

Introducing ORA

The Oxford University Research Archive (ORA)¹ has been in place as Institutional Repository (IR) for the University of Oxford since 2007, with the purpose of providing a permanent and secure online archive for research materials produced by the members of the University. This includes, but is not limited to the following:

- Journal articles
- Conference papers
- Working papers
- Technical reports
- Patents
- Book chapters
- Books
- Posters

The additional ORA-Data service² was launched in 2015 as part of the existing IR in response to external drivers, namely the expectations laid forth by the Engineering and Physical Research Council (EPSRC) with regard to research data management.³

ORA and its Open Access content⁴ is available to anyone with an Internet connection, though deposit to the archive requires University authentication. The repository, and its content, is managed and maintained by staff within the Bodleian Libraries⁵ although various stakeholders around the University are involved with decision making regarding its function.

Metadata records and files are received from multiple sources. At present most deposits are received from the University's Current Research Information System (CRIS) – Symplectic Elements (SE).⁶

SE is ORA's main source for records of traditional research publications (journal articles and conference papers) and grey literature such as working papers and reports. Theses and research data are received directly to ORA via in-house⁷ deposit forms. The source of record content for ORA currently influences the means by which deposits are reviewed and enhanced before a record (and file(s) where allowed) becomes available on the public website.

The ORA review team comprises an institutional repository librarian, a curator of digital research data, six metadata assistants, one research archive assistant and one senior research archive assistant, with the length of time in post varying from 1-5 years. The team has doubled in number over the last few years to account for the increase in number of deposits being received to the repository.

¹<https://ora.ox.ac.uk>

²<http://libguides.bodleian.ox.ac.uk/ora-data>

³<https://www.epsrc.ac.uk/about/standards/researchdata/expectations/>

⁴ Content available for download from the public pages. Other material may be stored in the archive under an embargo period awaiting release.

⁵<http://www.bodleian.ox.ac.uk/>

⁶ Information from the company Symplectic Ltd. can be found at: <http://symplectic.co.uk/products/elements/> and local to Oxford at: <https://www.admin.ox.ac.uk/researchsupport/awards/symplectic/>

⁷ Forms developed by ORA that form part of the application, not third-party application components (i.e. SE).

ORA Metadata Assistants are predominantly responsible for the review and enhancement of metadata associated with deposits made via SE and research data deposits, whereas the research archive assistant's main responsibility is for theses deposits.

In-house deposit and review

Creation and submission of records in-house, requires the user to complete an online deposit form which is broken into multiple stages. Each stage includes the option to add additional information as they proceed (Figure 1). The forms were developed with the intention of capturing sufficient information to facilitate the review of the deposit and to fulfil any requirements with regards to metadata from internal or external mandates.⁸

Figure 1. Snapshots of sections of the deposit forms for theses and data. Each deposit form has specific fields intended to facilitate and ease review of the records and also ensure compliance with external requirements for metadata.

The screenshot shows the 'About your award' section of the thesis deposit form. At the top, a progress bar indicates seven steps: 1. Upload files, 2. About your award, 3. About you, 4. Related publications, 5. Funders, rights and licences, 6. Making your thesis available, and 7. Review workflow. Step 2 is currently active. The form includes fields for:

- Title:** A text input field with the placeholder 'Example thesis title'.
- Subtitle:** A text input field.
- Abstract:** A text area containing the text: 'The Oxford University Research Archive (ORA) has been in place as Institutional Repository (IR) for the University of Oxford since 2007, with the purpose of providing a permanent and secure...'.
- Type of degree:** A dropdown menu with 'DPhil' selected.
- Date your leave to supplant (LTS) was granted:** A date input field.

 Below these fields is an 'Optional fields' section with the heading 'Enhance your deposit by making your thesis more discoverable via search engines.' It contains a button 'Add subject, keywords, and language information to your thesis' with a plus icon. At the bottom are navigation buttons: 'Back', 'Continue to Step 3 - About you', 'Next step', 'Save', and 'Check mandatory fields and Save'.

The screenshot shows the 'About your data' section of the research data deposit form. It features a similar progress bar to the thesis form, with steps: 1. Upload files, 2. About your data, 3. About you, 4. Related publications, 5. Data archiving and access, 6. Funders, rights and licences, and 7. Review workflow. Step 2 is active. The form includes fields for:

- Is your data in digital format?:** Radio buttons for 'Yes' and 'No'.
- Title:** A text input field with the placeholder 'Enter the title by which your dataset should be known and cited'.
- Subtitle:** A text input field.
- Documentation about your dataset and processes:** A text area with a plus icon for additional information.
- Publication information:**
 - Publisher name:** A text input field with 'University of Oxford' entered.
 - Year published:** A text input field with '2017' entered.
 - DOI:** A text input field.

 Below these fields is an 'Optional fields' section with the heading 'Enhance your deposit by making your data more discoverable via search engines.' It contains a button 'Add subject, keywords, dates, geographical location and language information to your data' with a plus icon. At the bottom are navigation buttons: 'Back', 'Continue to Step 3 - About you', 'Next step', 'Save', and 'Check mandatory fields and Save'.

⁸ EPSRC policy affecting research data has already been mentioned. For theses at Oxford it is a mandatory requirement as part of exam regulations to deposit an electronic version of the final examined work to ORA for all students who began their programme of study after 1st October 2007.

Once submitted the record moves to a separate review stage accessible only to ORA review staff preventing further edits to the submission by the depositor, although a function to refer the deposit back is available if needed.

ORA review staff are notified of a new deposit via email as the record reaches the review area. Deposits are 'claimed' by individual team members before beginning the review of a record to avoid conflicting reviewer edits. The review and enhancement of metadata for the records received via the in-house forms is made within the forms themselves. Correction and completion of information to the fields is done within each stage of the form.

Title fields for deposits are amended to conform to RDA (Resource Description and Access) standards and the abstract is enhanced with HTML (Hypertext Markup Language) and Unicode⁹ where necessary, so to render the abstract that displays on the record page as close to the papers content as possible. Where abstracts contain equations or molecular structures these are reproduced using HTML <table> tags (Figure 2) or where this may be too difficult to render, a JPG image file may be created and uploaded to the record page as a downloadable object. The image can then be referred to and displayed in the record abstract using the HTML syntax (Figure 3).

Figure 2. Rendered equation and HTML <table> mark-up used. The table <table>, table row <tr> and table cell <td> properties are used to determine the positioning of sections of the equation to compile it to the correct form (See ORA record: <https://ora.ox.ac.uk/objects/uuid:4b48ebc4-dd04-4741-8ab3-1ac8620f1cb0/>).

a)

$$|\nabla a|^2(b, \cdot) \leq C \frac{\|a(0, \cdot)\|_{\infty}^c - a^c(d, \cdot)}{b}, \text{ and } \frac{|\nabla a|^2}{a^{c_1}}, (d, \cdot) - \frac{\partial_b a}{a^{c_2}}, (d, \cdot) \leq \frac{C_1}{b}.$$

b)

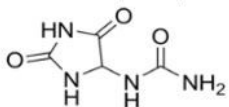
```
<table>
<tr>
<td nowrap>|&nabla;<em>a</em>|<sup>2</sup> (<em>b</em>,&middledot;) &le; <em>C</em> &nbsp;&nbsp;</td>
<td nowrap align=center>||<em>a</em>(0,&middledot;)||<sup><em>c</em></sup><sub style='position: relative; left: -
.4em;'>&infin;</sub> &minus; <em>a<sup>c</sup></em> (<em>d</em>,&middledot;)<hr noshade
size=1><em>b</em></td>
<td nowrap>,&nbsp;&nbsp;&nbsp;and &nbsp;&nbsp;&nbsp;</td>
<td nowrap align=center>|&nabla;<em>a</em>|<sup>2</sup><hr noshade
size=1><em>a<sup>c</sup><sub>1</sub></em></td>
<td nowrap>,&nbsp;&nbsp;&nbsp;(<em>d</em>,&middledot;) &minus; &nbsp;&nbsp;&nbsp;</td>
<td nowrap align=center>&part;<sub><em>b</em></sub><em>a</em></td>
size=1><em>a<sup>c</sup><sub>2</sub></em></td>
<td nowrap>,&nbsp;&nbsp;&nbsp;(<em>d</em>,&middledot;) &le; &nbsp;&nbsp;&nbsp;</td>
<td nowrap align=center><em>C</em><sub>1</sub><hr noshade size=1><em>b</em></td>
<td nowrap> &middledot;</td>
</tr>
</table>
```

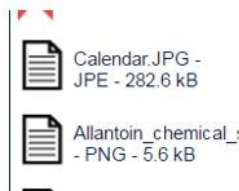
⁹<http://unicode.org/>

Figure 3. HTML src attribute used to display difficult compound structure as image imbedded to an abstract. 3a) example compound structure, 3b) example of HTML syntax to allow display.

a)

Abstract: Figure 3. HTML src attribute used to display difficult compound structure as image imbedded to an abstract. 3a) example compound structure [1], 3b) example of HTML syntax to allow display.





b)

```

```

Controls within the forms have some crossover with traditional cataloguing through the standards being used in the input of specific fields. FAST (Faceted Application of Subject Terminology) subject headings are provided as choices for subject information added to the forms, and Library of Congress (LoC) ISO (International Organization for Standardization) numbers are used for the language selected for the work, and date information.

Key aspects of traditional cataloguing and RDA standards such as Name Authority Records (NARSs) are not used. Author names and affiliations used to populate author information to the forms are obtained by querying the University's Core User Directory (CUD). As information is added to the field the user is able to select an existing University of Oxford author, which automatically populates institutional affiliation information (Division and department).

Due to not all authors being based at the University the field also allows for a free text entry. This adds to difficulty in linking works by a single author – a difficulty that is hoped to be overcome with the increased uptake of ORCID¹⁰ identifiers throughout the institution, and their incorporation into the underlying record metadata.

Metadata received from external sources – Symplectic Elements (SE)

ORA has been in receipt of metadata records and full text files from SE since 2008. In 2014 the Higher Education Funding Council for England (HEFCE) announced a Policy for open access in the post-2014 Research Excellence Framework (REF).¹¹ The Bodleian Libraries ran a pilot project in response to the policy to determine the most suitable deposit method for research publications affected by the policy, with the least burden on academic depositors.

¹⁰ <https://orcid.org/>

¹¹ All journal articles and conference published within a journal or conference proceedings with an ISSN accepted after 1st April 2016 are to be deposited to a repository within three months of acceptance. Updates to the policy have since seen this changed to within three months of publication up until 1st April 2018.

SE was selected as the deposit method for these items as the outcome of the pilot. The policy came into effect on 1 April 2016 and as a result the interaction between ORA and SE has increased significantly. Deposits have increased from an average of 80-100 received for review by ORA staff per month pre-April 2016 to an average of 1100 deposits per month post-April 2016.

Records in SE are created by a user completing a short deposit form and uploading a file (Figure 4) or by 'claiming' an existing record that has been obtained from an external datasource such as PubMed Central¹² or Web of Science.¹³ The deposit form was designed to be short and concise in order to make the deposit of an item quick and easy for academics, researchers and administrative staff. This does however mean that the additional content be added at review by ORA staff.

Review staff are notified of the deposit via email upon transfer of the file and metadata to ORA. Deposits made via SE are dealt with in a different review area to those received through the in-house forms, and involve editing the XML (Extensible Markup Language) directly.

There are multiple datastreams (XML and other metadata content files) for each record created for a record received from SE. These include:

- information of the related item within SE,
- a MODS (Metadata Object Description Schema) datastream,
- a DC (Dublin Core) datastream, and
- a citation datastream.

Enhancements and edits are made to the MODS datastream as part of the review process. This is done by incorporating the content of the MODS datastream into a new datastream called xml_ORA. Edits are then made to the xml_ORA datastream and the datastream is then preserved or 'locked' preventing further updates made to the record from SE overwriting the work done by reviewers.

¹² <https://www.ncbi.nlm.nih.gov/pubmed>

¹³ <http://apps.webofknowledge.com>

Figure 4. Screenshot of Oxford’s configuration of the SE deposit form for a new journal article deposit. The amount of information added to the form may be as minimal as the five mandatory fields (image used with kind permission from Symplectic Ltd. (<http://symplectic.co.uk/>)).

Full text files deposited along with the metadata are looked at as part of the review process. Information gathered from the file(s), and any information that may be sourced from a publication website is used to add additional information to the record (Figure 5).

Figure 5. Example of MODS xml_ORA content for journal article review.

```
MODS data
1 <mods:modsCollection xmlns:etd="http://www.ouls.ox.ac.uk/ora/modsextensions" xmlns:mods="http://www.loc.gov/mods/v3"
  xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance" xsi:schemaLocation="http://www.loc.gov/mods/v3
  http://ora.ox.ac.uk/access/mods-3.2-oxford.xsd http://www.ndltd.org/standards/metadata/etdms/1.0/
  http://www.ndltd.org/standards/metadata/etdms/1.0/etdms.xsd">
2   <mods:mods version="3.2">
3     <mods:titleInfo>
4       <mods:title>Example title record</mods:title>
5     </mods:titleInfo>
6     <mods:genre type="typeofwork">Journal article</mods:genre>
7     <mods:identifier type="pid">uuid:a07bd513-c947-4f55-3442-0dadb349b431</mods:identifier>
8     <mods:identifier type="urn">uuid:a07bd513-c947-4f55-3442-0dadb49b431</mods:identifier>
9     <mods:name authority="http://www.bodleian.ox.ac.uk/ora/authority" type="personal">
10       <mods:namePart type="given">John</mods:namePart>
11       <mods:namePart type="family">Smith</mods:namePart>
12       <mods:displayForm>Smith, JR</mods:displayForm>
13       <mods:namePart type="termOfAddress">Dr</mods:namePart>
14     </mods:name>
15     <mods:role>
16       <mods:roleTerm type="text">author</mods:roleTerm></mods:role>
17     <mods:affiliation type="institution">University of Oxford</mods:affiliation>
18     <mods:affiliation type="funding">Wellcome Trust</mods:affiliation>
19     <mods:affiliation type="funding">Royal Society</mods:affiliation>
20     <mods:affiliation type="grantNumber">102322/Z/23/Z</mods:affiliation>
21     <mods:identifier type="urn">uuid:65768350-4934-472d-a675-b8ab341b5d64</mods:identifier>
22   </mods:mods>
23 </mods:modsCollection>
```

An XML template is used to add the additional information into the xml_ORA datastream. The same RDA principles are used in reviewing the XML as within the in-house forms, but for fields such as date of publication the ISO format is applied manually copying from the XML template. The template covers common fields that are added or changed as part of the standard review process, such as publisher information, publication status and information regarding intellectual property rights, including copyright.

The current MODS schema has been customised for ORA specific review and is therefore many versions behind the latest available. Once edits to the metadata are completed updates to the MODS datastream are pushed from the xml_ora datastream. It is the MODS datastream that is used for the creation of the information displayed on the record page. DC metadata is automatically created using the MODS datastream. A record that has completed the review process is significantly more enriched with information than the original deposit, aiding the discoverability of the record and its usefulness as a resource (Figure 6).

Figure 6. a) A record submitted to review before reviewer enhancement. b) The same record after review.

<https://ora.ox.ac.uk/objects/uuid:a07bd513-c947-4f55-9742-0dad4b9b431>

Traditional cataloguing experience within the ORA review team is varied and is a result of training, education or previous employment. While undeniably helpful, this experience is not essential to undertake the review work. Experience with HTML is a benefit, but again not essential. Most of the in-house form review consists of data-entry in forms, and the underlying code provides the controlled vocabulary for the information entered. This is in contrast to the review of SE items where a template is used.

It is worth noting that review of deposits requires the review team to have extensive knowledge of policies affecting the dissemination of scholarly research, as well as being able to interpret publisher policies surrounding self-archiving.¹⁴ This is in addition to the knowledge of issues such as third-party copyright, the inclusion of sensitive content, and licencing.

A large amount of liaison with depositors is necessary during the review process, including communication about policies and requirements surrounding the deposit and access of a record, and any available full text.

¹⁴Often referred to as Green Open Access. Making an author accepted manuscript or Preprint available in a repository either immediately, after publication or on expiry of an imposed embargo period.

Onward sharing of metadata and future developments

As noted there are external policies that ORA needs to comply with in regards to metadata. These include policy regarding what information is recorded within the metadata and the format that this takes. For the HEFCE REF policy, one of these is the RIOXX Metadata Application Profile (RIOXX).¹⁵ RIOXX combines standard DC with additional RIOXX terms with the aim of aiding institutions in compliance with HEFCE metadata requirements.

ORA currently uses mappings from the MODS information to conform to RIOXX standards. An OAI-PMH (Open Archives Initiative Protocol for Metadata Harvesting) endpoint has also been developed to make record information available to internal and external services, including the Bodleian Libraries' own catalogue, SOLO.¹⁶

The ORA service is about to embark on a two-year project to upgrade most of its software, infrastructure and application stack design. An updated and scalable data model is being developed, which will see further improvement to ORA's compliance with endpoint standards (including OpenDOAR¹⁷, OpenAIRE¹⁸, UKRDDS¹⁹). It is anticipated that the current review process will see many changes as work begins, including the switch to writing metadata in RDF (Resource Description Framework) Triples. It is also planned that the review process will be performed through the use of further in-house forms, moving away from editing the xml datastreams directly.

¹⁵<http://rioxx.net/>

¹⁶<http://solo.bodleian.ox.ac.uk/>

¹⁷<http://www.opendoar.org/>

¹⁸<https://www.openaire.eu/>

¹⁹<https://www.jisc.ac.uk/rd/projects/uk-research-data-discovery>