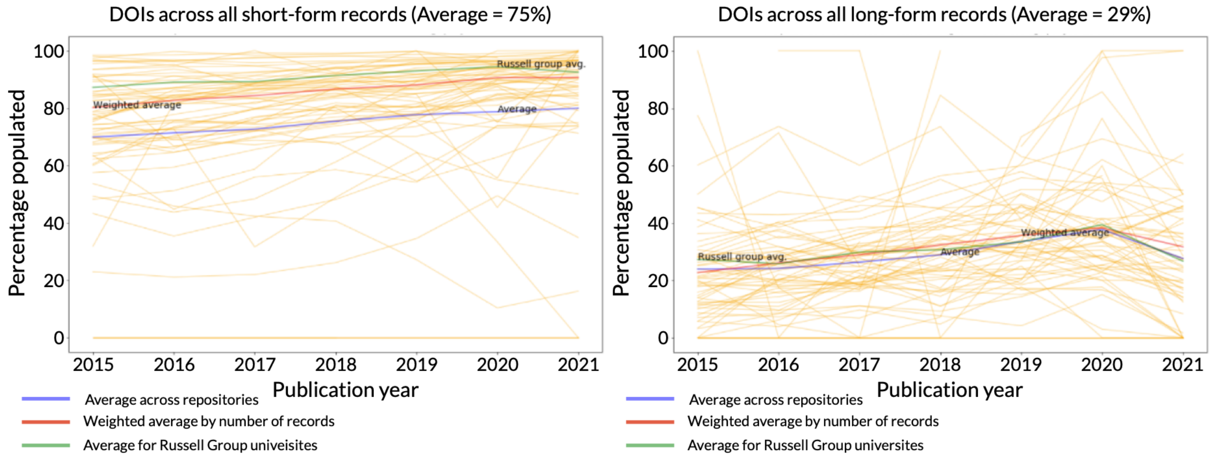


Figure 6: Percentage of DOI fields populated for short-form (research articles, conference proceedings, and reviews) and long-form (books, book chapters and monographs) content in RIOXX repositories in the UK. The blue line represents the average across repositories, the red line is the average weighted by the number of records and the green line is the average across Russell Group universities.

Inclusion of DOIs in repository records is important for monitoring compliance with the policy. Correctly identifying metadata allows the cross-referencing of the object in the repository with any version of record in a publisher's website. This is needed to check embargo periods and/or other technical requirements that may be added in the future for compliance route 2.

As shown in figure 6, DOI completeness for short-form or article content is relatively good, with an average across repositories of 75%. DOI metadata completeness is trending upwards almost across the board, although there are still some repositories that have very few or no DOIs, as shown by the density of orange lines in the plot.

For long-form content, the level of DOI completeness is much lower, at an average of 29% across RIOXX repositories. There are a number of factors that could explain this difference. Books are a more common form of output in the humanities and social sciences, where research metadata and bibliometrics are not as much a part of research culture as in the STEM subjects[28]. Also, many book publishers still don't register DOIs for books or book chapters.

A more in-depth study is required to investigate the best ways to ensure identifiability and discoverability of long-form content.

6.7.4 ORCID records (reqs 6,7)

ORCID IDs are a vital metadata element for understanding which people are involved in creating a piece of work, who wrote it, and who is responsible for maintaining it.

[28] Franssen, Thomas, and Paul Wouters. 'Science and Its Significant Other: Representing the Humanities in Bibliometric Scholarship'. Journal of the Association for Information Science and Technology 70, no. 10 (2019): 1124–37. https://doi.org/10.1002/asi.24206.

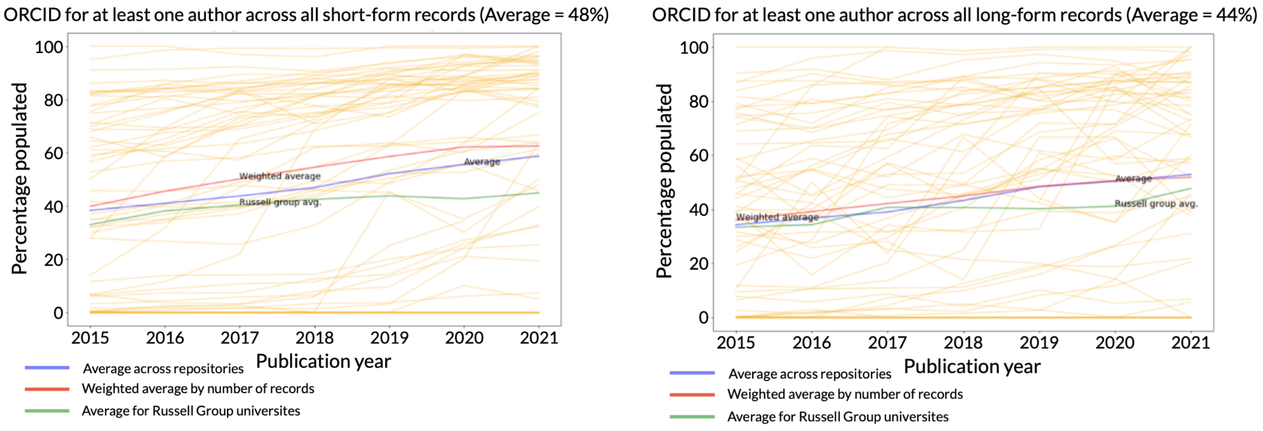


Figure 7: Percentage of short-form (research articles, conference proceedings, and reviews) and long-form (books, book chapters and monographs) content in RIOXX repositories in the UK that have at least one ORCID ID. The colour coding of the lines is the same as for figure 6.

Of short-form works, 48% of records have at least one ORCID ID associated with them (figure 7). As can be seen from the densities of orange lines, there is a general upward trend among many repositories, although levels of compliance are highly variable, with some still having no identifiable ORCID IDs. Interestingly, Russell Group university repositories generally underperform on this metric of metadata completeness compared with RIOXX repositories overall.

For long-form, the average across all repositories for at least one ORCID is 44%, only slightly lower than the average for short-form. The trends between the two content forms are broadly similar; there is a generally upward trend, with Russell Group universities having a slightly lower level than average.

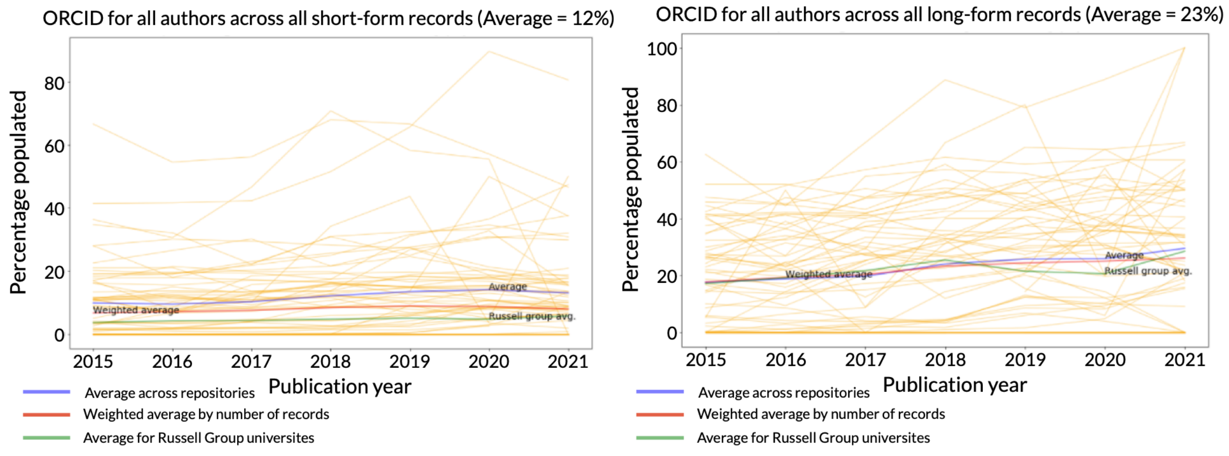


Figure 8: Percentage of short-form (research articles, conference proceedings, and reviews) and long-form (books, book chapters, and monographs) content in RIOXX repositories in the UK that have ORCID IDs for all authors. The colour coding of the lines is the same as for figure 6.

The percentage of records that have ORCID IDs for all authors is inevitably lower than the percentage with at least one ORCID ID. Figure 8 shows that adoption by this measure is fairly low (12% of records) and is climbing very slowly.

Institutional repositories face a different challenge to publishers in reaching full ORCID adoption. Where all authors of an output are at the same institution as the repository, the data protection challenges associated with authenticating ORCIDs are far less. Once a researcher has authenticated their ORCID with the institution, it should be possible to associate it seamlessly, with no need to obtain further permissions from the authors. On the other hand, such automation likely requires technical integration with other university systems and software. Anecdotally, integrations between institutional research management and repository systems are not of consistent quality across the HEI sector, indicating the need for investment. For outputs with authors spread across multiple institutions, there are other challenges. Repository managers have no ability to compel collaborating authors to visit their repository sites to enter their ORCIDs, a situation which is made more challenging if outputs are held at multiple repositories. As explained in section 7.1.3 , publishers routinely collect authenticated ORCID IDs only from the corresponding or lead author, so this will be the only ORCID sent through to the repository via the Publications Router, for instance.

For long-form content, the percentage of repository records where all authorship assertions have an ORCID is higher than for articles, at 23%. The distribution of orange lines, which represent individual repositories, is also more variable. It is tempting to think that for long-form content, like books and monographs, more progress has been made assigning ORCIDs to all co-authors. However, it is also the case that books and monographs tend to have fewer authors.

6.7.5 Licensing data (req 8)

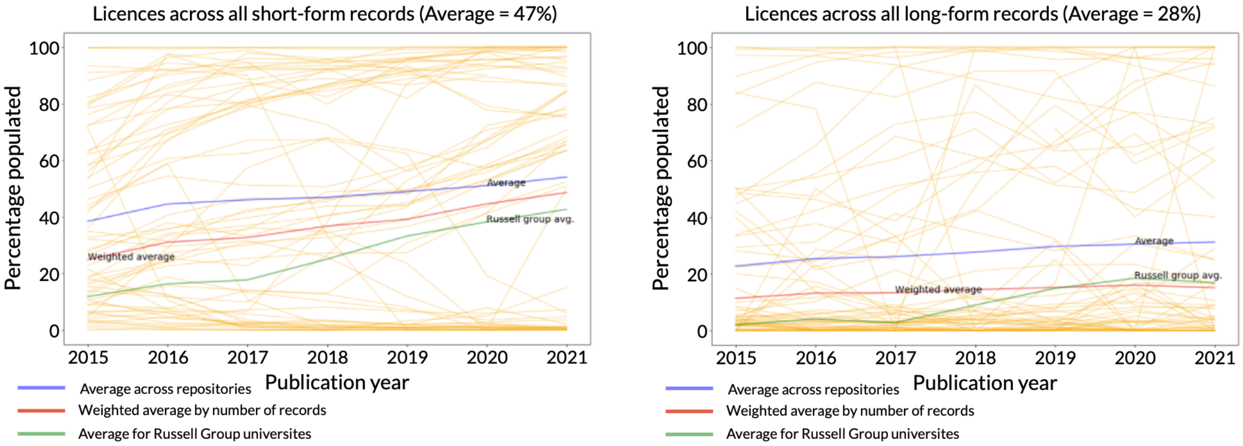


Figure 9: Percentage of licence fields populated for short-form (research articles, conference proceedings, and reviews) and long-form (books, book chapters, and monographs) content in RIOXX repositories in the UK. The colour coding of the lines is the same as for figure 6.

The average percentage of fields with licensing data is 47% across RIOXX repositories. The red line in figure 9 shows that the average when weighted by number of records is lower than the unweighted average (blue line), suggesting that smaller repositories do a better job of including licensing information than larger ones, in general. Similarly, repositories at Russell Group universities have lower levels of licence metadata completeness on average.

Again, care is needed when interpreting this data and further investigation is required to identify more granular trends. The density of orange lines gives a sense of the distribution across repositories. We can see from figure 9 that the densest parts of the graph are at very low and very high percentages. Repositories tend to therefore be either good at recording licensing information or poor, with few falling in between. The reasons for this split are unclear, but it may represent a distinction between mediated and unmediated deposit, or the fact that adding licensing information is sometimes part of a formal deposition workflow carried out by either researchers, an administrator, or a repository manager. This hypothesis requires further investigation.

Licences for long-form content are less frequent overall than those for articles (28%). Interestingly, Russell Group university repositories do comparatively well on this metric, though it is unclear why.

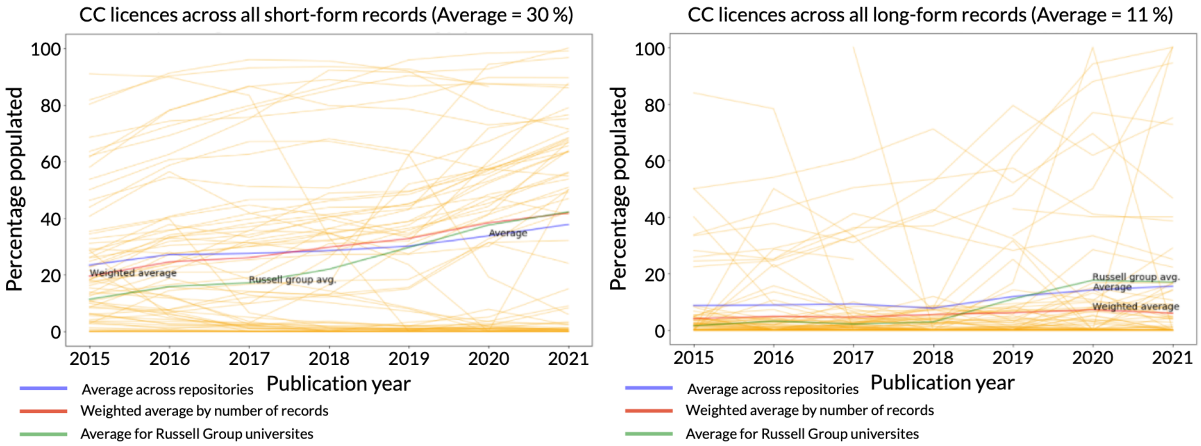


Figure 10: Percentage of licence fields populated and indicating CC licences, for short-form (research articles, conference proceedings, and reviews) content in RIOXX repositories in the UK. The colour coding of the lines is the same as for figure 6.

It would be reasonable to expect that the majority of content in institutional repositories should be under some form of open licence, given that they are often used as a means to establish open access policy compliance through route 2.

Figure 10 shows that the number of recognisable Creative Commons licences found in the RIOXX repository data is 30% on average across repositories. The red, green, and blue lines show that, on average, Russell Group university repositories are trending the same way as the broader community. The density of orange lines shows a generally orderly upward trend across repositories, although there remains a population with very few or no identifiable Creative Commons licences. The gradual change that is happening within many institutions (as opposed to the all-or-nothing appearance of general licensing metadata compliance) may imply that a steadily increasing number of individual authors are making a positive choice to publish using a Creative Commons licence and to enter that information into the repository themselves.

6.7.6 Funding information (reqs 1,13)

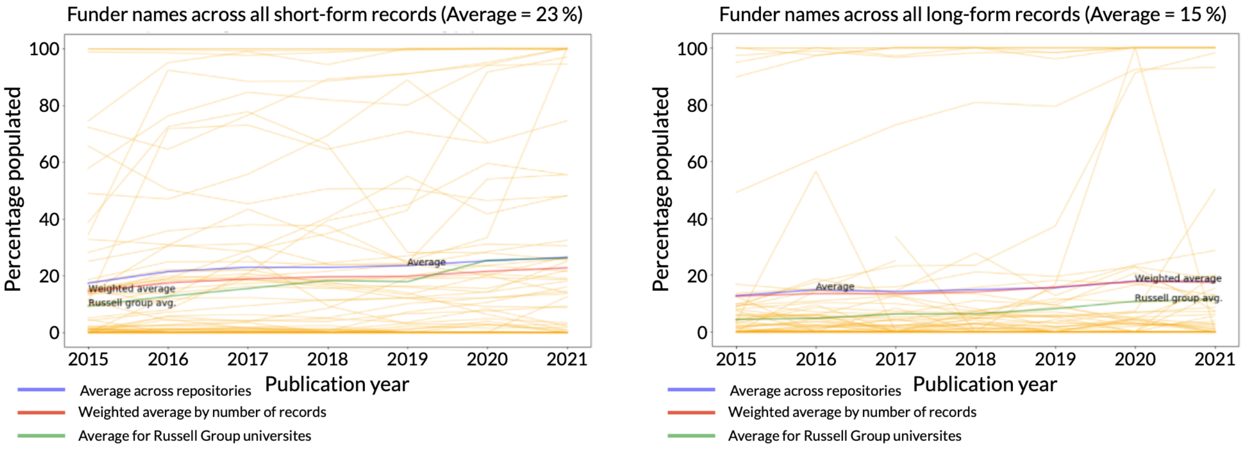


Figure 11: Percentage of funder name fields populated for short-form (research articles, conference proceedings, and reviews) and long-form (books, book chapters, and monographs) content in RIOXX repositories in the UK. The colour coding of the lines is the same as for figure 6.

Funding information in repository metadata is a policy requirement, because it is necessary for correct attribution and acknowledgement of UKRI funding of projects.

Figure 11 shows that, while there are a small number of repositories that have very strong funder name metadata compliance, most repositories have poor levels of metadata population for this field, as indicated by the distribution of orange lines.

The averages are all trending upwards, although at a very modest rate. As with ORCID IDs, funder data will sometimes be held by the universities' research management function, often in their research information management system. Again, increased compliance in this area could be driven by investment in, and support for, technical integrations.

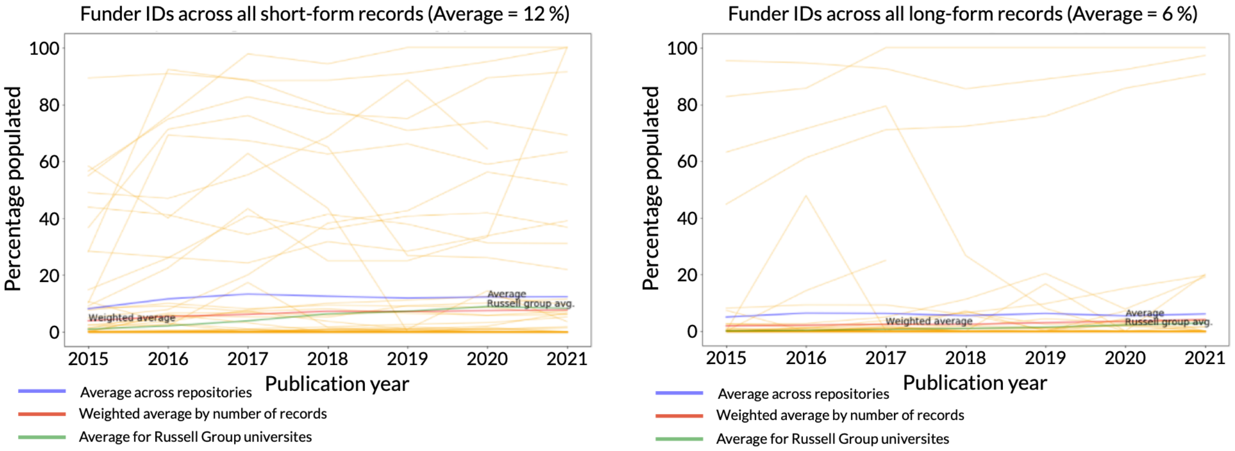


Figure 12: Percentage of funder ID fields populated for short-form (research articles, conference proceedings, and reviews) and long-form (books, book chapters, and monographs) content in RIOXX repositories in the UK. The colour coding of the lines is the same as for figure 6.

Funder names in metadata records can help identify funding sources and meet the requirements of the policy. However, applying funder IDs is a more robust mechanism that would lead to greater accuracy and less confusion in funder attribution. The use of a persistent ID, like the Crossref funder ID prevents many errors due to naming variations and simple typos.

Figure 12 shows the proportion of records that have funder IDs. The trends are similar to those identified for funder names but the percentages are significantly lower (12% for short-form and 6% for long-form). Funder identifiers are comparatively new compared to DOIs for outputs and ORCID IDs. Time and concerted effort are required to drive adoption of funder identifiers by funders, institutions, and repositories.

7 Organisational and capability challenges

7.1 Interview findings

Key findings arising from the interviews fell into four main types: general (high-level, expressed by a number of interviewees), service providers, repositories, and publishers.

7.1.1 General

The policy is generally being well received by the research-supporting communities. There is a sense that UKRI has prepared the industry well with messaging and consultations, and is exhibiting a willingness to meet with potentially affected stakeholders. Consequently, the interviewees were pleased to have been invited to participate in this project, showed a high level of awareness of the policy, and were forthcoming in their responses. There were reports, however, that researchers themselves have mixed feelings about how the policy might affect their working practices, with their enthusiasm correlating with their career stage (typically, the earlier the career stage, the more enthused the researcher).

Almost all interviewees asked for practical instructions on how to comply with the policy. Requests included: a frequently asked questions collection; advice on which schemas to use; and on which fields need to be populated, by whom and at what point in the publication process.

Some broader, perhaps longer-term, questions were also raised, such as:

- What will ‘compliance’ look like? (The chief concern here is with the ‘all author ORCID IDs’ requirement, which is currently agreed to be too difficult)
- How should international stakeholders engage with a nation-specific policy (in terms of information flow, international metadata and other standards, international publications)?

7.1.2 Service providers[29]

This group welcomed the timing of the policy announcement, which allows some time to prepare before staff capacity and attention is consumed by the next REF exercise. Currently, UK institutions are using a variety of tools and platforms to track outputs and activities, including Eprints, HAPLO,

[29] This group included organisations such as Jisc, repository aggregation and platform providers, and Crossref.

RIOXX, DSpace, Unpaywall, CORE Repository Dashboard, PURE, Symplectic, Converis, OAI-PMH, ResourceSync, and Researchfish. This means that the pathways to meeting policy requirements are diverse, contain various pros and cons with respect to accommodating the new policy, and may need some specific consultation with groups like CoSector and ARMA to support them - particularly the less research-intensive institutions.

One interviewee mentioned feedback from an informal post-REF poll of ORCID community member institutions, which indicated that research managers only pay attention to mandates, recognising them as the strongest requirement signals from the policy-makers. If 'recommended' activities tend to be ignored then, the UKRI OA policy should require clear mandates to ensure change in the metadata landscape, which would provide the most useful support for stakeholder organisations.

There are already a number of partially effective tools that can be used to support compliance and reporting for open access. However, there are complexities involved no matter which tools are chosen to respond to the UKRI policy, and providers are keen to build some specific recommendations for which systems should be used for what pieces of the workflow, and by whom. These may also depend upon differing levels of research intensity and discipline requirements across the sector. Plug-ins were suggested as a partial solution, but repositories will still need help implementing these, which will necessitate some community work and possible investment. There is also a separate quality management issue (see later in this section), which is beyond the capacity of tools or platforms to remedy.

One key consideration is how to balance the relative capabilities and governance structures of OpenAIRE and RIOXX. Although the former was primarily developed to support reporting on European Commission-funded outputs and the latter to support RCUK),they can be mapped to one another to enable data structured using one profile to be 'translated' correctly into the other. Such mappings will need to be kept up to date. Furthermore, metadata itself has a general governance and ownership issue (standards help the exchange of data, but do not guarantee accuracy or completeness, and there are many 'sources of truth' for the actual claims made about an article) that needs to be addressed in order to build trust and ensure the long-term success of the policy.

According to some interviewees, the Jisc Publications Router has facilitated the implementation of the RCUK and previous REF open access requirements by encouraging publishers to send specific, relevant information for the purpose of populating repositories. To maximise its efficacy now, there should be a specific audit of its capabilities and interactions with other systems. Some particular points of clarification that our interviewees noted are:

- Which dates are needed?
- How should the different licensing situations be expressed? (For example, the Green open access route for the accepted manuscript is not well expressed by publishers)
- What happens when embargoed material is de-embargoed?
- Will there be a review of initial compliance and non-compliance; not just the levels but also how non-compliance is happening (technical, publisher, loopholes by researchers, etc)?
- How is versioning to be handled?

As an international organisation, Crossref would be happy to engage with its members and work with UKRI with some supportive messaging. Although Crossref would need to decline any UK specific requirements with respect to its schema, it is interested in framing this as an opportunity to improve licence data and would be willing to do some community work to encourage the supply of better quality metadata.

They are also willing to work with the JATS4R team to expedite recommended updates to their advice. In terms of the roll-out of the policy, Crossref suggested that the guidance could include advice on maintaining and updating the metadata, as policies and urls can change over time. More generally, they would also welcome any additional emphasis on including information about robust preservation programmes in the metadata.

JATS4R can help with developing standards and recommendations based on known use cases.[30] However, the JATS4R interviewee warned that requiring too much data that cannot be complied with - specifically collecting all author ORCID IDs for each output - could result in bad information being entered into the system. Like Crossref, JATS is a global initiative, so a single country would be unable to unduly influence its decisions. However, the UKRI policy is a legitimate locus for describing use cases on which to base standards and recommendations. The JATS4R interviewee noted that the policy does not include mention of XML, which in their opinion is key to making metadata machine-readable. In short, 'if it ain't readable, it won't be found'.

7.1.3 Publishers

Publisher interviewees were aware that members of their community would likely experience the policy differently depending upon their size, subject areas, published content types, and comparative exposure to the UK market. There is a range of preparedness and a disparity of resources. Learned societies - whether self-publishing or contracting out to commercial partners - are already challenged and this may increase the pressure. There is a lack of clarity on how transformative agreements will work in relation to the licensing requirements, which points to a more general issue of complexity within a system. Anything that can be done to systematise, automate, or otherwise clarify the requirements would be well received. Specifically, this could include increased use of persistent identifiers, and ongoing community work to build best practices, standards, and knowledge.

7.1.4 Repositories

The repository interviewees admitted that their sector is not well set up to cope with the requirements of the new policy for a range of reasons. Changes may need to be made to RIOXX v3 to enable users to comprehensively report compliance across all the UKRI policy requirements, which would take around six months (if the current review cycle can be leveraged). Many repositories would prefer a patch or plug-in to be able to implement the new RIOXX version. They also understand the need to do some work in order to provide the funding acknowledgement, as currently this is not being routinely added to the repository records, and RIOXX isn't expressing this information in a useful way.

Institutions relying on Current Research Information Systems (CRISs) are likely to experience particular problems as CRISs can be extremely slow to deliver updates. They are dependent on their vendor, who may not be wholeheartedly enthusiastic about the requirement if it relates only to a

[30] The process involves pulling together a group of experts plus the JATS representatives who then submit a recommendation for review. This should be based on what people are already doing.

small percentage of its international customer base, or does not align with its own investment priorities. This could delay timely implementation and could signal the need for a workaround for some institutions as well as dialogue with the vendors themselves.

There was general agreement that the more the system can be automated, the better it will work. There could be some issues with handling the metadata itself as some metadata records are still being treated as a proprietary resource by commercial entities in parts of the scholarly communications cycle, resulting in the question: at what point in the process does the CC-0 licence for the metadata kick in?

7.2 Additional considerations

As well as the key findings required for the scope of this project, some other questions and comments emerged that are worth capturing as, although outside the scope of this report, they may form part of ongoing dialogue with stakeholders going forward:

- Questions about the costs of preparing for and delivering the policy - does anyone know how much any of this will cost?
- It was suggested that the community should be as fully involved as possible — while this could result in slower progress initially, there would be better long-term results for the policy
- How is measuring the success of rights retention-policies — known to be of interest to UKRI and other funders — being discussed with other Plan S funders?
- How do preprints fit into the policy, can there be some clarification on this?

8 Concluding remarks

> *"Everything we're talking about, including using aggregators as correctional facilities, relies on the standard that's transferring this data to be as good as it can be to express all data, as CORE (or someone like CORE) won't be able to correct anything if they're harvesting data that isn't complete in the first place and that's going to be frustrating for everyone involved. Because it's already frustrating when they're harvesting it and then being told to correct data that's already been corrected somewhere else."*
>
> \- *Repository service provider*

Currently, a great deal of effort is deployed in filling gaps in the metadata record. Companies, products, and services have grown up to correct omissions or to remedy data loss. As this quote from one of our interviewees exemplifies, there are challenges in bringing together metadata from varying sources which may be incomplete or follow conflicting standards. There needs to be a two-way flow of information so that corrections made at any stage in the process of reporting compliance can be passed back upstream and corrected at source. This is inevitably a more complex process than simple aggregation, and is currently an ideal rather than a strategic objective.

Luckily, there is no need for UKRI to 'boil the ocean' to address these challenges. A relatively short set of metadata fields are required to show compliance. Many of these are already supported by relevant standards or application profiles. There are enthusiastic and committed communities of practice who will be able to remedy the gaps that have been identified, and to help UKRI communicate the need to adopt these updated standards to the wider community. Helping organisations and systems to be ready to comply is achievable.

In this section, we summarise the major challenges that have emerged during our research and analysis, and indicate potential remedies and solutions.

8.1 Challenges arising

For publishers, extensions to JATS are not structurally major. Project IDs need to be added to JATS and Crossref. Preservation location, self-archiving policies, and I4OC participation are also missing, but are not really article-level features, and can be verified at the journal or publisher level instead, so adding them to JATS may not be the optimal path.

Engagement with JATS4R is the best place to start. Aspects of the policy could be better tackled with guidance and recommendations on how to use the existing standards rather than by updating the standard itself. This is a well-defined process, which JATS4R is best positioned to facilitate. With an efficiently focused 'ask' and expert community support, this could shorten the timeline to remediation dramatically.

For repositories, there are two distinct sets of challenges. RIOXX needs to be extended to include data availability statements and the authentication status of ORCID IDs, and to support a wider range of PIDs for projects, grants etc. Given that RIOXX v3 is under active development, this is an opportune time to make changes to the application profile.

The second challenge is far greater. After more than a decade of work, only half of the repositories aggregated by CORE are RIOXX-compliant. Increasing adoption levels is likely to take longer than adapting the underlying metadata structures. It is beyond the scope of this report to provide a detailed analysis of the situation, but challenges are likely to include: different platforms; the range of versions of those platforms being used; and the variability in support for technical updates (some have local implementations that are supported by internal staff resource, others are cloud-hosted and supported by a service provider under a stringent service level agreement, some are proprietary, some rely on global volunteer communities who may not see the UKRI policy as a priority, and so on). All these variables lead to the conclusion that closing the gaps could be a protracted process.

In terms of the content of the required fields, work may be needed to help repository and publisher platforms make best use of the PIDs that the policy calls for, especially grant IDs. We note that Crossref has plans to enable more links between grant IDs and various content types, which may add value to the use and integration of these PIDs in more systems. For licences, we can see that there has been poor adoption of the ALI standard across the network, so this is a critical area where guidance and good practice should be extended.

Respondents across the board stated that the requirement for ORCID IDs for all authors and contributors was a problem. Publishers often only engage directly with the corresponding author, and there is no mechanism in place for gathering authenticated ORCID IDs from large groups of named contributors. Authentication of an ORCID ID demands that the owner of the ID signs in to their ORCID account from whatever platform they are using to submit an article and then approves the connection of their ID to the work in question. Corralling large, dispersed groups of authors, many of whom will not have been directly involved in drafting or submitting an article, is seen as an impossible, time-consuming task. To date, we have found indications from just two publishers that they are requesting ORCID IDs for all authors on papers submitted to their journals, the Journal of Medical Internet Research, and ScienceOpen.[31]

Repository managers and service providers all indicated that they have even less consistent contact with authors than publishers do, and no power to compel authors employed at other institutions, or in other countries, to visit the repository site to add their ID, let alone to do this for the multiple repositories that may house a copy of the article. There was a general assumption that repositories would be able to harvest a full set of authenticated ORCID IDs from publishers. This assumption, as noted above, is mistaken.

Many repositories do not create PIDs for the AAMs and other versions of content contained within them (there is a related issue of different PIDs being used for what may be essentially the same content: the AAM, of which there may be a number of versions, and the published version). There is a current NISO working group looking at article versioning that may help with this issue.[32] More than 120 UK repositories have access to DataCite membership services via the British Library-supported UK DataCite consortium, but they primarily use this to obtain DOIs for data rather than the other types of content DataCite currently supports. If DataCite extended their schema alongside RIOXX (many of the same additions are required) they could provide a useful additional pathway to demonstrating compliance, with the additional benefit of plugging UK repository data into a global discovery network with the ability to gather authenticated ORCID IDs to serve the DataCite-ORCID auto-update service. Some repositories use OAI identifiers for a range of objects including AAMs, however, OAI IDs do not have an associated metadata schema, so their use does not address the central challenge of metadata availability or interoperability.

In terms of the potential burden on the community of the policy, some specific challenges emerged during our interviews. Some learned societies are already struggling, and the policy represents an additional burden which they lack capacity to address. As noted throughout this report, Crossref is key to supporting the publishing community in compliance and reporting, but they also have serious capacity constraints,and are struggling to make changes to their formal XML document structure (also known as ‘Document Type Definition’ or DTD).

In addition, the cost of complying will fall disproportionately on less research-intensive institutions and small publishers, for whom the volumes of activity reported will be low and, consequently, the level of investment needed to support compliance may be difficult to justify. Larger organisations, with a substantial research income, are much more likely to have advanced research management systems or software available to them to help with policy compliance and reporting, although our research shows that metadata quality in these systems is not currently good enough. The OA Switchboard[33] will make it easier to engage some small publishers, and may also be useful for organisations in tracking publications and eligibility checks.

[31] Meadows, Alice; Haak, Laurel (2017): ORCID Open Letter - One Year On Report. ORCID. Journal contribution. https://doi.org/10.6084/m9.figshare.4828312.v1
[32] http://www.niso.org/standards-committees/jav-revision
[33] https://www.oaswitchboard.org/

There is a cultural issue amongst certain publishers and information providers, for whom, according to the publisher interviewees, in-house generated and cleaned metadata is seen as content in its own right, rather than as an open resource. A related issue is the risk that proprietary systems are employed to fill gaps in metadata coverage post hoc by investing in 'black box' solutions that rely on machine learning and AI to fill in the gaps. Such services are not reproducible or scrutinised, and tend to create lock-in to specific companies or platforms.

8.2 Remedies

There are clear processes for the remediation of challenges and gaps, such as working with JATS4R and Crossref for publisher data to support route 1 compliance; and engaging the RIOXX committee in UKCoRR to extend and enhance the RIOXX application profile. A campaign to support open source journal and repository platforms in implementing the technical underpinnings of these systems will accelerate adoption and roll-out, although it is impossible to predict when (or if) 100% coverage will be achieved. However, a programme of support for these activities will lower costs and improve efficiencies across the board. It will also cement the goodwill that the policy-making process has engendered.

We recommend that UKRI focuses on the specific, and reasonably short, list of fields required to be technically capable of showing compliance with the policy. The repository and publisher communities should be supported with additional consultation related to compliance implementation and information needs. Detailed breakdowns of the practical implications of pathways to compliance were suggested by several respondents as a much sought first step in this direction, as this would help the community to focus their efforts.

Guidance notes for different stakeholders should be prepared, based on an assessment of who will be doing what (from standards bodies to service providers to researchers).

Technically, the policy calls for OpenAIRE guidelines or RIOXX compliance. Our analysis of the available metadata and the challenges in analysing information at scale suggest that it may be wise to focus on a revised RIOXX profile as an application profile for UK universities. With support for its adoption, and updated and extended crosswalks, it could then be used to support OpenAIRE compliance as well.

There is also a need to use one or more metadata registries for aggregating UK repository metadata. CORE would be well placed to do this, if RIOXX were adopted consistently. DataCite could also serve that purpose if there are well-described crosswalks from RIOXX to DataCite, enabling it to function in an equivalent way to Crossref for publishers. The levels of coverage and adoption offered by the current consortium would also need to be extended. The use of tools to help institutions identify issues in metadata, fix them, and monitor their compliance (such as the CORE Repository Dashboard) should be encouraged as a mechanism to improve metadata quality across repositories.

There is a range of measures that UKRI and the community can take in the short term (such as updating standards guidance or adjusting application profiles); the medium term (developing specifications for RIOXX implementation, or publishing and updating metadata alongside grant IDs); and the longer term (driving up PID adoption and coverage for emerging PIDs such as projects and grants).

Together, these interventions and collaborations suggest that, with the right communication, investment, and guidance, and a solid programme setting out a roadmap of practical support for efficient, affordable policy compliance reporting, UKRI is in position to leverage community goodwill and appetite for change in scholarly communications. Pragmatism around timelines for compliance, and recognition of the challenges posed by specific demands (such as that for ORCID IDs for all authors and contributors) can be built on the evidence emerging from these collaborations, and help to ensure that obstacles to affordable, automatic reporting and compliance checking are eliminated as early as possible.

Appendix A - schema review

The following sources of documentation were used to identify schema components and tags.

Route 1 (Publisher metadata)	NLM-JATS	https://jats.nlm.nih.gov/publishing/tag-library/1.3/
	Crossref	https://www.crossref.org/documentation/schema-library/
Route 2 (Repository metadata)	DataCite	https://schema.datacite.org/meta/kernel-4.4/doc/DataCite-MetadataKernel_v4.4.pdf
	RIOXX3	https://www.rioxx.net/profiles/v3-0-beta-1/
	OpenAIRE	https://openaire-guidelines-for-literature-repository-managers.readthedocs.io/en/v4.0.0/application_profile.html

The table on the next two pages shows the tags for each metadata item of interest for the four metadata schema and application profiles that we analysed. When no tag was available according to the documentation, the cell is coloured light red to indicate a gap in the metadata standard with respect to the policy requirements.

Requirement	Content type	for Publishers	for Repositories	JATS	Crossref	DataCite	RIOXX	OpenAIRE
Acknowledge UKRI funding	Articles / long-form	Mandatory	Mandatory	<funding-source>	<funding-source>	<funderName>	<funder_name>/<funder_id>	<oaire:funderName>
Date of deposit	Articles	NA	Mandatory			<date datetype=”Available”>	<dcterms:dataAccepted> &<rioxxterms:publication_date>*	<datacite: date type="X">
Data Access Statement	Articles	Mandatory	Mandatory	<sec sec-type="data-availability">	<sec sec-type="data-availability">			
ISSN	Articles	Mandatory	Mandatory	<issn>	<issn>	<relatedIdentifier relatedIdentifierType="ISSN”>	<dc:source>	<datacite:relatedIdentifier relatedIdentifierType="ISSN">
Article level PID (DOI, URN, Handle)	Articles	Mandatory	Mandatory	<article-id pub-id-type="doi">	<DOI>	<identifier identifierType="DOI">	<dc:identifier>	<datacite:alternateIdentifier alternateIdentifierType="DOI”>
ORCID ID (all authors)	Articles	Mandatory	Mandatory	<contrib-id contrib-id-type="orcid">	<ORCID>	<nameIdentifier schemeURI=”...” nameIdentifierScheme="ORCID">	<rioxxterms:author **id="URL"**>	<datacite:nameIdentifier>
Is ORCID ID authenticated	Articles	Encouraged	Mandatory	<contrib-id contrib-id-type="orcid" **authenticated="true"**>	<ORCID **authenticated="true"**>			
Licence (non-proprietary format)	Articles / long-form	Mandatory	Mandatory	<ali:license_ref>	<license>	<rights xml:lang="en-US" schemeURI="..." rightsIdentifierScheme=... rightsIdentifier=... rightsURI="..."/>	<ali:license_ref>	<datacite:rights>
Preservation location (Portico etc)	Articles	Mandatory	NA		<archive name="X"/>	<contributor contributorType="distributor">		
Self-archiving policy (registered in	Articles	Mandatory	NA					

Sherpa-Romeo)								
Citation data in I4OC (http://opencitations.net/)	Articles	Mandatory	NA					
PID for funders	Articles	Encouraged	Encouraged	<institution>	<funderef_data>...<assertion name="funder_identifier">	<funderIdentifier funderIdentifierType="...">	<funder_id>	<funderIdentifier>
PID for research performing orgs	Articles	Encouraged	Encouraged	<institution-id>	<institution><insitution_id>	<affiliation affiliationIdentifier="URL..." affiliationIdentifierScheme="...">		
PID for grant	Articles	Encouraged	Encouraged	<award-id>	<fundref_data>...<assertion award-number="X">			<oaire:awardNumber>
PID for project	Articles	Encouraged	Encouraged					
Final version or AAM	long-form	NA	Mandatory	<article-version>	<resource content_version="X">		<rioxxterms:version>	<oaire:version>

Appendix B - prevalence of ISSNs in repository data

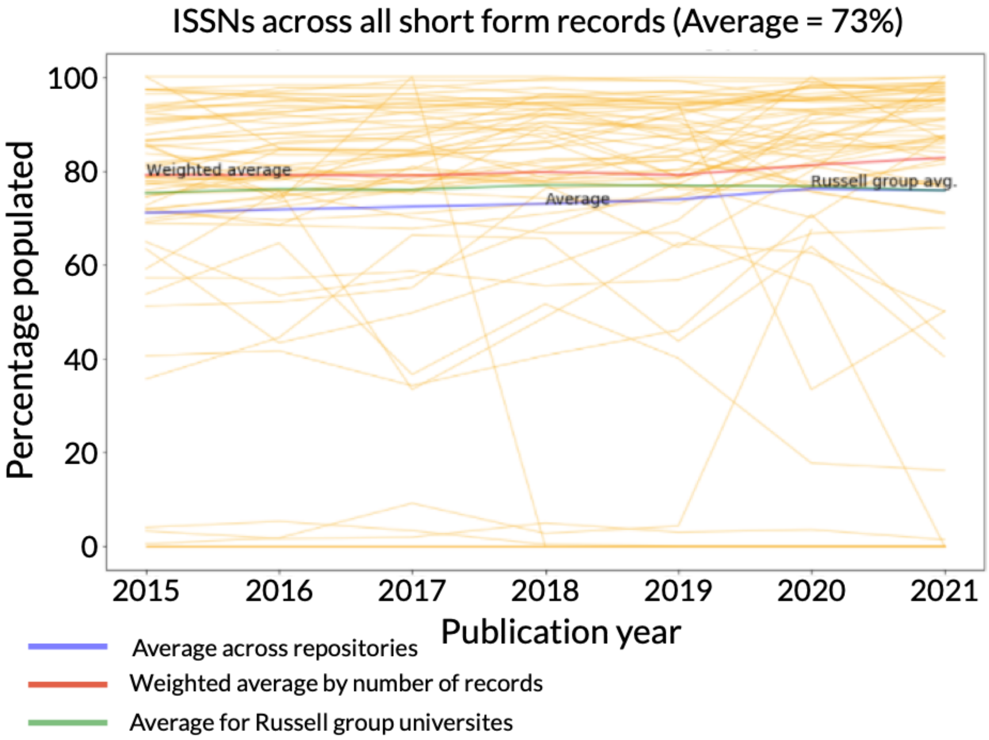


Figure B1: Percentage of ISSN fields populated for short-form (research articles, conference proceedings, and reviews) content in RIOXX repositories in the UK. The orange lines represent each individual repository, giving a sense of the variation between institutions.

Unlike publishers' metadata, repositories do not have a journal-centric publication workflow so, unsurprisingly, the level of metadata compliance for ISSN is not 100%. However, it is fairly high with an average across institutions of 73% and trending slowly upwards on average. The average compliance rate for Russell Group universities seems to be trending slightly downwards, but this may just be a random fluctuation rather than indicative of an issue.

ISSN metadata compliance for long-form content is not shown, as books and monographs generally don't have ISSNs.

Appendix C - sample interview outline

1. Interviewee Name
2. Short descriptor prepared in advance - stakeholder type, role, technological or other specialist knowledge, any evidence of their interaction with the policy to date
3. Briefly explain the project
4. Draft question set (note: we look for engagement, awareness, information about challenges, and potential suggestions for expediting implementation so use these questions to guide the general conversation rather than as a standard rote process)
    a. How does the new policy potentially impact their work (services offered, volume of work, new solutions or problems)?
    b. Do they see potential issues for other stakeholders in the ecosystem? Will it change how their organisation interacts with other entities?
    c. Are there interactions or other interventions they will need which involve third parties, e.g. Crossref?
    d. Do they think existing frameworks (like RIOXX) or services (like the Publications Router) are ready to support the policy?
    e. How much of a problem will this be for international research groups and other cross-cutting entities?
    f. Do they need anything more in the way of information or other support?
    g. Informally, how do they feel the new policy is being received?

Appendix D - infrastructure to support UKRI in green OA compliance monitoring

Please see attached separate file:

Appendix D - Infrastructure-to-support-ukri-oa-policy-with-CORE.pdf