Practice of CrossRef extensible markup language coding and more advanced information for CrossRef deposits

Rachael Lammey
CrossRef, Oxford, United Kingdom

Abstract
CrossRef is an association of scholarly publishers that develops shared infrastructure to support more effective scholarly communications. This infrastructure is centred around the Digital Object Identifier (DOI), and the associated bibliographic metadata that publishers deposit with CrossRef. To deposit this metadata, publishers need to provide the relevant information to CrossRef in extensible markup language (XML). Some publishers have a high level of XML expertise, and some are less familiar with formatting XML and need more assistance in order to register DOIs with CrossRef. There are some basics of depositing which can be picked up quickly, with the help of CrossRef tools, but there is a lot of information that publishers can provide via their XML to supplement the basic deposit information. This article will cover some of the more advanced information that publishers can deposit with CrossRef when they register DOIs, to help them participate in additional CrossRef services and provide additional value for the scholarly research community.

Keywords
Coding; CrossRef; Practice; Extensible markup language

Introduction
Publishers interested in learning about depositing the minimum required metadata with CrossRef can refer to the earlier article 'The basics of CrossRef extensible markup language', also published in Science Editing [1]. This article will provide details on the additional information that can be deposited within the CrossRef extensible markup language (XML) including references (to facilitate CrossRef’s Cited-by Linking service [2]), Open Researcher and Contributor IDs (ORCIDs), abstracts and full-text links and license information (which will facilitate participation in CrossRef Text and Data Mining Services [3]). It also provides brief information on the machine interfaces that publishers can use to efficiently upload XML to the CrossRef deposit system in a more automated way.
**Depositing and Distributing References**

In addition to article metadata, CrossRef members may also deposit reference lists. Reference deposit is recommended and there is no extra charge associated with it, but it is required for publishers who want to participate in Cited-by Linking. CrossRef Cited-by Linking is a service that allows publishers and their authors to discover how their articles are being cited. This information can be incorporated into a publisher’s platform so that this is available for their readers who can use it to find connections between related content. There is no charge for publishers to participate in this service, however in order for a CrossRef member to discover what publications cite their content, they must in turn submit metadata listing the works that their own publications cite.

In practice, this is not hard to do. Reference metadata can be easily included within normal DOI deposits, and it is now even possible to deposit references for Cited-by Linking through the cut-and-paste Simple Text Query interface [4]. Although there are other services that provide cited-by facilities (e.g., Google Scholar, Scopus, ISI and some hosting providers.), CrossRef Cited-by Linking complements these other services because: 1) It is controlled and managed by the CrossRef membership. As such, it enables direct primary publisher-to-publisher linking without the use of intermediaries. 2) It is not tied to any particular “metric”, although the metadata provided by the service could be used by others to either validate existing metrics or as the foundation for creating new metrics. 3) It is not constrained by discipline and, as such, may prove useful for discovery of otherwise hard-to-find interdisciplinary citations and for use by publishers in fields that are underserved by the other services. 4) It is not constrained by content type. CrossRef can accept reference data for journals articles, monographs, reference works, etc. 5) The CrossRef Cited-by Linking service is built on top of the DOI infrastructure and, as such, is very precise. 6) CrossRef is finding that those who seek to use CrossRef’s metadata services (CMS) [5] are increasingly interested in collecting reference metadata as well. CrossRef is concerned that, if they cannot get citation metadata from CrossRef, they will resort to using the much less accurate metadata that they can gather through screen scraping and services like Google Scholar.

Reference-only deposits are really nothing more than the list of references in an article’s bibliography. Under CrossRef’s normal linking service, members deposit an article’s metadata and then—in a separate process—query for the DOIs of the references [6]. In order to do that querying, the references must be extracted from the article. Reference-only deposits simply combine the identity of the article (the DOI in the metadata deposit) with its list of references.

References can be deposited by publishers in two ways. The first way is as part of their standard CrossRef metadata deposits along with the relevant article information. To do this, the XML metadata should be structured as shown below [7]:

```xml
<doi_data>
  <doi>10.50505/test_20051229930</doi>
  <resource>http://www.crossref.org</resource>
</doi_data>

<citation_list>
  <citation key="ref1">
    <journal_title>Current Opinion in Oncology</journal_title>
    <author>Chauncey</author>
    <volume>13</volume>
    <first_page>21</first_page>
  </citation>
</citation_list>
</journal_article>
</journal>
</body>
</doi_batch>
```

Publishers who do not wish to modify their existing metadata deposit process to include the deposit of reference information may instead deposit the reference lists in a separate reference-only deposit. A special schema [8] has been developed which allows for the deposit of references for an article (or indeed other items of additional metadata) that already has a CrossRef DOI. For publishers who want to deposit references in this way instead, they can follow the XML format shown below:

```xml
<doi_citations>
  <doi>10.5555/small_md_0001</doi>
  <citation_list>
    <citation key="ref1">
      <journal_title>Current Opinion in Oncology</journal_title>
      <author>Chauncey</author>
      <volume>13</volume>
    </citation>
  </citation_list>
</doi_citations>
```
In both approaches, an article’s references are described using the <citation_list> XML element. This element contains one or more <citation> elements (this is also common to both methods of depositing references). The current elements for citation tagging are: <issn>: ISSN (International Standard Serial Number) of a series (print or electronic), <journal_title>, <volume>: first author of an article or other publication, <first_page>: journal issue, <first_page>, <cYear>: year of publication, <article_title>: journal article, conference paper, or book chapter title <isbn>, <series_title>: title of a book or conference series, <volume_title>: book or conference proceeding title, <component_number>: the chapter, section, part, etc. number for a content item in a book.

CrossRef also accepts unstructured citations, which can be deposited using the <unstructured_citation> tag. These are citations for which no structured data is available, sometimes because they have been extracted from portable document formats (PDFs) or other systems that cannot tag the individual elements of the information extracted. CrossRef’s ability to process unstructured citations is limited, just because it means that it is more difficult to match them to related articles. As such, publishers should also include the DOI wherever possible (using the <doi> tag) when depositing reference information as matches may be made using the DOI as a key point of similarity.

By default, references deposited for a DOI are only distributed to the current owner of the DOI or to Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH) subscribers i.e., as those participating in advanced CMS when given permission by the current publisher. OAI-PMH is a protocol for metadata harvesting and is how advanced CMS participants retrieve the CrossRef metadata they subscribe to [9].

Publishers may also now choose to include deposited references in unified extensible markup language format (UNIXML) query results as well as OAI-PMH results. UNIXML querying returns the exact data submitted by the publisher of the DOI and not processed by the CrossRef system. This differs from other query result formats, like OAI-PMH that return data that has been processed by the CrossRef. If publishers elect to include their deposited references in the UNIXML query results, this makes the references publicly available. A publisher can contact support@crossref.org if they would like reference distribution enabled for the prefixes they manage. CrossRef encourages reference distribution for all members.

### Depositing Open Researcher and Contributor IDs within CrossRef Extensible Markup Language

ORCIDs [10] are unique identifiers for researchers, and CrossRef supports the deposit of ORCIDs for authors. Publishers can ask authors to provide them with their individual ORCID ID upon submission or acceptance of their work, and then deposit them with CrossRef with the rest of the article metadata. The presence of ORCIDs in the CrossRef XML enriches the information that can be discovered about an article and enable a reader to disambiguate between the author and another researcher who may share the same name.

Combined with CrossRef’s FundRef [11] service, the presence of ORCIDs will also allow funding agencies to tie agency funded research publications directly to researchers. In the long-term, widespread use of ORCIDs in CrossRef deposits could even let agencies start to develop publication key performance indicators (KPIs) for researchers that they fund.

An author's ORCID ID can be deposited with their details within the CrossRef XML as shown below:

```xml
<contributors>
  <person_name sequence="first" contributor_role = "author">
    <given name> Josiah </given name>
    <surname> Carberry </surname>
    <ORCID authenticate = "true"> http://orcid.org/0000-0002-185-0097 </ORCID>
  </person_name>
</contributors>
```

Publishers can also show if the ORCID has been authenticated or checked for validity within the XML. This is shown by the <ORCID authenticated = "true"> tagging.

### Depositing Abstracts with CrossRef

CrossRef changed their deposit schema in mid-2013 to support the deposit of abstracts within the XML. It is optional to include abstracts in CrossRef deposits, but if publishers do include them then they will be included in CrossRef’s established metadata distribution services—CMS and other CrossRef application programming interfaces (APIs). These services
already are very effective at disseminating bibliographic meta-
data with the option to include references. Including abstracts
in these services will benefit both members (some of whom
are eager to include abstracts) and CrossRef, as it will enhance
the information on content available via CMS.

Another stakeholder in academic research—funding agen-
cies—have an interest in building custom portals that high-
light agency-funded research. In order to provide users of
these portals with the best experience, agencies will want,
where possible, to display abstracts of publications along with
their standard bibliographic metadata.

If publishers want to deposit abstracts, these must conform
JATS is a set of XML elements and attributes used to tag jour-
nal articles that differs in some aspects from the CrossRef
XML. Many publishers already use JATS XML and it made
sence for CrossRef to use the JATS abstract formatting for that
reason. To deposit abstracts, publishers should add tags as per
the example shown below [13]:

```xml
< /publication-date >
...
</jats:abstract></jats:p>

We describe a method for cloning nucleic acid molecules onto the surfaces of 5 &times; 03BC;m micro-beads rather than in biological hosts. A unique tag sequence is attached to each molecule, and the tagged library is amplified. Unique tagging of the molecules is achieved by sampling a small fraction (1&times;0025;) of a very large repertoire of tag sequences. The resulting library is hybridized to microbeads that each cary &times; 2248;10 < jats:sup > 6 </jats:sup > strands complementary to one of the tags. About 10 < jats:sup > 5 </jats:sup > copies of each molecule are collected on each microbead. Because such clones are segregated on microbeads, they can be operated on simultaneously and then assayed separately. To demonstrate the utility of this approach, we show how to label and extract microbeads bearing clones differentially expressed between two libraries by using a fluorescence-activated cell sorter (FACS). Because no prior information about the cloned molecules is required, this process is obviously useful where sequence databases are incomplete or nonexistent. More importantly, the process also permits the isolation of clones that are ex-
pressed only in given tissues or that are differentially ex-
pressed between normal and diseased states. Such clones
then may be spotted on much more cost-effective, tissue-
or disease-directed, low-density planar microarrays.

Further details can be found in the CrossRef Schema Docu-
mentation of the < abstract > element [14].

Recording Links to Full Text of Documents, etc.

For some time, CrossRef member publishers have needed to
be able to record links to the full text of the content to which a
DOI refers. Additionally, publishers may want to offer differ-
ent versions (e.g., Accepted Manuscript or Version of Record)
and different representations (e.g., PDF for viewing, XML for
text and data mining, etc.) of the content tailored for specific
applications.

The < resource > element in CrossRef metadata is most of-	en used to record an hypertext transfer protocol (HTTP)
uniform resource identifier (URI) pointing at the publisher's
landing page for the publication identified by the CrossRef
DOI in question. However, the CrossRef schema has long
supported the recording of multiple < resource > elements in
order to enable, for example:

- Multiple resolution [15]: a service that allows multiple
URLs to be allocated to a DOI, in case content is legiti-
mately co-hosted on more than one platform.
- Search engine indexing: the ability to associate the DOI of
the item with the location at which it was indexed. Search
services often index content by where it was found (URL)
which may not be a stable/reliable key. Adding the DOI as
an index key adds stability.
- CrossCheck indexing [16]: CrossCheck is CrossRef’s pla-
giarism screening service [17], and providing full-text
links in the CrossRef metadata enables the content to be
indexed in the CrossCheck database so that other works
can be checked against it.

Funders are also interested in making sure that the full text
content of the best available version of an article is made avail-
able for reading, automated processing and archiving.

CrossRef has extended the ability to record multiple < re-
source > elements in order to allow the recording of URIs
which point to the full text of content identified by the Cross-
Ref DOI. The publisher can record multiple representations
of the full text (e.g., PDF, XML, plain text) using the new
mime_type attribute and then, through their access control
systems, control who is able to reach which representation of
an article and under which conditions. It is important to note
that, by recording a < resource > that points to the full text, a
publisher is not necessarily guaranteeing that the URI will be
accessible—this would still be managed by their access con-
trol systems.
The `<collection>` property should be used as a container for one or more items each holding a DOI or a resource (URI) which is related to the DOI in the `<doi_data>` element. Publishers can use the `<resource>` element for recording links to the full text, as per the example below [18]:

```xml
<doi_data>
  <doi>10.555/pubdate1</doi>
  <resource>http://www.yourrl.org/article1.html</resource>
  <collection property="crawler-based">
    <item crawler="altavista">
      <resource>http://www.yourrl.org/article1_altavista.html</resource>
    </item>
    <item crawler="google">
      <resource>http://www.yourrl.org/article1_google.html</resource>
    </item>
    <item crawler="iparadigms">
      <resource>http://www.yourrl.org/article1_crosscheck.html</resource>
    </item>
  </collection>
</doi_data>
```

A collection must be qualified by a property attribute or the multiple resolution attribute to show how it is intended to be used. Property attributes can be list-based: i.e., takes users to an interim page and presents the list of items to the user (via Multiple Resolution), or crawler-based: i.e. identifies resource to be crawled by the specified crawlers.

Information on how to record links to full-text in the deposit schema, as part of enabling the new CrossRef text and data mining service, is available on the text and data mining support site at: https://apps.crossref.org/docs/tdm/full-text-userguide.html. Full text links for text and data mining can be used to point to the full-text of the piece of content in all of the forms that the publisher can provide it in. For example, if a publisher has the full-text of a paper available in PDF and XML format, they can specify both in their XML so that a researcher wanting to text and data mine the content can choose the format they wish to harvest it in. This is called Content Negotiation [19] and is supported by CrossRef and by Datacite [20], another DOI registration agency.

### License Information

One of the main drivers behind the FundRef service was that many funding bodies are required to report on the public availability of the results and publications arising from their funder-financed research. Funders are therefore interested in understanding how publications related to funded research are licensed.

As such, CrossRef added the additional element `<ai:program>` to support the Access Indicators schema. Access indicators for published content are most commonly included in metadata deposits in the form of license information, so that a publisher can show which license(s) a piece of content is published under and therefore it's permitted usage. License information metadata collected by CrossRef includes:

- license URL element (license_ref)
- `start_date` attribute, optional, date format YYYY-MM-DD
- applies_to attribute, optional, allowed values are:
  - vor (version of record)
  - am (accepted manuscript)
  - tdm (text mining)

Access Indicators may be included in a metadata deposit or submitted as a resource-only deposit.

To deposit license information, publishers must use the `<license_ref>` element. The value of the `<license_ref>` element must be a stable HTTP URI which points to a human-readable document that either includes, or guides the reader to, any copyright and/or licensing information related to the CrossRef DOI of the content. The URI should point either to a location on the publisher's site or to the stable location of any well-known licenses such as those of the Creative Commons.

It is also acceptable to record a `<license_ref>` URI as a "placeholder." If a publisher is still working out specific licensing terms, the URI recorded in the metadata can point to a blank page or even a simple re-assertion of the document's copyright. This at least indicates an intent on the part of the publisher to eventually clarify licensing information.

Use of the `<license_ref>` element is best explained through examples. The `<license_ref>` for content licensed under the popular CC-BY [21] license, would look like this:

```xml
<license_ref>http://creativecommons.org/licenses/by/3.0/deed.en_US</license_ref>
```

Where as the Journal of Psychoceramics (a fake journal used by CrossRef for demonstration purposes) might record that their content is licensed under a proprietary license like this:  

```xml
<license_ref>http://www.psychoceramics.org/license_v1.html</license_ref>
```

A publisher can also deposit multiple `<license_ref>` elements, so the following would indicate that a document was available under a dual license (e.g., one for commercial applications and one for non-commercial applications):

```xml
<license_ref>http://www.psychoceramics.org/non-commercial_license_v1.html</license_ref>
<license_ref>http://www.psychoceramics.org/commercial_license_v1.html</license_ref>
```
Publishers may also want to record that a document is under embargo. In other words, that it is available under access control and a proprietary license for a set period of time, after which it is available under an open license. Publishers wishing to record embargoes can use the optional start_date attribute on the <license_ref> element.

For example, the following records that the content is under a proprietary license from its date of publication on February 3, 2014 and that it is under a CC-BY license a year later on February 3, 2015:

```xml
<license_ref start_date="2014-02-03">http://www.psychoceramics.org/license_v1.html</license_ref>
<license_ref start_date="2015-02-03">http://creativecommons.org/licenses/by/3.0/deed.en_US</license_ref>
```

It is important to note that the value of the start_date element must be recorded using the format YYYY-MM-DD The start_date attribute can be combined with multiple <license_ref> elements to indicate that a document is under a proprietary license during an embargo, but that it is then under a dual (commercial/non-commercial) license a year later:

```xml
<license_ref start_date="2014-02-03">http://www.psychoceramics.org/license_v1.html</license_ref>
<license_ref start_date="2015-02-03">http://www.psychoceramics.org/non_commercial_license_v1.html</license_ref>
<license_ref start_date="2015-02-03">http://www.psychoceramics.org/commercial_license_v1.html</license_ref>
```

There is no corresponding end_date attribute for the <license_ref> element. This is because including end dates could introduce ambiguities. For example:

- Open Licenses, such as CC, do not have “end dates”.
- With end dates, it would be possible to inadvertently record “gaps” between licenses.

By recording that another license takes effect in the future, a publisher is informing the user of the current restricted license is only for the embargo period. In short, the publisher is recording the intent to change the license when the embargo is done. Providing additional metadata for a current publication at some future date is an additional chore for the publisher that might well be overlooked. Again, this may not be something that publishers go into a lot of detail with when starting out with CrossRef, but they should consider adding this information to their CrossRef deposits in the longer term to provide useful information to a large number of interested parties in the publishing industry and beyond.

### Using Hypertext Transfer Protocol to Upload Files to CrossRef

This article has mainly focused on extra metadata that can be included in CrossRef deposits. However, another aspect of depositing with CrossRef that more technical users may be interested in is being able to bulk deposit files programmatically with CrossRef. The following information is more technical, and publishers can refer back to the related article, ‘The basics of CrossRef extensible markup language’ [1] for more manual means to deposit files.

Uploading files (for deposit or for bulk queries) are submitted using HTTP post with the encType: multipart/form-data. Multi-part/form-data is a more complicated encoding, but one that allows entire files to be included in the data. The URL for all submissions is http://doi.crossref.org/servlet/deposit. The parameters required within the encoding are shown in Table 1 [22].

The CrossRef website shows some sample transactions that publishers can use to format their own submissions as below:

```plaintext
POST http://doi.crossref.org/servlet/deposit?operation=doMDUpload&login_id=USER&login_passwd=PSWD&area=live HTTP/1.1
Accept: image/gif, image/x-xbitmap, image/jpeg, image/pjpeg, */*
```

<table>
<thead>
<tr>
<th>Form field</th>
<th>Description</th>
<th>Possible values</th>
<th>Mandatory</th>
<th>Default</th>
</tr>
</thead>
<tbody>
<tr>
<td>Operation</td>
<td>Depends on submission type</td>
<td>doMDUpload: For metadata (XSD) submissions doDOIcitUpload: For DOI citations or resources submissions doQueryUpload: For query submissions doDOIQueryUpload: For DOI-to-metadata query submissions</td>
<td>NO</td>
<td>doMDUpload</td>
</tr>
<tr>
<td>SubType</td>
<td>Subtype for metadata submissions</td>
<td>cm: For conflict management submissions</td>
<td>NO</td>
<td>N/A</td>
</tr>
<tr>
<td>Login_id</td>
<td>CrossRef supplied login</td>
<td>N/A</td>
<td>YES</td>
<td>N/A</td>
</tr>
<tr>
<td>Login_passwd</td>
<td>CrossRef supplied password</td>
<td>N/A</td>
<td>YES</td>
<td>N/A</td>
</tr>
<tr>
<td>Fname</td>
<td>Submission contents</td>
<td>N/A</td>
<td>YES</td>
<td>N/A</td>
</tr>
</tbody>
</table>

This program uploads files to the CrossRef system. If the given file has a 'list' extension it is considered to be a file with a list of files to upload. Any other extension is considered a single file upload. If the file is a directory then all files in the directory will be uploaded.

The default upload option is DEPOSIT, but alternatively it can be set to DEPOSIT_REFS for reference deposit. Other optional arguments are: -h (host) -hp (port) -ph (proxy host) -pp (proxy port), and users should note that -h defaults to doi.crossref.org, and -hp defaults to port 80. Again, for further information on this, users can contact support@crossref.org.

**Conclusion**

CrossRef offers a host of services to publishers that hinge on the registration of DOIs and article metadata within the CrossRef deposit system. The basic bibliographic metadata for a piece of content is useful for a host of purposes, the main one being persistent discoverability of publisher content. However, providing additional optional metadata such as references, full-text links, ORCIDs, abstracts, and license information serves to increase the functions that the metadata can perform and enable publishers to participate in additional CrossRef services like Cited-by linking and CrossCheck (via the deposit of as-crawled URLs). Participation in these extra services then adds value for various stakeholders; researchers, funding bodies, libraries and third parties like search engines. This document has aimed to describe what this extra information can encompass, and how publishers can add this supplementary information to their CrossRef metadata to engage with the increasing scope of the scholarly communications industry.

**Conflict of Interest**

No potential conflict of interest relevant to this article was reported.

**References**

4. Crossref. Simple text query form [Internet]. Lynnfield: CrossRef; 2013 [cited 2014 May 21]. Available from:
http://www.crossref.org/SimpleTextQuery/
21. Creative Commons. About the licenses [Internet]. Mountain View: Creative Commons [cited 2014 May 22]. Available from: https://creativecommons.org/licenses/by/3.0/