Catalogue & Index | Issue 209 (2024)

Metadata & Discovery Group

Enhancing the Oxford University Research Archive

compliance, consistency, and challenges in supporting researchers

Tom Mitchell 0009-0007-5451-0486
Open Access and Repository Supervisor, Bodleian Libraries, University of Oxford

Tuula Lindgren 0009-0001-7105-7819
Open Access and Repository Supervisor, Bodleian Libraries, University of Oxford

Jason Partridge 0000-0001-9819-9991
Open Access Service Manager, Bodleian Libraries, University of Oxford

Received: 27 Nov 2024 | Published: 16 Dec 2024

ABSTRACT

This paper examines recent developments made to the Oxford University Research Archive (ORA), which has been in place as the University of Oxford's institutional repository, supporting Oxford researchers, since 2007. Over the past four to five years, ORA has undergone significant updates, including the implementation of a new data model, the adoption of enhanced metadata standards, and the integration of automation tools. These changes have streamlined workflows for repository staff by reducing manual tasks, improving metadata consistency, and ensuring compliance with both institutional and external policies, thereby benefiting both researchers and repository administrators. The paper concludes by exploring potential future directions for the continued development of ORA in alignment with emerging researcher needs and policy landscapes.

KEYWORDS institutional repository; metadata standards; metadata enhancement; data model; automation; open access; policy compliance; system development

CONTACT Jason Partridge jason.partridge@bodleian.ox.ac.uk Bodleian Libraries

Introduction

The Oxford University Research Archive (ORA) has served as the University of Oxford's institutional repository since 2007[1]. In this role, it supports researchers to ensure the widest possible access to research outputs including publications (e.g. journal articles, book sections, reports), research theses and data, and so-called grey literature (e.g. working and discussion papers), in line with institutional objectives. ORA has gone through several interfaces and systems and, as such, several significant changes to the processes by which research objects are created, added, and enhanced. In 2020, ORA systems moved from a workflow of editing raw XML code (using MODS

[1] https://ora.ox.ac.uk/about

metadata schema) to a form-based entry interface, encompassing the use of richer standards, controlled vocabularies, and identifiers. Additionally, a 22-month project was completed in September 2024, delivering over 100 improvements across 10 milestones, impacting on repository workflows in the processing and sharing of deposited research.

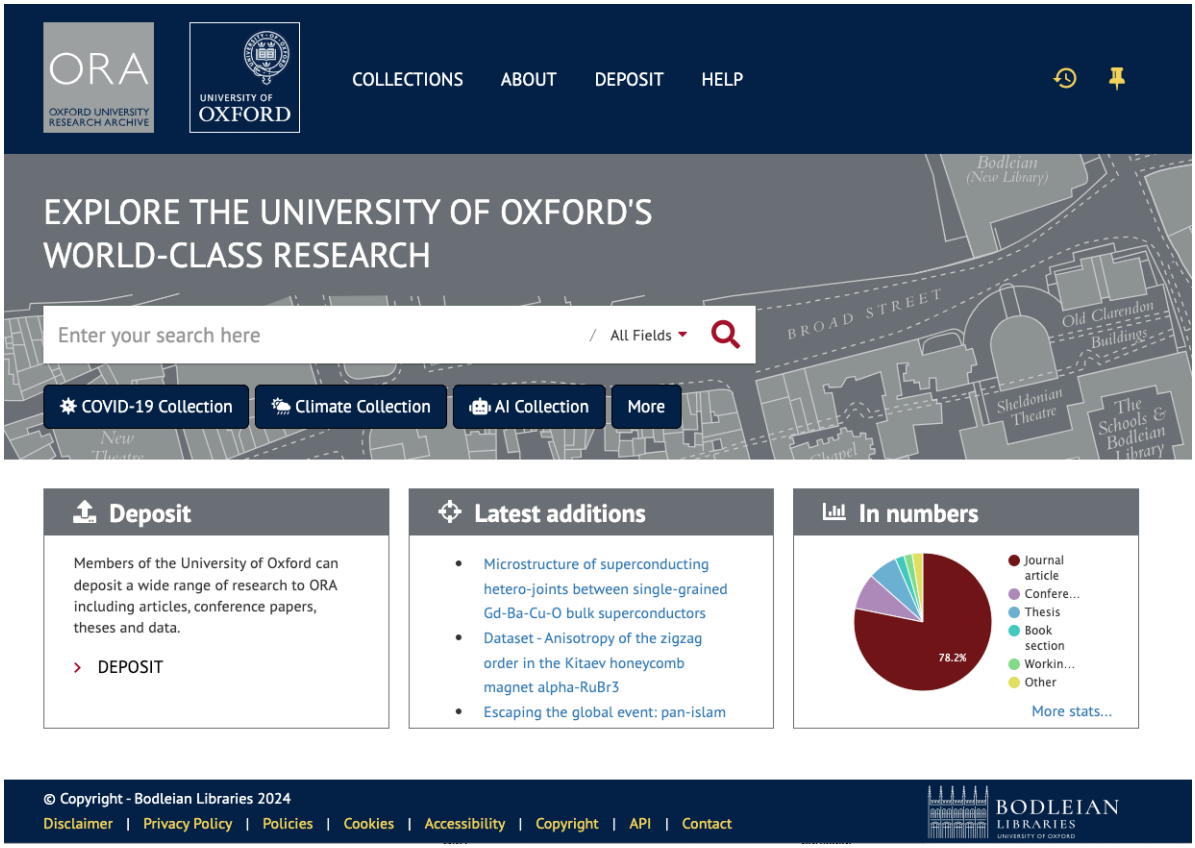

Figure1: The public homepage for the Oxford University Research Archive (ORA)

Repository and metadata management literature details the importance of good quality descriptive metadata in the discoverability of research, aiding visibility (Yang, 2016), indexing (Chiu, Chen and Cline, 2023; Editage Insights, 2024), citation, understandability, and reuse (Kemp, Dean and Chodacki, 2018). It also plays an important role in archiving and digital preservation (Digital Preservation Coalition, n.d.; Płoszajski, 2017). The importance of good quality descriptive metadata has also been highlighted in funder policy (Coalition S (n.d.); UK Research and Innovation (UKRI) (2022)), which often points to specific metadata standards or principles such as FAIR (GO FAIR (n.d.)), and is supported by institutional statements surrounding the access and longevity of research (University of Cambridge (n.d.)) A key role for ORA, therefore, is to ensure that metadata describing Oxford research to both humans using the public interface and machines using ORA's application programming interface (API) is of high quality. These goals were key to the recent improvements to ORA's 'review' system.

ORA metadata review process, checking, enhancing information

While aiming for high quality metadata, ORA also seeks to make depositing research outputs as easy as possible for researchers, understanding the pressures which already exist on their time. Most item types are deposited via Symplectic Elements (Elements)[2] and require only minimal manual entry of data for authors or a departmental administrator depositing on their behalf. Additional 'optional' fields are available for the depositor to add more information should they desire. If a paper is already published and available on a publisher website or through a source such as CrossREF[3], then Elements can collect metadata about the publication and present it to Oxford users - matching to an author based on criteria such as name, email address, or ORCID. Within the Elements interface authors may then only need to "claim" an output and attach a file to send it to ORA. Theses and datasets require slightly more effort upon deposit, as ORA registers DataCite[4] Digital Object Identifiers (DOIs) for these items, and both repository staff and creators must be happy that metadata is correct before this can be done. Nonetheless, every effort is made to reduce the burden on depositors.

The job of checking the items deposited to ORA, correcting, and enhancing the metadata describing research outputs, is that of the repository's review staff. Their role is to ensure all currently available bibliographic information (such as acceptance date, volume number, abstract, etc.) is added to a record before making it publicly available (often requiring further "check backs" to add information not yet available at time of first review). The review process takes place within the ORA review interface, built using Samvera open-source software called Hyrax. This has been developed as a tabbed set of review 'steps' within a form to allow the enhancement of metadata and description of files – a task previously done by editing XML code directly.

Whilst the previous method of editing XML for ORA records provided staff with flexibility in metadata creation, this left more room for error and for deviation in the standard of review or presentation of content. The current review interface allows for more conformity in review and the establishment of a shared standard that can be checked and measured within the team.

ORA Data Model

The underlying guiding structure to ensure consistency and control in the reviewing process is the ORA Data Model[5]. The data model manages the field structure and value entry for the ORA review interface, but is also applied to the ORA public interface, the repository storage layer, and API output, allowing for uniformity and alignment of the model across the whole ORA system.

[2] https://www.symplectic.co.uk/theelementsplatform/
[3] https://www.crossref.org/
[4] https://datacite.org/
[5] http://dx.doi.org/10.5287/bodleian:pr22x1bjE

For the purposes of the review interface this means that metadata fields such as title, contributors (authors, etc.), host title (such as journal title or book title), and dates (such as copyright year) are consistently labelled, ensuring that citation references can be built for each repository object in consistent formats and styles across various item types (journal article, conference paper, etc.).

The Data Model also allows for controlled vocabularies to be used to ensure that naming and terminology is uniform, for example for fields such as peer review status (Peer Reviewed, Reviewed Other, Not Reviewed), or re-use licence (CC-BY, CC-BY-ND-NC, etc.).

Where available, the Data Model utilises metadata frameworks or identifiers as lookups for value standards. An example of this is in the look-up of subjects to the FAST schema[6] or in funder names - using Research Organization Registry (RoR)[7] identifiers. It also implements ISO standards, such as ISO 639[8] for language codes (e.g. eng, English) and ISO 8601 for dates[9] (YYYY-MM-DD), promoting consistency across records.

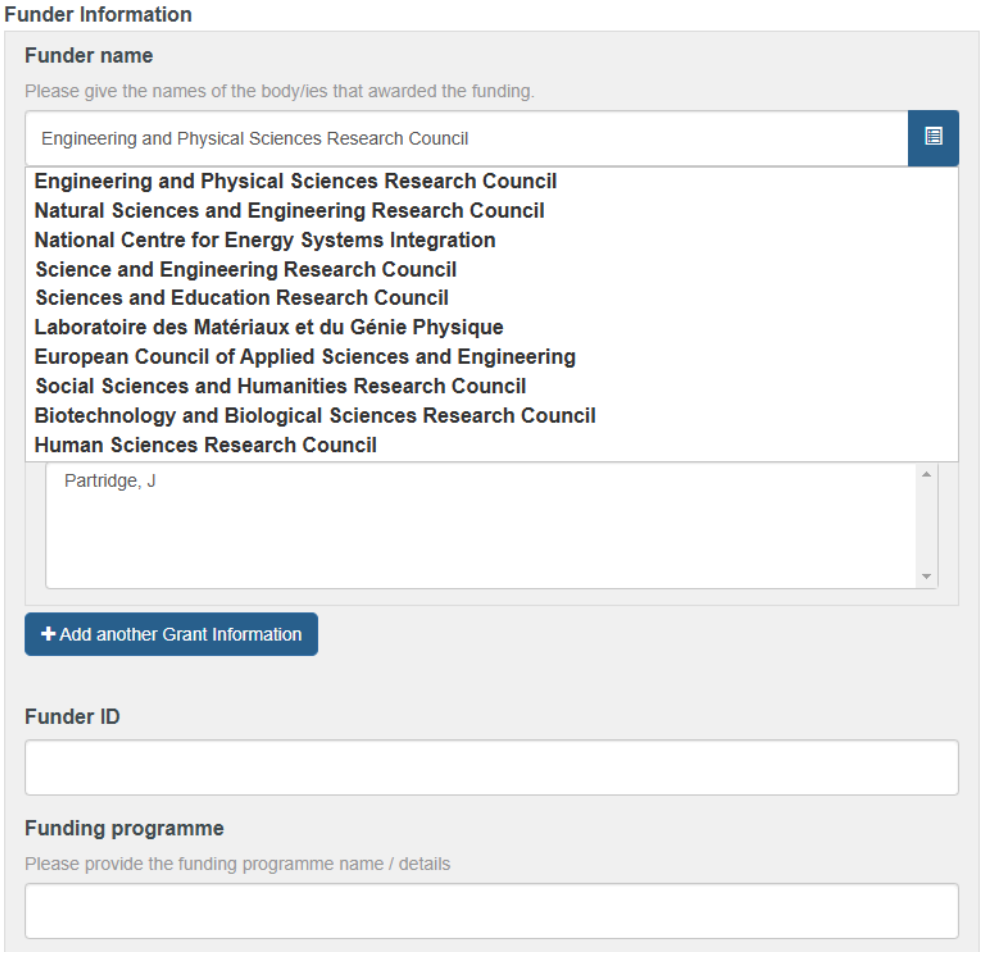

Figure 2: Selecting a funder using a lookup in the ORA review system

[6] https://fast.oclc.org/
[7] https://ror.org/
[8] ISO - ISO 639 — Language code
[9] ISO - ISO 8601 — Date and time format

Using these standards helps ORA keep metadata uniform and compatible with other systems, making it easier to share data and ensure interoperability.

On a local level the ORA review interface links with central user directories which allows for user details (names, email addresses, university department information, ORCID, SSO, etc.) to be added automatically to an object by a look-up using name or email address during the review process.

To further assist with consistency within the review process and to ensure that a certain standard of review is upheld across the review team, Open Access and Repository Supervisors undertake regular checks on public ORA records, asking staff to correct/add anything missed upon review, and check the private metadata to ensure that all reporting is correct.

Automation in ORA metadata review forms

Another element of the recent development project has been introducing automation to the review processes. Some of these features are relatively minor, such as buttons to “decapitalise” values in fields such as title (to conform to Resource Description and Access (RDA) standards) and keywords. At this stage, the decapitalisation is not particularly smart (see ‘Future thoughts’ below), and when using the button, staff must still take the time to ensure proper nouns and abbreviations are properly capitalised, but this is usually faster than manually changing individual letters at the start of every word in a title or keyword set.

A button has also been introduced to the review forms to automatically add text to a “public records note” field, generating an automated statement declaring that an available file is an “accepted manuscript” version, and that the final version may be downloaded from [publisher] at [DOI link]. Further buttons have also been used in the review forms to allow reviewers to simply click to follow a URL or DOI, and to perform searches, such as a Google search of a title, or to look up a journal title or ISSN with Open Policy Finder[10], quickly taking staff to information regarding publisher or journal embargo policies. Whilst these are small time savings within the review process, due to the volume of deposits that require processing, over time this accumulates to a significant reclaim of working hours.

A more substantial feature developed is a process to convert Word document files to PDFs and rename files to match ORA’s standardisation (Name_et_al_2024_First_three_words.pdf). In cases where only one file is present this action now happens on the ingestion of a deposit, but it can also be manually activated by use of a button within the review form. This can cut the time required to review a deposit by several minutes.

The result of these changes to the review interface and review workflow is a 25% increase in the throughput content processed by review staff between the beginning of the project and project close.

[10] https://openpolicyfinder.jisc.ac.uk/. Previously Sherpa Services (ROMEO, JULIET, FACT)

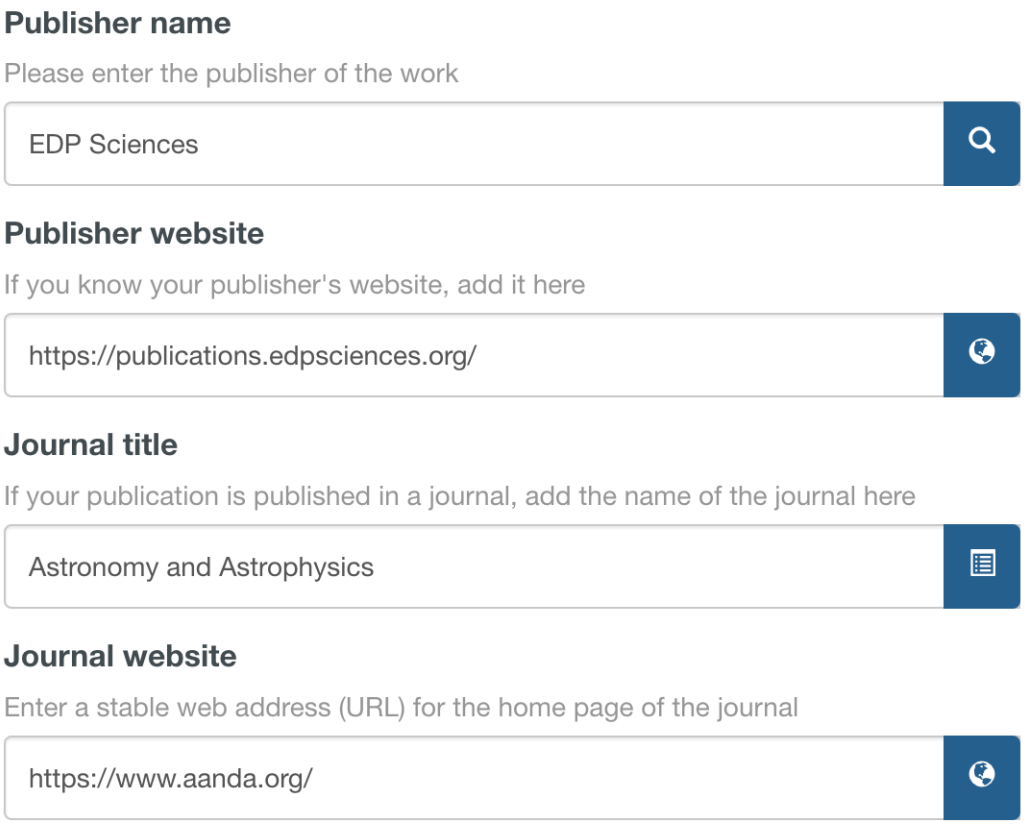

Figure 3: Time-saving buttons alongside the metadata fields in an ORA review form

Open access compliance and interoperability

Alongside improving the metadata review process, ORA has made efforts to improve interoperability and meet open access requirements set by frameworks like the Research Excellence Framework (REF)[11], OpenAIRE[12] and Plan S[13], as well as funders such as UK Research and Innovation (UKRI)[14]. These mandates require research outputs to be openly accessible with interoperable metadata to support accurate compliance reporting.

To meet these requirements, ORA has implemented RIOXX v2[15] standards to metadata and has plans to move to v3 in early 2025. ORA's metadata infrastructure has been updated to ensure that persistent identifiers (PIDs), open access status, re-use licences, and funder details (using RoR) are in line with technical standards set by UKRI and Plan S.

[11] https://www.ukri.org/who-we-are/research-england/research-excellence/research-excellence-framework/
[12] https://guidelines.openaire.eu/en/latest/
[13] https://www.coalition-s.org/addendum-to-the-coalition-s-guidance-on-the-implementation-of-plan-s/principles-and-implementation/
[14] https://www.ukri.org/publications/ukri-open-access-policy/
[15] https://www.rioxx.net/profiles/

ORA shares metadata using the Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH), providing it in various metadata schemas and standards that meet the requirements of scholarly portals like OpenAIRE, DART, and BASE[16]. Outputting metadata to recognised standards and schema also improves compatibility with open access infrastructures, and open access aggregator services like CORE[17], which in turn further impacts the discoverability of University of Oxford research. Since ORA's recent OAI-PMH work, more downloads and interactions with ORA items have been recorded, showing an increase of 172% in download activity (as recorded by IRUS UK) in the last two years (2022-24).

ORA has also incorporated automated harvesting services to streamline repository management and support open access compliance. The Jisc Publications Router[18], a service provided by Jisc, delivers Oxford-affiliated full-text articles directly from publishers to ORA, with a primary focus on gold open access and a selection of green open access content. This automated process reduces the need for manual deposit and update, ensuring that research outputs are more quickly and reliably accessible within the repository, while simultaneously further reducing the strain on researchers' time. Meanwhile, CORE provides ORA with automatic updates for open access content indexed from repositories worldwide[19].

Future challenges

ORA continues to develop its systems and workflows to maintain compliance with funders such as the UKRI[20] and Plan S technical[21] requirements for open access repositories as they continue to change. Funder policies and requirements also continue to have an impact on shaping how ORA collects, stores, and shares content, and the anticipated REF 2029 guidelines will again bring this into consideration.

Work on the next development project has already commenced, further exploring automation options to reduce manual workload and streamline processes for both repository staff and Oxford researchers. This includes integrating tools like Unpaywall[22] and OpenAlex[23] to retrieve open access full text versions of articles that have been identified as having an author affiliated to Oxford. Investigation is also underway into the use of artificial intelligence to improve metadata quality, such as automating the de-capitalisation of output titles and keywords upon ingest, or to extract metadata from PDF files.

[16] https://base-search.net/
[17] https://core.ac.uk/
[18] https://www.jisc.ac.uk/publications-router
[19] http://dx.doi.org/10.5287/ora-nb1bawday
[20] https://www.ukri.org/publications/ukri-open-access-policy/uk-research-and-innovation-open-access-policy/#section-annex-2:-technical-requirements-for-research-articles
[21] https://www.coalition-s.org/technical-guidance_and_requirements/
[22] https://unpaywall.org/
[23] OpenAlex https://openalex.org/ last accessed 21 November 2024

Identifiers continue to play an important role in both ensuring consistent naming and standardisation of metadata and names, but also in automation workflows. Email addresses, ORCIDs and Scopus IDs have recently been identified as having the biggest impact at Oxford with regards to automation, and work is being done to ensure that these are captured across the suite of services collecting researcher information and being made available to Elements and ORA. Building on the current ORCID usage, we hope to improve integration of ORCID identifiers with Jisc Publications Router to capture broader researcher data, such as affiliations and detailed contributions to collaborative research projects.

ORA continues to support Oxford's academic community, but to do so it must remain agile in development, flexible to change, and forward thinking in solution design.

References

Chiu, T., Chen, H. and Cline, E. (2023) 'Metadata implementation and data discoverability: A survey on university libraries' Dataverse portals', *The Journal of Academic Librarianship*, 49(4), p. 102722. doi: 10.1016/j.acalib.2023.102722.

Coalition S (n.d.) Principles and implementation of Plan S. Available at: https://www.coalition-s.org/addendum-to-the-coalition-s-guidance-on-the-implementation-of-plan-s/principles-and-implementation/ (Accessed: 14 November 2024).

Digital Preservation Coalition (n.d.) Metadata and Documentation. Available at: https://www.dpconline.org/handbook/organisational-activities/metadata-and-documentation (Accessed: 14 November 2024).

Editage Insights (2024) 'Role of metadata in enhancing discoverability and impact of scholarly research', *Editage Insights*, 10 September. Available at: https://www.editage.com/insights/role-of-metadata-in-enhancing-discoverability-and-impact-of-scholarly-research.

GO FAIR (n.d.) FAIR Principles: Findable, Accessible, Interoperable, and Reusable. Available at: https://www.go-fair.org/fair-principles/ (Accessed: 14 November 2024).

Kemp, J., Dean, C. and Chodacki, J. (2018) 'Can richer metadata rescue research?', *The Serials Librarian*, 74(1–4), pp. 207–211. doi: 10.1080/0361526x.2018.1428483.

Płoszajski, G. (2017) 'Metadata in long-term digital preservation', in Chen, J. and Wang, Y. (eds.) *Studies in Computational Intelligence*. Berlin: Springer, pp. 15–61. doi: 10.1007/978-3-319-51801-5_2.

UK Research and Innovation (UKRI) (2022) Open Access Policy: Technical Requirements. Available at: https://www.ukri.org/wp-content/uploads/2022/12/UKRI-301122-UKRI-open-access-policy-technical-requirements.pdf (Accessed: 14 November 2024).

University of Cambridge (n.d.) Digital Preservation Policy. Available at: https://www.lib.cam.ac.uk/files/pol_cul_digitalpreservationpolicy.pdf (Accessed: 14 November 2024).

Yang, L. (2016) 'Metadata effectiveness in Internet discovery: An analysis of digital collection metadata elements and Internet search engine keywords', *College & Research Libraries*, 77(1), pp. 7–19. doi: 10.5860/crl.77.1.7.