science editing

pISSN 2288-8063
eISSN 2288-7474

Sci Ed 2014;1(2):76-83
http://dx.doi.org/10.6087/kcse.2014.1.76

Training Material

The basics of CrossRef extensible markup language

Rachael Lammey

CrossRef, Oxford, United Kingdom

Abstract

CrossRef is an association of scholarly publishers that develops shared infrastructure to support more effective scholarly communications. Launched in 2000, CrossRef's citation-linking network today covers over 68 million journal articles and other content items (books chapters, data, theses, and technical reports) from thousands of scholarly and professional publishers around the globe. CrossRef has over 4,000 member publishers who join as members in order to avail of a number of CrossRef services, reference linking via the Digital Object Identifier (DOI) being the core service. To deposit CrossRef DOIs, publishers and editors need to become familiar with the basics of extensible markup language (XML). This article will give an introduction to CrossRef XML and what publishers need to do in order to start to deposit DOIs with CrossRef and thus ensure their publications are discoverable and can be linked to consistently in an online environment.

Keywords

Citation-linking network; CrossRef; Digital Object Identifier; Extensible markup language

Received: May 23, 2014
Accepted: July 29, 2014

Correspondence to Rachael Lammey
rlammey@crossref.org

ORCID
Rachael Lammey
http://orcid.org/0000-0001-5800-1434

This article was presented at the workshop entitled 'Application of advanced information technology to scholarly journal publishing' during the 12th EASE General Assembly and Conference, Split, Croatia, July 12-13, 2014.

Introduction

CrossRef's general purpose is to promote the development and cooperative use of new and innovative technologies to speed and facilitate scholarly research. CrossRef's specific mandate is to be the citation linking backbone for all scholarly information in electronic form. CrossRef is a collaborative reference linking service that functions as a sort of digital switchboard. It holds no full text content, but rather effects linkages through CrossRef Digital Object Identifiers (DOIs), which are tagged to article metadata supplied by the participating publishers. The end result is an efficient, scalable linking system through which a researcher can click on a reference citation in a journal and access the cited article.

Reference linking via CrossRef hinges on linking via the DOI. CrossRef DOIs are most frequently found in the reference lists at the end of scholarly articles and link persistently to other academic literature. However, a DOI is not a standalone object. Publishers register their DOIs with CrossRef by providing bibliographic information regarding the piece of content to Cross-

Ref, which distinguishes it from any other published work. This information is most commonly delivered to CrossRef in extensible markup language (XML) format. This article will cover the basics of CrossRef XML, or what publishers need to provide to CrossRef in order to register DOIs and therefore use them to maintain consistent links to their online content.

Getting Started

Although it may seem obvious, the first step a publisher needs to take in order to start assigning CrossRef DOIs to the content that they publish is to become a member of CrossRef. By becoming a member, a publisher is committing to maintaining DOIs for their content, and linking their references via DOIs and adhering to other criteria such as what the DOI response page should show.

Details on how to join are available via the CrossRef website at http://crossref.org [1] and there is an annual fee for membership. This is based on the annual revenue of the publisher. Once a publisher is a CrossRef member, they will be given a DOI prefix and log-in information for the CrossRef deposit system so that they can start to register DOIs.

What Is a Digital Object Identifier?

Similar to a bar code for a physical object, a DOI is a unique alphanumeric string assigned to a digital object, such as an electronic journal, article, report, or thesis. Each DOI is unique and serves as a stable, persistent link to the full-text of an electronic item on the Internet (Fig. 1) [2].

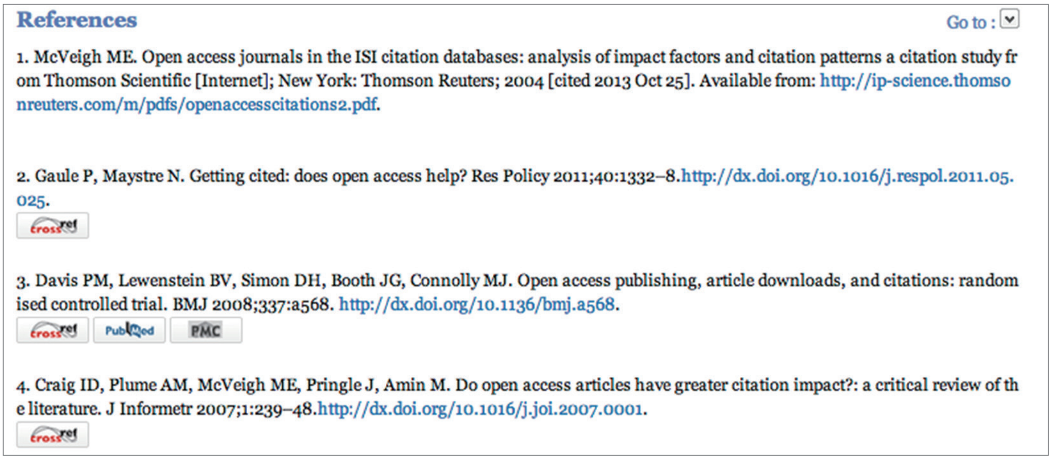
The advantages of assigning DOIs to content and linking references via the DOI are link persistence; unlike uniform resource locators (URLs), DOI links continue to function even if content moves or changes ownership. DOIs also aid content visibility and accessibility as CrossRef helps drive traffic to content by making it discoverable for linking and easier to link to. A single agreement with CrossRef serves as a linking agreement with all participating publishers. As such, it enriches the end-user experience, the scholarly research process, and the utility of published resources.

DOIs are the only widely adopted persistent identifier for scholarly works. DOI names appear in printed materials and online as links. A DOI name consists of two segments. The first is the prefix, a unique numeric string beginning with the numeral 10 assigned by CrossRef to the publisher that submitted the information about the digital object. The second is the suffix, an alphanumeric string or series of strings used internally by the publisher to identify the digital object.

In this sample DOI name: 10.6087/kcse.2014.1.13, '10.6087' is the prefix (in this case, for the publisher Korean Council of Science Editors) and 'kcse.2014.1.13' is the publisher-assigned suffix for the particular item (in this case, indicating that it is from the journal *Science Editing* and was published in 2014). DOIs will often be displayed as links, so the DOI above may be represented as: http://dx.doi.org/10.6087/kcse.2014.1.13. CrossRef encourages display of DOIs in this way for three reasons. 1) Users will more easily recognize CrossRef DOIs as an actionable link, regardless of whether they know about DOIs. 2) Users who do not know how to right-click on the link and choose "Copy Link", will still be able to easily copy the http URI. 3) Programs (e.g., bots, etc.) will recognize the DOI as a link.

When a publisher joins CrossRef, they are given a DOI prefix, but they can choose the pattern or system for the suffix themselves.

The DOI suffix has a very flexible syntax. It can be any alphanumeric string, consisting of a single node or multiple nodes. A node is a portion of a character string. A single node has no delimiters (periods, colons, pipes, and so on), for ex-



References

1. McVeigh ME. Open access journals in the ISI citation databases: analysis of impact factors and citation patterns a citation study fr
om Thomson Scientific [Internet]; New York: Thomson Reuters; 2004 [cited 2013 Oct 25]. Available from: http://ip-science.thomso
nreuters.com/m/pdfs/openaccesscitations2.pdf.

2. Gaule P, Maystre N. Getting cited: does open access help? Res Policy 2011;40:1332–8.http://dx.doi.org/10.1016/j.respol.2011.05.
025.

3. Davis PM, Lewenstein BV, Simon DH, Booth JG, Connolly MJ. Open access publishing, article downloads, and citations: random
ised controlled trial. BMJ 2008;337:a568. http://dx.doi.org/10.1136/bmj.a568.

4. Craig ID, Plume AM, McVeigh ME, Pringle J, Amin M. Do open access articles have greater citation impact?: a critical review of th
e literature. J Informetr 2007;1:239–48.http://dx.doi.org/10.1016/j.joi.2007.0001.

Fig. 1. Reference linking using the Digital Object Identifier (From Jeong GH et al. Si Ed 2014;1;24-6) [2].

ample: 123456. A character string with multiple nodes must include a delimiter (a period, colon, pipe, and so on) between each one, for example: 12.34.56. Each suffix must be unique within a prefix.

Because the DOI is an opaque string intended to remain unique and persistent throughout changes in ownership and location of the content, a publisher does not need to include any specific or descriptive information in the DOI. Such information forms the metadata associated with each DOI, which is submitted along with the DOI and URL. As such, if a publisher chooses to include such bibliographic information in a DOI string, it will have no meaning within the CrossRef or DOI system. Existing identifiers can also be used for the DOI suffix, such as an ISBN (International Standard Book Number) or existing internal numbering scheme.

Full guidelines advising on best practice for DOI suffixes are available at http://help.crossref.org/#establishing_a_doi_suffix_pattern [3], but in short, they suggest making the suffix concise, unique, case insensitive (DOIs are case-insensitive: 10.5555/ABC123 is the same DOI as 10.5555/abc123), consistent and extensible so that they could be used for parts of articles as well such as figures, tables, graphs but still reference the parent DOI of the article itself.

Depositing with CrossRef

To register a DOI with CrossRef, the DOI that a publisher has chosen for a piece of content should be deposited into the CrossRef system with some basic bibliographic information about the piece of content it is being assigned to.

Depositing metadata to CrossRef involves the creation of XML according to the CrossRef deposit schema. The deposit schema sets out the structure that the XML must adhere to in order to be accepted by CrossRef. This XML is submitted to the CrossRef system via public or machine interfaces. During the submission process, DOIs and metadata are added to the CrossRef system, and DOIs are registered with the Handle resolver which deals with the management and resolution of persistent identifiers.

The basic process for depositing with CrossRef consists of these steps:

1. Creating XML using the CrossRef deposit schema (non-technical users may use the Web Deposit form).
2. Verifying the XML that has been created.
3. Uploading the XML (via a web interface or programmatically).
4. Submitting the deposit.
5. Reviewing the submission logs to verify the success of the deposit.

Step 1: Creating Extensible Markup Language

There is a minimum set of information that should be deposited with CrossRef via the XML. This is basic bibliographic information on the piece of content, which serves to distinguish it from other works, and also serves the purpose of trying to stop publishers depositing more than one DOI for one piece of content.

There are details on the minimum metadata that should be submitted for each type of content here: http://help.crossref.org/#elements [4]. However, this can be worked through in a step-by-step way. Many publishers who are new to CrossRef may not have expertise in XML and as such, may start depositing using the web deposit form [5] which is freely available and helps compile the XML needed to deposit DOIs for their content (Fig. 2).

The web form enables publishers to select what type of content they are depositing (book title, journal article, and conference proceedings) and then enter the information on it using free-text. Information that is required is marked with an asterisk and users will not be able to proceed with deposit without providing the minimum information necessary.

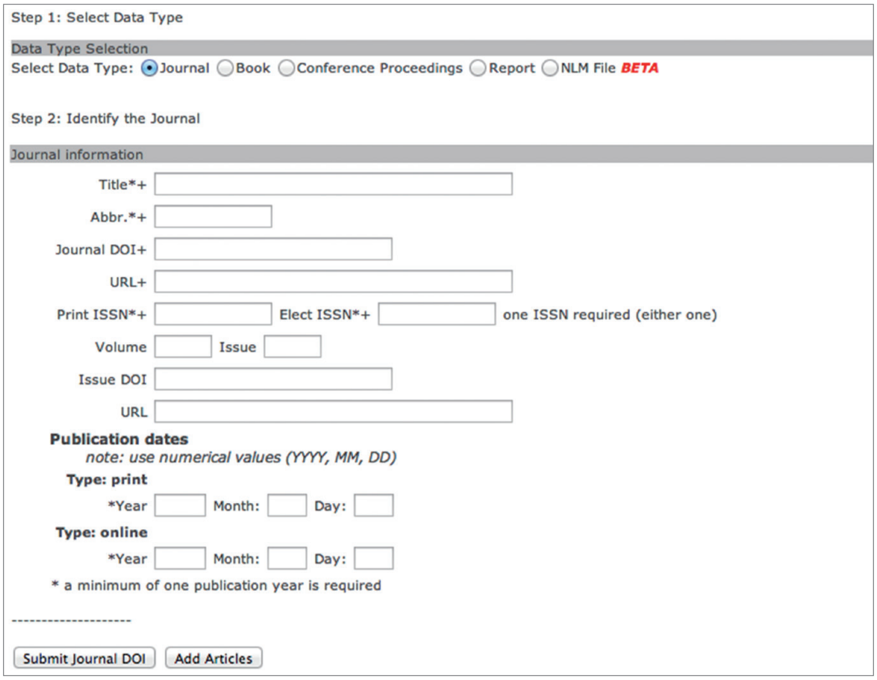
If a publisher selects 'Journal' then the first thing they will be prompted to enter is the journal title, abbreviated title, ISSN (International Standard Serial Number) and publication dates.

When they have done this, they can then use the 'add articles' button to register DOIs for articles from the journal they have just named. By the time they reach the next page to add these articles, the system is already starting to compile the XML needed for their CrossRef deposit (Fig. 3).

The publisher or editor can then add the bibliographic information for the article; title, the DOI they have chosen, the URL where the article is located and they can also add information on all authors of the paper (all authors should be entered for completeness of the record) and also the page numbers if applicable. There is also a button that users can click to fill out CrossMark metadata for the submission. The insertion of CrossMark metadata will be covered in a later document.

When the user has entered this information, they can choose to add another article or finish to deposit the articles with CrossRef. Choosing 'Finish' will prompt the user to enter their user name and password, provided by CrossRef when they joined. They should then enter an email address. This is where the results of the deposit will be sent. The article metadata can then be deposited with CrossRef and users will see a page and receive an email to say their deposit has been successfully received by the system.

The upload process is very basic and performs no data validation at the time of upload. The 'Success' acknowledgment displayed after submission step simply indicates that your file



Step 1: Select Data Type

Data Type Selection

Select Data Type: ● Journal ○ Book ○ Conference Proceedings ○ Report ○ NLM File **BETA**

Step 2: Identify the Journal

Journal information

Title*+

Abbr.*+

Journal DOI+

URL+

Print ISSN*+ Elect ISSN*+ one ISSN required (either one)

Volume Issue

Issue DOI

URL

Publication dates

note: use numerical values (YYYY, MM, DD)

Type: print

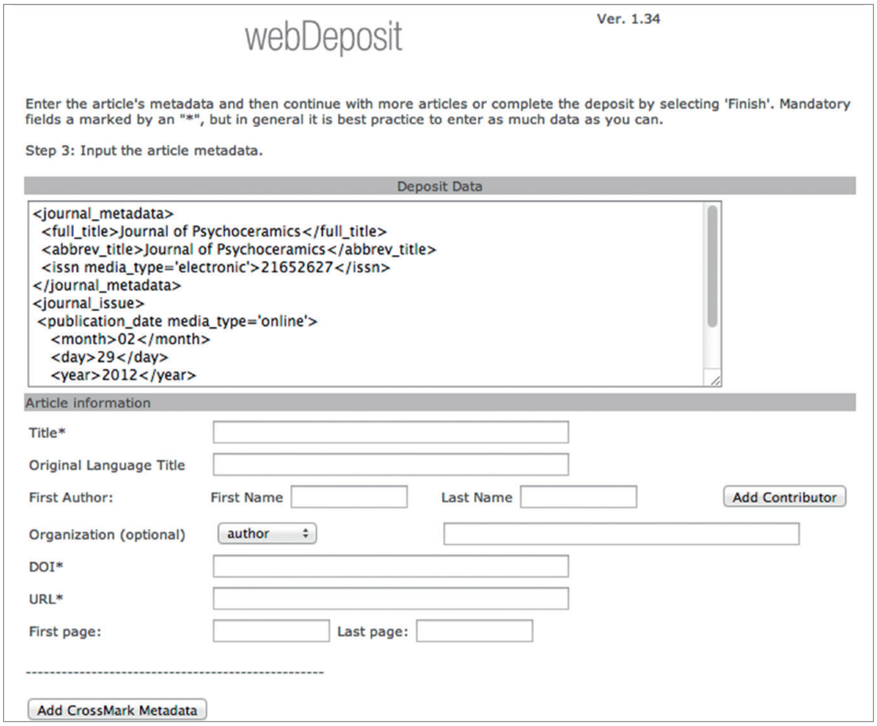
*Year Month: Day:

Type: online

*Year Month: Day:

* a minimum of one publication year is required

Fig. 2. The first page of the web deposit form on the CrossRef website.



webDeposit

Ver. 1.34

Enter the article's metadata and then continue with more articles or complete the deposit by selecting 'Finish'. Mandatory fields a marked by an "*", but in general it is best practice to enter as much data as you can.

Step 3: Input the article metadata.

Deposit Data

```
<journal_metadata>
  <full_title>Journal of Psychoceramics</full_title>
  <abbrev_title>Journal of Psychoceramics</abbrev_title>
  <issn media_type='electronic'>21652627</issn>
</journal_metadata>
<journal_issue>
 <publication_date media_type='online'>
   <month>02</month>
   <day>29</day>
   <year>2012</year>
```

Article information

Title*

Original Language Title

First Author: First Name Last Name

Organization (optional) author

DOI*

URL*

First page: Last page:

Fig. 3. The 'Deposit Data' section at the top of this page on the web deposit form shows the extensible markup language that is being compiled for CrossRef deposit.

has been received. Each uploaded file then goes into a queue to await processing which checks the XML submitted using a parser. This step verifies that the XML is well formed and conforms to the rules of the CrossRef schema. It also performs certain logic checks on the data in the file, for example publication title ownership is enforced. This means that

CrossRef recognizes a single publisher as owning a title and thus only DOIs using the prefix of that publisher may be assigned to the publication.

For a publisher or editor who is relatively new to XML, making a deposit in this way can be helpful as it clearly shows what information should be entered. Then, upon completion of their deposit they are sent the XML file of the deposit via email. The file contains the information shown below:

Journal section of CrossRef XML

```
< ?xml version = “1.0” encoding = “UTF-8”? >
< doi_batch xmlns = “http://www.crossref.org/schema/
  4.3.0” xmlns:xsi = “http://www.w3.org/2001/XMLSche-
  ma-instance” version = “4.3.0” xsi:schemaLocation =
  “tinyhippos-injected”/ >
< script id = “tinyhippos-injected”/ >
< head >
< doi_batch_id > -58cd1699144411615d6-6cc1 < /doi_
  batch_id >
< timestamp > 201405130618 < /timestamp >
< depositor >
< name > creftest < /name >
< email_address > rlammey@crossref.org < /email_ad-
  dress >
< /depositor >
< registrant > WEB-FORM < /registrant >
< /head >
< body >
< journal >
< journal_metadata >
< full_title > Journal of Psychoceramics < /full_title >
< abbrev_title > Journal of Psychoceramics < /abbrev_title >
< issn media_type = “electronic” > 02643561 < /issn >
< /journal_metadata >
< journal_issue >
< publication_date media_type = “print” >
< month > 08 < /month >
< day > 13 < /day >
< year > 2008 < /year >
< /publication_date media_type = “online” >
< mnoth > 08 < /month >
< day > 14 < /day >
< year > 2008 < /year >
< /publication_date >
< journal_volume >
< volume > 5 < /volume >
< /journal_volume >
< issue > 11 < /issue >
< /journal issue >
```

Article section of CrossRef XML

```
< journal_article publication_type = “full_text” >
< titles >
< title >
 Toward a Unified Theory of High-Energy Metaphysics:
 Silly String Theory
< /title >
< /titles >
< contributors >
< person_name sequence = “first” contributor_role =
 “author” >
< given_name > Josiah < /given_name >
< surname > Carberry < /surname >
< /person_name >
< /contributions >
< publication_date media_type = “print” >
< month > 08 < /month >
< day > 14 < /day >
< year > 2008 < /year >
< /publication_date >
< pages >
< first_page > 1 < /first_page >
< last_page > 3 < /last_page >
< /pages >
< doi_data >
< doi > 10.5555/12345678 < /doi >
< resource >
  http://psychoceramics.labs.crossref.org/10.5555-1234
  5678.html
< /resource >
< /doi_data >
< /journal_article >
< /journal >
< /body >
< doi_batch >
```

So the system has taken the text entered and transformed it to XML, which is helpful for the purposes of this article and also for the publisher to see how the elements they entered in the form appear in CrossRef XML format.

CrossRef encourages publishers to save the XML file sent via email post-deposit, as using it is the easiest way to perform updates of the data entered using the form. Instead of re-entering all the metadata, a publisher can edit the XML and resubmit using the system interface found here: http://doi.crossref.org/ [6].

Explaining the Basic Extensible Markup Language Elements

To explain the basic XML displayed, it may be useful to dissect what is shown in journal and article section of CrossRef XML and look at some of the individual elements entered.

The sections that the system adds are some standard elements to encase the deposit, for example:

<head> The container for information related to the DOI batch submission. This element uniquely identifies the batch deposit to CrossRef.

<body> The container for the main body of a DOI record submission. The body contains a set of journal, book, conference proceedings or stand alone component records. It is not possible to mix genres within a single DOI submission. It is possible to include records for multiple journals, books, conferences, or stand alone components in a single submission.

<doi_batch> Top level element for a metadata submission to CrossRef. This element indicates the start and end of the XML file. The enclosed information, "http://www.crossref.org/schema/4.3.0" xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance" version="4.3.0" xsi:schemaLocation="http://www.crossref.org/schema/4.3.0" http://www.crossref.org/schema/deposit/crossref4.3.0.xsd, shows the version of the schema you want the XML you are depositing to be checked against, and where that schema is located.

The <doi_batch_id> is automatically generated using the web-form, but for publishers not using the web form, you will need to enter this ID that uniquely identifies the DOI submission batch. It will be used as a reference in error messages sent by the deposit system, and can be used for submission tracking. The publishers decides on and enters this number themselves and should ensure that this number is unique for every submission to CrossRef. It should be more than four characters long.

The <timestamp> indicates version of a batch file instance or DOI. Again, this is automatically generated by the web form and is used to uniquely identify batch files and DOI values when a DOI has been updated one or more times. Every time a DOI is deposited it must be given a timestamp, the value of which must increment with subsequent updates. This value is a string of text that gets interpreted as a number. The recommended format is YYYYMMDDHHMM (ex: 200810021422). If a publisher wants to redeposit a DOI using the same XML, they should always increment the timestamp to a greater number so that the system recognises it as a more recent deposit, and will generate an error message if this is not done.

The <depositor> section is quite self-explanatory, but provides Information about the organization submitting DOI metadata to CrossRef. The name The name placed in this element should match the name under which a depositing organization has registered with CrossRef i.e. Elsevier, AIP. The email section refers to the e-mail address to which batch success and/or error messages are sent. It is recommended that this address be unique to a position within the organization submitting data (e.g., "doi@...") rather than unique to a person. In this way, the alias for delivery of this mail can be changed as responsibility for submission of DOI data within the organization changes from one person to another. The registrant is the organization that owns the information being registered.

It is also necessary to add the declaration: <?xml version="1.0" encoding="UTF-8"?> at the top of the form. Version="1.0" means that this is the XML standard this file conforms to, and encoding="utf-8" means that the file is encoded using the UTF-8 Unicode encoding.

The rest of the information contained is taken directly from the web form and tagged appropriately. Note that the URI where the content is located is described as a <resource>.

If a user is new to CrossRef, or if they intend to deposit a new content type, it is recommended that they verify the format and structure of an XML file before submitting it as a deposit to the system. Using these methods is quicker than verifying XML by trial and error after uploading it. There are two methods that can be used to do this. The first is to use the XML parser on the CrossRef website: http://www.crossref.org/02publishers/parser.html [7] which will quickly validate the files uploaded and let the user know how many DOIs it has found in the file. It will also bring up error messages for any badly-formatted XML but it will not deposit the XML in the CrossRef system.

CrossRef also has a test system at http://test.crossref.org [8]. This system functions identically to the live system but uses a test database and does not register DOIs. However, the test system is useful if a publisher wants to test large numbers of files or new titles. Once a publisher is satisfied that their XML is formatted correctly, they can submit it to CrossRef.

Reviewing Submission Logs

After a deposit is processed, the email address listed in the XML deposit will receive an email indicating the results (in an XML format), which lists the status of each DOI contained in the file. Note that while many DOIs in a file may successfully get deposited, individual DOIs may fail. Submission logs must be examined, and any flagged problems should be corrected and the file(s) resubmitted.

The submission log email will look like this:

```
<?xml version="1.0" encoding="UTF-8"?>
<doi_batch_diagnostic status="completed" sp="ds4.
  crossref.org">
  <submission_id>1366152368</submission_id>
  < batch_id > -58cd1699144411615d6-6ef0 < /batch_id >
  < record_diagnostic status = “Success” >
    < doi > 10.5555/test05052014 < /doi >
    < msg > Successfully updated in handle < /msg >
  < /record_diagnostic >
  < batch_data >
    < record_count > 1 < /record_count >
    < success_count > 1 < /success_count >
    < warning_count > 0 < /warning_count >
    < failure_count > 0 < /failure_count >
  < /batch_data >
< /doi_batch_diagnostic >
```

The section highlighted shows that the deposit was a success, and the success count shows how many DOIs were deposited from the XML file with the corresponding submission ID. However, if the deposit has failed, you will get an email like this:

```
< ?xml version = “1.0” encoding = “UTF-8”? >
< doi_batch_diagnostic status = “completed” sp = “ds4.
  crossref.org” >
  < submission_id > 1366152357 < /submission_id >
  < batch_id > -58cd1699144411615d6-6ef1 < /batch_id >
  < record_diagnostic status = “Failure” msg_id = “4” >
  < doi > 10.5555/test05052014 < /doi >
  < msg > Record not processed because submitted version:
    201405051257 is less or equal to previously submitted
    version (DOI match) < /msg >
  < /record_diagnostic >
  < batch_data >
    < record_count > 1 < /record_count >
    < success_count > 0 < /success_count >
    < warning_count > 0 < /warning_count >
    < failure_count > 1 < /failure_count >
  < /batch_data >
< /doi_batch_diagnostic >
```

And the < msg > section will tell you why the deposit failed. In the case shown above, a publisher tried to redeposit a DOI but without incrementing the timestamp. They should increment the timestamp element in the original XML and redeposit.

There are a number of common error messages that correspond to the most common deposit mistakes. These are listed, with the appropriate follow-up actions on the CrossRef Help site: http://help.crossref.org/#suberrors [9].

Deposit Tips

So that submissions can be processed efficiently, the file sizes of deposits should stay under 150 kilobytes. And overall file size should never exceed 1.5 megabytes. During times of very heavy loads, deposits may take several hours to reach the top of the queue and large files can take an hour or more to process. A publisher can track their submission's progress and, if necessary, request that CrossRef staff move it up in the queue. When the deposit is successful, the DOI will then start to work and resolve to the URI listed in the resource section of the XML.

Conclusion

Publishers and societies join CrossRef predominantly to allocate CrossRef DOIs to their content. All other CrossRef services such as Cited-by linking, CrossMark, CrossCheck, and FundRef are dependent on the DOI and associated bibliographic metadata. This article covers the basic information for publishers and editors who are starting out with XML and aims to answer some of the basic questions they may have. More complex aspects of CrossRef XML and depositing will be covered in a later article.

Conflict of Interest

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

References

1. CrossRef. Join crossref [Internet]. Lynnfield: Crossref; 2013 [cited 2014 May 10]. Available from: http://crossref.org/01company/join_crossref.html/
2. Jeong GH, Huh Sun. Increase in frequency of citation by SCIE journals of non-Medline journals after listing in an open access full-text database. Sci Ed 2014;1:24-6. http://dx.doi.org/10.6087/kcse.2014.1.24
3. CrossRef. Establishing a DOI suffix pattern [Internet]. Lynnfield: Crossref; 2013 [cited 2014 May 10]. Available from: http://help.crossref.org/#establishing_a_doi_suffix_pattern/
4. CrossRef. Required, recommended, and optional elements [Internet]. Lynnfield: Crossref; 2013 [cited 2014 May 12]. Available from: http://help.crossref.org/#elements/
5. CrossRef. Webdeposit ver. 1.34 [Internet]. Lynnfield: Crossref; 2013 [cited 2014 May 12]. Available from: http://www.crossref.org/webDeposit/
6. CrossRef. Deposit system interface [Internet]. Lynnfield: Crossref; 2013 [cited 2014 May 12]. Available from: http://doi.crossref.org/
7. CrossRef. Metadata quality check [Internet]. Lynnfield: Crossref; 2013 [cited 2014 May 12]. Available from: http://

www.crossref.org/02publishers/parser.html/

8. CrossRef. Crossref test deposit system [Internet]. Lynnfield: Crossref; 2013 [cited 2014 May 12]. Available from: http://test.crossref.org/
9. CrossRef. Error and warning messages [Internet]. Lynnfield: Crossref; 2013 [cited 2014 May 12]. Available from: http://help.crossref.org/#suberrors/