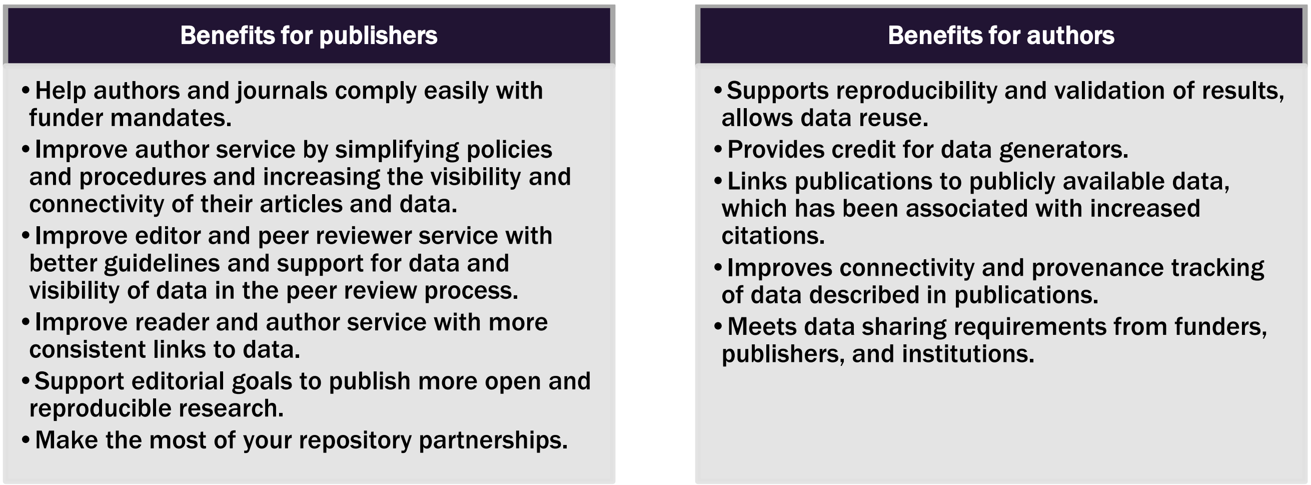
STM Advancing trusted research

Scholix: linking supporting research data to scholarly literature

Introduction

Scholix is a framework that links publications to data, and vice versa, in a simple and persistent way. In a system in which data plays an increasingly important role, adopting the Scholix framework makes data more FAIR and provides several advantages to publishers and authors (Figure 1).



Benefits for publishers

- Help authors and journals comply easily with funder mandates.
- Improve author service by simplifying policies and procedures and increasing the visibility and connectivity of their articles and data.
- Improve editor and peer reviewer service with better guidelines and support for data and visibility of data in the peer review process.
- Improve reader and author service with more consistent links to data.
- Support editorial goals to publish more open and reproducible research.
- Make the most of your repository partnerships.

Benefits for authors

- Supports reproducibility and validation of results, allows data reuse.
- Provides credit for data generators.
- Links publications to publicly available data, which has been associated with increased citations.
- Improves connectivity and provenance tracking of data described in publications.
- Meets data sharing requirements from funders, publishers, and institutions.

Figure 1.The benefits of linking literature to datasets.

This document offers a simple step-by-step guide for publishers who want to support data-to-literature links using the Scholix framework by sharing information with Crossref. Additional information on how to set up Scholix can be found on stm-researchdata.org.

What is Scholix?

Scholix is designed as a way to streamline the process of linking literature to datasets, and for that information to then be exposed for discovery services. It is not a piece of software, but a description of how links between literature and research data should be created and exchanged using a standard format. In simple terms, Scholix sets out how data repositories, publishers and others should provide information about data-to-literature links through‘ community hubs like Crossref and DataCite. The community hubs, in turn, share the link information with each other through a common, open exchange mechanism using the Scholix format. Crossref does this via their Event Data service (Figure 2). This allows the community to build tools like Scholexplorer (scholexplorer.openaire.eu/#/) and DataCite Commons (commons.datacite.org/) that allow users to navigate the connections between scholarly objects.

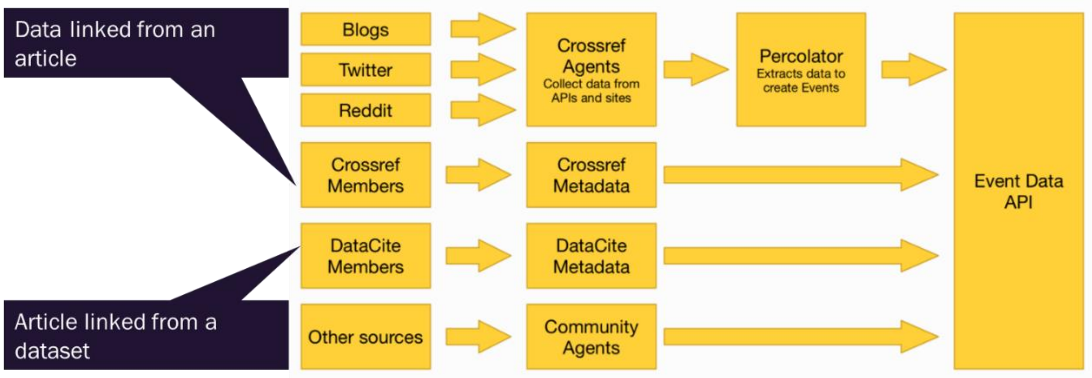


Figure 2. Data-to-literature links from articles or datasets are sent to Crossref or DataCite, as relevant. That information is passed to Event Data, which extracts data-to-literature links into a separate service that lets them be easily discovered.

This document goes into a little more detail on how scholarly publishers can provide Scholix-compliant information through Crossref by sharing the data-to-literature link information as part of the standard content registration process.

Capturing data-to-literature links

Data policies

It is good practice to have a clear editorial policy expressed in the Information for Authors outlining your stance on open data and data-to-literature links. The Research Data Alliance developed a framework containing six policy tiers and 14 features, which is described in detail on the STM Research Data website at stm-researchdata.org/getting-started/.

The policies all recommend the inclusion of a Data Availability Statements (DAS) within the published article. DAS are a simple, consistent way to provide human- and machine-readable information about the availability of data associated with an article, and about how the authors have complied with relevant funder, institution or journal data policies.

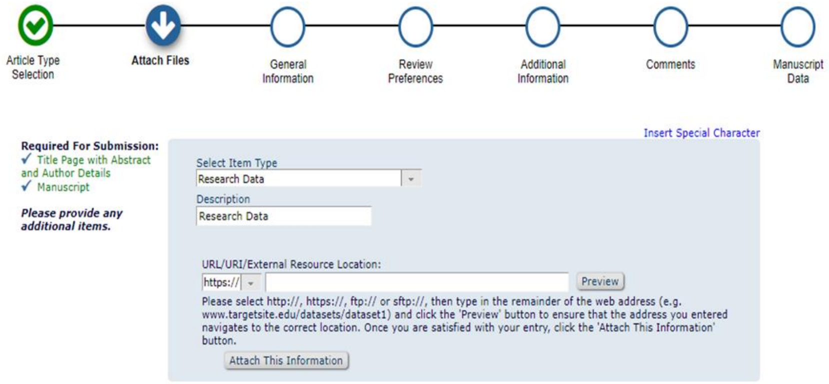
During submission

There are two main methods for capturing information about data-to-literature links during the article submission process:

- As part of the submission form in response to direct questions, by providing a link, DOI, accession number, or other identifier to datasets held in repositories, or providing a DAS (Figures 3 and 4).
- Within the submitted manuscript by providing a DAS, which may include a link or identifier as part of the text, providing data references in the main reference list or in a separate data reference list, or by providing data in appendices or supplementary files associated with the submission.

Data-to-literature links in the submission form

As shown in Figure 3, submission systems can typically be configured to allow authors to specify their datasets during submission. In the case of Aries Editorial Manager, they can do this by specifying what type of link they are creating, providing a title or description for the link, and then adding the URL – in the case of our example, we can see the author is linking to research data. DOIs are also links (e.g. http://doi.org/10.5334/dsj-2020-005), but some systems also allow authors to add the DOI or another identifier without the prefix (e.g. 10.5334/dsj-2020-005).



Required For Submission:
✓ Title Page with Abstract and Author Details
✓ Manuscript

Please provide any additional items.

Select Item Type: Research Data

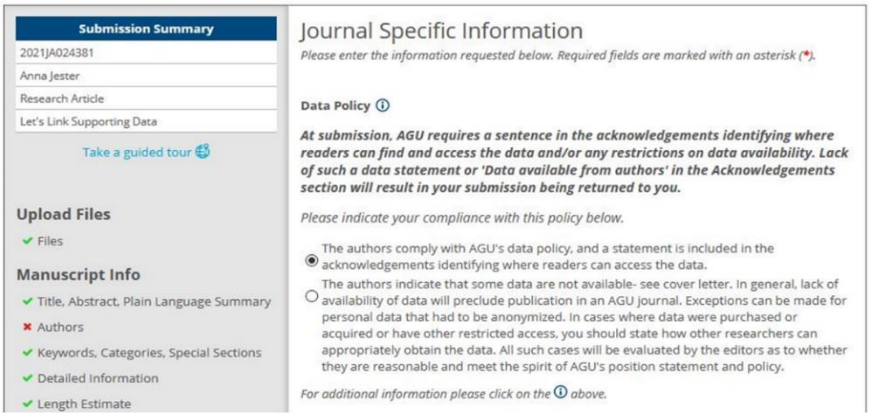
Description: Research Data

URL/URI/External Resource Location: https://

Please select http://, https://, ftp:// or sftp://, then type in the remainder of the web address (e.g. www.targetsite.edu/datasets/dataset1) and click the 'Preview' button to ensure that the address you entered navigates to the correct location. Once you are satisfied with your entry, click the 'Attach This Information' button.

Figure 3. A screenshot from Aries Editorial Manager showing how authors may provide a link to datasets.

The other way for authors to provide information about data-to-literature links is by publishers displaying standard DAS wording and asking authors to confirm which version applies to their article, as shown in the EJ Press example (Figure 4).



Submission Summary

2021JA024381

Anna Jester

Research Article

Let's Link Supporting Data

Upload Files

- ✓ Files

Manuscript Info

- ✓ Title, Abstract, Plain Language Summary
- ✗ Authors
- ✓ Keywords, Categories, Special Sections
- ✓ Detailed Information
- ✓ Length Estimate

Journal Specific Information

Please enter the information requested below. Required fields are marked with an asterisk (*).

Data Policy ⓘ

At submission, AGU requires a sentence in the acknowledgements identifying where readers can find and access the data and/or any restrictions on data availability. Lack of such a data statement or 'Data available from authors' in the Acknowledgements section will result in your submission being returned to you.

Please indicate your compliance with this policy below.

● The authors comply with AGU's data policy, and a statement is included in the acknowledgements identifying where readers can access the data.

○ The authors indicate that some data are not available- see cover letter. In general, lack of availability of data will preclude publication in an AGU journal. Exceptions can be made for personal data that had to be anonymized. In cases where data were purchased or acquired or have other restricted access, you should state how other researchers can appropriately obtain the data. All such cases will be evaluated by the editors as to whether they are reasonable and meet the spirit of AGU's position statement and policy.

For additional information please click on the ⓘ above.

Figure 4. A screenshot from EJ Press showing how authors may indicate data policy compliance through a pre-configured DAS.

Where information about data-to-literature links is provided by the author as part of the submission form, the system needs to be configured to export that information to the production system or vendor as part of the article metadata package at acceptance. This is something your operations team or system supplier will be able to do for you.

Data-to-literature links in the manuscript file

In many cases, authors will have included a DAS in their manuscript file or potentially cited datasets and included the information in their reference lists. You may wish to include an instruction to this effect in your Information for Authors pages, and potentially include the presence of a DAS in journal editorial office checks that happen at submission.

Where information about data-to-literature links is provided within the manuscript file, no changes are required to ensure that this information is provided to the production system or vendor.

Production

If you are marking up articles in the JATS XML schema, you can add information to your DAS and to data references that will make the data-to-literature links explicit. Full instructions for marking up DAS in JATS 1.2 can be found online at jats.nlm.nih.gov/archiving/tag-library/1.2/chapter/tag-data-avail.html.

In the example given here, the DAS includes a link to a dataset in Figshare that has been marked up as a reference.

```
<back>
...
<sec sec-type="data-availability">
  <title>Data Availability Statement</title>
  <p>The data analysis file is available in Figshare:</p>
  <ref-list>
    <ref id="data001">
    <label>D1</label>
    <element-citation publication-type="data" specific-use="isSupplementedBy">
    <name><surname>Garner</surname><given-names>Jane</given-names></name>
    <name><surname>Wakeling</surname><given-names>Simon</given-names></name>
    <name><surname>Jamali</surname><given-names>Hamid R</given-names></name>
    <name><surname>Hider</surname><given-names>Philip</given-names></name>
    <name><surname>Mansourian</surname><given-names>Yazdan</given-names></name>
    <name><surname>Lymn</surname><given-names>Jessie</given-names></name>
    <name><surname>Randell-Moon</surname><given-names>Holly</given-names></name>
    <data-title>Australian Public Library COVID Survey</data-title>
    <source>Figshare</source>
    <year iso-8601-date="2021">2021</year>
    <pub-id pub-id-type="doi" assigning-authority="figshare"
    xlink:href="https://doi.org/10.6084/m9.figshare.14183060.v1">
    https://doi.org/10.6084/m9.figshare.14183060.v1</pub-id>
    </element-citation>
    </ref>
  </ref-list>
</sec>
```

The use of the `@publication-type` attribute allows publishers to specify that a reference is to a dataset, but as `@publication-type` is optional publishers may also choose to specify that a reference is to a codeset, a methodology, etc. Similarly, the `@specific-use` attribute is flexible enough to specify the types of relationship between literature and datasets. The examples above use `@specific-use=“IsSupplementedBy”`, but you could specify that an article simply references a dataset by using `@specific-use=“References”`.

This inclusion of a data reference in the DAS makes it very simple for humans to access the information they need, while also facilitating machine readability. Another option is to provide data-to-literature links as part of the article reference list – that is, to treat the dataset in the same way as any other citation. Finally, some journals provide a separate data reference list, marked up in the same way as the main article reference list. Regardless of inclusion in the main reference list or a data reference list, we recommend you require a DAS in every article.

Note: References to datasets may be marked up as `<mixed-citation>` instead of `<element-citation>`, if that is your production team’s preferred style, either within the DAS or in a reference list, per the example below:

```
<ref-list>
  ...
  <ref id="27">
  <label>27</label>
  <mixed-citation publication-type="data" specific-use="isSupplementedBy">
    <name><surname>Garner</surname><given-names>Jane</given-names></name>
    <name><surname>Wakeling</surname><given-names>Simon</given-names></name>
    <name><surname>Jamali</surname><given-names>Hamid R</given-names></name>
    <name><surname>Hider</surname><given-names>Philip</given-names></name>
    <name><surname>Mansourian</surname><given-names>Yazdan</given-names></name>
    <name><surname>Lymn</surname><given-names>Jessie</given-names></name>
    <name><surname>Randell-Moon</surname><given-names>Holly</given-names></name>
    <data-title>Australian Public Library COVID Survey</data-title>
    <source>Figshare</source>
    <year iso-8601-date="2021">2021</year>
    <pub-id pub-id-type="doi" assigning-authority="figshare"
    xlink:href="https://doi.org/10.6084/m9.figshare.14183060.v1">
    https://doi.org/10.6084/m9.figshare.14183060.v1</pub-id>
  </mixed-citation>
  </ref>
  ...
</ref-list>
```

Crossref

While JATS markup is important, it is only by sending data references to Crossref that data-to-literature links from your content will appear in services based on the Scholix framework. How you do this will be guided by where your data-to-literature links fall within the article.

Full instructions for Crossref markup can be found at crossref.org/documentation/reference-linking/data-and-software-citation-deposit-guide/.

Data-to-literature links structured as references

Whether your data-to-literature link appears in the DAS, the main article references list or a separate data reference list, if you have marked it up using the JATS structure listed above you will be able to share the necessary information with Crossref.

Note that your Crossref deposits might be sent directly from your production vendor, from your publication platform, or by your in-house team.

There are two options for sharing these data references. The first is to convert your structured markup from JATS to the Crossref schema like this:

```
<citation_list>
  ...
  <citation key="27" publication-type="data">
    <doi>10.6084/m9.figshare.14183060.v1</doi>
    <cYear>2021</cYear>
    <author>Garner J, Wakeling S, Jamali HR, Hider P, Mansourian Y, Lymn J, Randell-Moon H</author>
    <data_title>Australian Public Library COVID Survey</data_title>
  </citation>
  ...
</citation_list>
```

Alternatively, you can choose to provide unstructured citations as shown below.

```
<citation_list>
  ...
  <citation key="27" publication-type="data">
  <doi>10.6084/m9.figshare.14183060.v1</doi>
  <unstructured_citation>Garner J, Wakeling S, Jamali HR, Hider P, Mansourian Y, Lymn J, Randell-
  Moon H. “Australian Public Library COVID Survey.” Figshare 2021</unstructured_citation>
  </citation>
  ...
</citation_list>
```

Note that the capacity to provide information on the type of publication (`publication-type=“data”`) will be supported in their upcoming schema release. Data-to-literature links can currently be provided without this tag, and will still be correctly identified by Crossref, provided the <doi> of the dataset is tagged in the metadata as shown above.

Alternative markup

The previous section describes how to mark up data references for Crossref based on their upcoming 2021 schema release.

An alternative method that is also supported is to assert a relationship within the metadata sent to Crossref, as shown here:

```
<related_item>
  <description>Data Availability Statement. The data analysis file is available in
  Figshare.</description>
  <inter_work_relation relationship-type="data" identifier-type="doi">10.6084/m9.figshare.14183060.
  v1</inter_work_relation>
</related_item>
```

Establishing data and software citations via the attribute `@relationship-type` enables precise tagging of the dataset and its specific relationship to the research results published. To tag the data and software citation in the metadata deposit, Crossref asks for the description of the dataset and software (optional), dataset and software identifier and identifier type (DOI, PMID, PMCID, PURL, ARK, Handle, UUID, ECLI, and URI), and relationship type.

In general, Crossref suggests using the `relationship-type=“references”` for data and software resources. To specify that the data or software resource was generated as part of the research results, use `relationship-type=“isSupplementedBy”`. Being this specific is optional, but can support scientific validation and research funding management.