CORE and the Oxford University Research Archive (ORA)

Supporting Open Access through automated data gathering

Introduction

CORE harvests and aggregates information of research papers collected from institutional and subject repositories, and open access and hybrid journals[1], and makes the content available via an API (Application Programming Interface). The CORE API[2] offers a wealth of metadata and full text content from its many data providers.

For ORA (Oxford University Research Archive)[3], the use of the CORE API offers an opportunity to enhance workflows and streamline the process of reviewing and curating articles for inclusion in the repository.

The University of Oxford is subscribed to CORE as a Supporting Member[4] makes details of ORA repository content available for inclusion in CORE as a data provider[5].

The CORE API

The CORE API is an open source[6] API for accessing data on millions of documents from multiple data sources, including CrossRef[7], DataCite[8], BASE[9], PubMed Central[10], arXiv[11], and more. The API can also be used to access data from institutional repositories, such as ORA.

It enables institutions to quickly access and search data from multiple databases and provides a unified view of the different sources. The CORE API is easy to use and allows access to metadata fields for each document, including title, authors, date of publication, DOI, and other publication information.

The API also provides access to additional features, such as the ability to search for documents by DOI or title.

---

[1] https://core.ac.uk/about
[2] https://core.ac.uk/services/api
[3] https://ora.ox.ac.uk/
[4] https://core.ac.uk/membership
[5] https://core.ac.uk/data-providers/88
[6] https://en.wikipedia.org/wiki/Open_source
[7] https://www.crossref.org/
[8] https://datacite.org/
[9] https://www.base-search.net/
[10] https://www.ncbi.nlm.nih.gov/pmc/
[11] https://arxiv.org/

ORA Review Processes

With the changing landscape and policies impacting on how institutional repositories manage and process content, such as the Plan S initiative[12], it is increasingly important for a repository to know when research is published or updated with publication information.

The Wellcome Trust[13] and UKRI (UK Research and Innovation councils)[14] both have open access policies[15] that requires researchers to act in a certain way to ensure that content can be shared at the point of publication. The University of Oxford is also operating a Rights Retention policy[16] supporting self-archiving and open access to research that may otherwise be subject to an embargo. In order for a repository to be able to adequately release research at this point, publication needs to be known.

Current processes within ORA are that once a member of the University of Oxford has deposited to the repository a review process is undertaken. Deposit could be made as early as submission, though more commonly this is at the point of acceptance for publication.

The review process involves repository staff checking the content that has been deposited, checking publisher and funder policy affecting the sharing of the content, and enhancing metadata and bibliographic information regarding the deposit.

Where this involves the release of the full text upon publication, review staff mark the record for checking after a set period has passed – commonly one month from the date of deposit. This can therefore require the ORA reviewer to 'check-back' an object multiple times to wait for the paper to be published and to gather the necessary publication information.

Integrated Development

The CORE API version 3[17] allows for a number of metadata to be collected, including date of deposit (within the data provider repository), authors, acceptance date (for publication), abstract, title, DOI (DIgital Object Identifier), download URL, published date, and publisher.

To utilise the information available from CORE, ORA uses a RAKE[18] task to identify a set of candidate records for automatic update from the CORE API data. This has been limited to objects in the repository that are in a specific review 'state' or criteria - e.g. those not already considered 'complete'. The CORE API can then be used to provide an update to those records adding in updates to both record and full text file metadata where available.

Updates from the CORE API have so far been limited to information of publication, predominantly DOIs and publication date. However, criteria are being determined to allow for further metadata to be

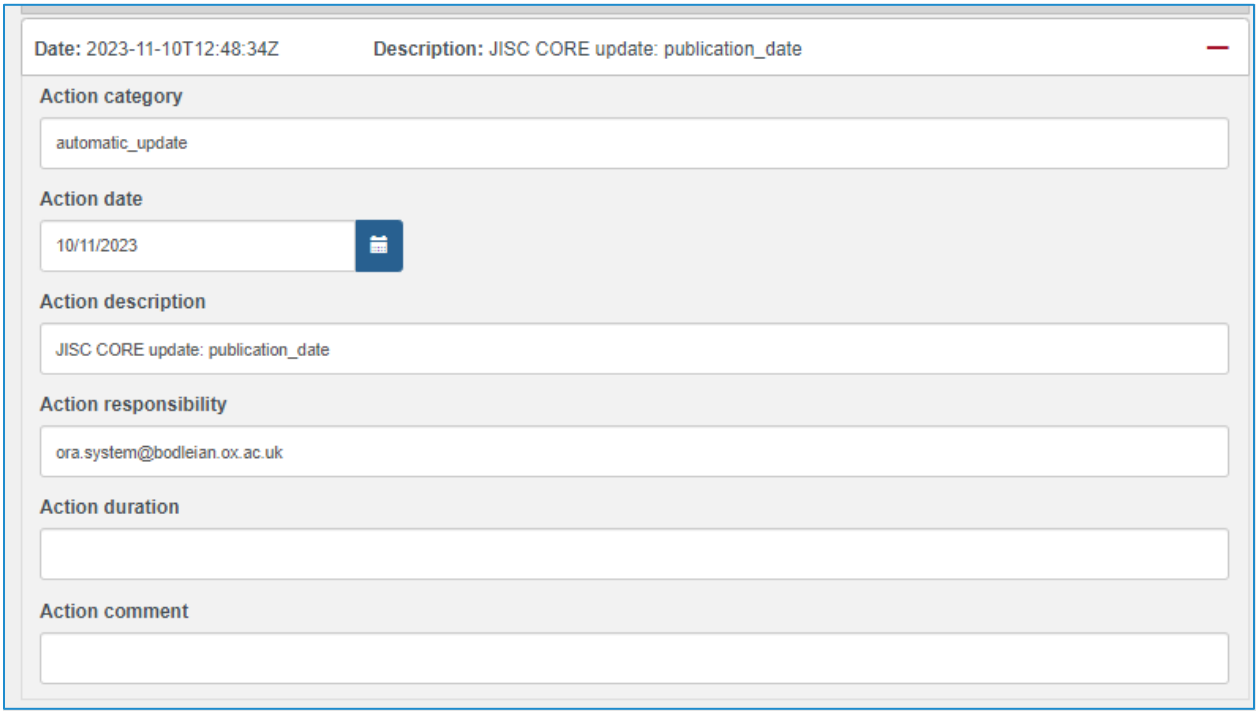
[12] https://www.coalition-s.org/
[13] https://wellcome.org/
[14] https://www.ukri.org/
[15] https://wellcome.org/grant-funding/guidance/open-access-guidance/open-access-policy / https://www.ukri.org/publications/ukri-open-access-policy/
[16] https://openaccess.ox.ac.uk/home-2/rights-retention/
[17] https://api.core.ac.uk/docs/v3
[18] https://en.wikipedia.org/wiki/Rake_(software)

accepted to ‘empty’ fields in a deposited record as well as indication where an existing populated field has conflicting or differing information.

Once an object has been updated with information from CORE this is flagged to the ORA review team by highlighting the object for an earlier check back (based on the new publication date information) and labelled to show that the update came from CORE.

An “updated by CORE” tag is added to the ORA record in the review interface

An audit trail is also added to the history of the object so that it can be determined when and what information has been updated by the CORE API query.



A history log entry is created for the updates added by the CORE API query providing detail on when and what was updated

Obtaining these programmatic updates help ORA to ensure accurate and timely release of files from embargo, supporting the University of Oxford in compliance with the requirements of funders, such as the Wellcome Trust and UKRI, and Rights Retention policies.

Criterion, technical challenges, and limitations

Harvested metadata from CORE is currently being utilised to complement existing review processes rather than completely replace them, i.e. action would still need to be taken by a staff member to accept and ‘publish’ any changes.

It was also determined that the automatic updates would only be added to objects that had not yet been reviewed or were in 'active' review, or awaiting 'check back', not to objects marked as 'complete' or otherwise discounted (e.g. duplicates).

An additional consideration made in what metadata is accepted to an ORA record was made for publication date - detailed below.

Metadata quality

As CORE harvests data from a number of data providers it is necessary to consider the accuracy of the data harvested. It is not uncommon for some metadata provided by sources to implement defaults or supply incorrect date information. For example, some source dates default to 1st January of a given year (or the 1st day of a given month) when only a year is available.

In order to gather the most accurate and relevant updates to the publication date field the following criteria was added:

if the record has not been reviewed: take all fields if not set; take more granular fields over less granular (e.g. if a publication date is 2023-01-01 in ORA, or 2023-06-01 (generic year or month), but the new publication date is 2023-06-13, take the new date)

The accuracy and completeness of the metadata within CORE relies on the information being supplied by the data providers. In testing it was also discovered that updates were required to ORA's own API output to prevent records incorrectly being marked with having a publication date to CORE where this did not exist.

API limitations

In setting up the query to the API there are also technical and content limitations. In terms of technical limitations this relates to the http headers used for the API and the number of search tokens allowed (150 per 5 minutes). These were primarily overcome by implementing a 'sleep' between API calls.

For bibliographic content, other limitations are based on the number of data providers, but also how frequently these are updated or harvested. In testing it was found that some DOIs within ORA that were being used to match within the CORE API did not yet exist in CORE.

CORE notes that whilst they try their "... best to have full coverage of DOIs by keeping synchronised with CrossRef and exposing and comparing DOIs from the repositories, however, we still don't have full coverage...".

None of the limitations or challenges were significant or insurmountable in continuing to make the API connection to CORE.

Conclusion

The CORE API offers flexibility and customisation, allowing developers to tailor their methods for obtaining publication information to their specific needs and requirements. For ORA Review staff, this means that the metadata available via the API can be integrated into existing workflows and processes.

Automated updates to metadata in ORA support efficiencies being made to workflows for the ORA Review staff, allowing information such as dates of publication, DOI, and other publication information for objects awaiting review or check back on publication to be added programmatically where possible.

However, due to the quality of some of the metadata being harvested this does not yet replace the need for manual checking and lacks some accuracy. This should be something that the whole UK sector of research institutions should look to offer improvement to – in what level of quality is being output for CORE and other services to harvest via API.

For now, the primary use of this automated mechanism for ORA is to provide an indication of publication and to bring ORA objects into the 'check back' process earlier than they may otherwise have been flagged to.

Next steps

Further development is now being explored as to how the links to full text content captured by CORE, where files are held in repositories and other providers of metadata to CORE, outside of ORA can be used to automatically identify and obtain content to ORA without a deposit being made at the home institution.

ORA (Oxford University Research Archive) *is the institutional repository for the University of Oxford. ORA was established in 2007 as a permanent and secure online archive of research materials produced by members of the University of Oxford.*

ORA aims to provide access to the full text of as much of Oxford's academic research as possible. This includes articles, conference papers, theses, research data, working papers, posters, and other content types.

More information can be found at: https://ora.ox.ac.uk/about

CORE *provides access to the world's largest collection of open access research papers, collecting and indexing research from repositories and journals. It is a not-for-profit service dedicated to the open access mission and one of the signatories of the Principles of Open Scholarly Infrastructures POSI.*

CORE serves the global network of repositories and journals increasing discoverability and preventing misuse of their content; making metadata records uniquely identifiable and resolvable with decentralised PIDs; supporting data providers in adopting good practices by providing tools for metadata validation, content management, enrichment and OA compliance; and facilitating machine access to open research.

More information at https://core.ac.uk/about

Jason Partridge, Bodleian Libararies
11 July 2023