NCBI Bookshelf. A service of the National Library of Medicine, National Institutes of Health.

Journal Article Tag Suite Conference (JATS-Con) Proceedings 2016 [Internet]. Bethesda (MD): National Center for Biotechnology Information (US); 2016.

ORCID iD Throughput in Publishing Workflows

Laurel Haak,[1] Paul Donohoe,[2] Véronique Kiermer,[3] Helen Atkins,[4] John Lees-Miller,[5] and Craig Raybould[6].

[1] Executive Director, ORCID
l.haak@orcid.org
[2] Senior Developer, SpringerNature
Paul.Donohoe@macmillan.com
[3] Executive Editor, PLOS
vkiermer@plos.org
[4] Director, Publishing Services, PLOS
hatkins@plos.org
[5] Co-founder, Overleaf
john.lees-miller@overleaf.com
[6] Chief Process Engineer, Hindawi Publishing Corporation
craig.raybould@hindawi.com

ORCID iDs are unique persistent identifiers for authors and other contributors in the research community, part of a community effort to support authenticated connections between authors and their works and thereby address the name ambiguity problem in scholarly communications. In this paper, we will examine use cases from organizations who are early adopters of ORCID iDs, all of whom have an ORCID throughput to Crossref; each of these organizations use different publishing platforms and manage XML flows differently. Our goal in this paper is to identify and improve community awareness of effective metadata management practices.

Overview

Increasingly, publication processes are relying on the use of an identifier infrastructure to enable workflows and information processing. Starting with Crossref (crossref.org), which provides a persistent identifier for the paper itself—and connects to an infrastructure to ensure persistent access to the source document—identifiers are becoming an accepted and, in some cases, required component of publishing. DataCite (datacite.org) provides a persistent identifier for datasets, and ORCID (orcid.org) for researchers. There are also identifiers for organizations, samples (http://www.geosamples.org/aboutigsn), reagents (RRI, https://scicrunch.org/resources), and equipment (see Southampton University example, http://orcidpilot.jiscinvolve.org/wp/hei-based-projects/) in various stages of adoption by the community.

Identifiers support persistent and unique identification of the target item, whether that is a person, a location, or an object. In combination with XML-formatted structured fields and APIs, they can be used to replace typed data entry, thus ensuring standard cross-platform expression of items. There are several challenges in using identifiers. There is a hesitation to adopt identifiers until a sufficient proportion of the community is using them, and about creating a burden for authors. There is concern about interactions between identifier schemes for the same item-for example, identifiers for organizations, of which several are currently used in publishing but are not interoperable, diminishing the effectiveness of the schemes. And, there is a concern about long-term access to identifiers and related metadata, both in terms of access to proprietary schemes and stability of the organization providing the identifier.

In addition to these issues, is a question about mechanics: when should identifiers be inserted into publication processes, how, and by whom. In this paper we examine these questions, using ORCID identifiers as a case-in-point. We describe current workflows at Hindawi, Nature, and PLOS, discuss gaps, and examine opportunities provided by enabling technologies such as the XML-based manuscript preparation system provided by Overleaf to improve the collection of identifiers.

Publishing practice

In the majority of cases, authors submit a text document, formatted to publisher specifications but otherwise unstructured. Author and organization names are listed as text strings; reagents, samples, and equipment are embedded in methods sections; and funding support is usually listed in acknowledgement sections. Without structure, it is difficult to discover these items within the text of the document. The resulting ambiguity makes it more difficult to use the information in the paper, with implications for accurate attribution, metrics, and reproducibility of research.

Publishers are looking to address these issues by adopting information processing standards. Already, many publishers have implemented XML-based workflows for processing manuscripts, and are using standard expressions, such as JATS, to describe paper metadata. Within this structure, publishers are starting to use ORCID iDs, a standard identifier for researcher names.

Collecting ORCID iDs from authors

As a researcher-controlled identifier, ORCID iD collection requires the author to register for and approve collection of their identifier using an API-mediated OAuth permission process. The authenticated iD method involves placing a button on the publisher site, and when the user clicks it, (s)he authenticates to ORCID by typing his/her username and password, and ORCID passes back to the publisher system the authenticated ORCID iD associated with that individual. This ensures that the person and the identifier are associated, and that the identifier is collected without typographical errors. It provides the opportunity to ask for permissions to transfer data between systems, important given differing national privacy regulations. In addition, this method being used by universities, funders, as well as publishers, so the process is getting more and more familiar to researchers. The registration/login process has been streamlined to enable swift completion, for ease of use when inserted into existing workflows such as manuscript submission. Authors can register and/or authenticate in about 30 seconds. More details about this process (with screen shots) can be found here, and can be performed with either the public or member ORCID API.

Manuscript submission systems at Nature, PLOS, and Hindawi allow corresponding authors to connect their ORCID iD to their local author profile. Using an API-mediated process kicked off when an author is registering or updating their profile on the publisher's site, the author logs in or registers on the ORCID site and this returns the validated ORCID iD and a permission token to the publisher. The identifier and associated validated flag are then attached to the author profile in the publishing system, and through that to papers that the author submits. PLOS subsequently allows the author to sign-in to the submission platform using their ORCID login credentials.

Some manuscript submission platforms allow the corresponding author or editorial staff to mediate the collection of co-author ORCID iDs; the methods for this vary from a non-advised manual entry of ORCID iD by or on behalf of the co-author to a messaging system that prompts the co-author to use an API-mediated collection process. Particularly problematic is a search-fetch method. Name-based searches do not always return expected results, for all of the reasons for which ORCID was founded in the first place: common names, changed names, name variations. In addition, only information marked as public may be available for such a search, it can be quite difficult even for the iD-holder to discern which ORCID iD belongs to him/her. It is too easy for someone to mistakenly choose the wrong iD, thinking that it is theirs. When an incorrect iDs becomes part of sources that are considered authoritative, the entire community puts in doubt the validity of the iDs association to an individual, greatly negating the benefit of having these unique iDs.

The authenticated iD method has you placing a button on your site, and when the user clicks it, (s)he authenticates to ORCID by typing his/her user name and password, and ORCID passes back the authenticated ORCID iD that is associated with that individual. The benefits are several - you know the individual is in control of the iD, and are sure that the iD used is the one that the user is actively using. In addition, this is the same method being used by publishers, universities and other funders, so the process is getting more and more familiar to researchers. More details about this process (with screen shots) can be found here, and can be performed with our public API: http://members.orcid.org/funder-workflow#application.

For all authors, voluntary provision of ORCID iDs has been running at less than 10% of the submitting population. To raise awareness and adoption, several publishers, including Nature, PLOS, and Hindawi, are planning to require corresponding authors provide their ORCID ID during the submission process, some at time of submission and others at acceptance.

Adding identifiers into manuscripts

Some publishers are starting to support manuscript submission in XML format. This pushes the collection of identifiers forward into the manuscript writing process. In this scenario, authors using an XML-based text editor can connect their ORCID iD to their paper prior to submitting it to a journal.

One example of such a system is Overleaf, an online LaTeX-enabled text editor that supports real-time collaboration and produces typeset output in the background as the author is typing. Authors who have access to the document can connect their ORCID iD, affiliation data, and datasets. An overlay system allows authors to manage author and affiliation names. Overleaf uses ORCID's API permission and authentication workflow, ensuring that each author can connect only their own ORCID iD. Authors who have access to the paper can remove and reorder author records using the Author Overlay (see Figure 1), which is a graphical editor for authorship information embedded directly in Overleaf's editable rich text view of the article. If an author has not yet linked their ORCID iD, they can do so directly from the manuscript via the Author Overlay.

At some journals, authors can submit an XML-formatted paper directly from the text editor into a manuscript submission system; such functionality is supported by ScholarOne, Editorial Manager, and eJournalPress. Publisher- and journal-specific pre-submission checks can automatically determine whether required metadata, including ORCID iDs, are present for the manuscript, and alert the corresponding author if any checks fail. Once submission requirements are met, article content and article metadata can be passed directly into the manuscript tracking system for final review and approval by the author. The exact format in which the data are passed can be customized on a per-publisher and per-journal basis. Overleaf passes article metadata to Editorial Manager and eJournalPress in JATS XML. In particular, the ORCID iD is passed using the contrib-id element with type "orcid". The @authenticated flag may also be passed. For example:

```
<contrib contrib-type="author">
    <contrib-id contrib-id-type="orcid">http://orcid.org/0000-0003-3492-8854</contrib-id>
</contrib>
```

If authors are later asked to resubmit following review, article metadata, including ORCID iDs, can be preserved from the first submission and updated in each later submission, using the same processes.

Passing identifiers into production systems

For accepted manuscripts, associated ORCID iDs are exported into the production system. Obtaining the ORCID iD at submission or acceptance means publishers can automate more of the production processes. Below is shown the standard XML used by PLOS to export contributor metadata to production.

```
<contrib contrib-type="author">
    ...

<contrib-id contrib-id-type="orcid" specific-use="authenticated">0000-0002-4565-0280</contrib-id>
    ...
</contrib>
<contrib contrib-type="author">
    ...
    <contrib-id contrib-id-type="orcid">0000-0000-0000-1111</contrib-id>
    ...
</contrib>
```

For papers submitted as PDF or Word documents, the collection of identifiers occurs in parallel to manuscript submission and is not integrated into the manuscript until after acceptance. For papers submitted in XML, identifiers can flow through systems as one with the manuscript.

At Nature, author metadata including ORCID IDs is exported (via e-Journal Press) as an XML file separate from the manuscript metadata file, which is then included in the article package sent to typesetters, who convert the package into a composed article XML file. PLOS employs a similar workflow, using Editorial Manager. The validation flag and the ORCID iD are included in the metadata export file, which is sent to their composition/XML vendor. All of the authors must sign off on any authorship changes. PLOS has indicated it will ensure any new corresponding author will provide an authenticated ORCID iD once their ORCID requirement is effective. Using an automated check, Hindawi imports an ORCID ID into the word file if the author name in the manuscript matches the name in author's online ORCID record. Currently, if an author does not have an ORCID ID, Hindawi encourages them to register when sending the galley proofs. A finalization team checks certain points on the article prior to publication. The team checks if the author has added an ORCID iD to their account and will manually add it to the article if present. Upon acceptance, the system will automatically check again to ensure the ORCID iD present in the article is correct and the name is matched on the ORCID registry. If it is not, the ORCID iD is removed. Following implementation of their ORCID requirement, Hindawi will not transfer the article from their peer review system to production without an ORCID iD on the corresponding author's account. This will remove the manual steps above.

At this point it should be reiterated that names for an individual may change or be expressed with variation across their career, so it is important for any checking process to ensure that the identifier was collected using an authenticated pathway rather than a fetch or manual entry process.

Indexing identifiers

When the production team finishes work on the manuscript it is converted to JATS XML format, which is used for the creation of the PDF, HTML, and EPUB. At PLOS, the composition/XML vendor includes both authenticated and non-authenticated ORCID IDs from the submission system export in the `<contrib><contrib-id>` markup of the article XML. The default @authenticated value is "false", so is excluded for non-authenticated ORCID iDs, and @authenticated="true" is applied for authenticated ORCID iDs. If an author with an ORCID iD is also a principal award recipient in the funding data, they identify the ORCID iD in the `<principal-award-recipient><contrib-id>` elements as well.

The post-publication distribution of article data to third-party abstracting, indexing and distributing organizations is fully automatic. ORCID markup is unalterered from the JATS article XML. Article metadata, including the ORCID iDs, are extracted and XML files in CrossRef and PubMed formats are created. Crossref requires all ORCID iDs to express an @authenticated="true" or @authenticated="false" designation. PubMed's DTD does not support an authentication flag, and authenticated and non-authenticated ORCID iDs are passed to PubMed without distinction. These files are packaged and distributed within an hour of publication (PLOS and Nature) or daily (Hindawi) to PubMed, CrossRef, and other recipients. Figure 2 shows the transformation of ORCID markup as it passes from the PLOS composition vendor to its syndicates.

Along with general indexing preparation, publishers should ensure that validated identifiers are included in their Crossref metadata deposit. Including the author's ORCID iD enables auto-update of the their ORCID record. This benefits researchers: they need only include their iD at manuscript submission (enter once), and grant permission to Crossref to update their record once, and then information about the published work -- with their iD -- can easily flow into connected systems, such as funder reporting systems and university repositories (re-use many times). Including the ORCID identifier in usual metadata deposits acts as a default "opt-in" for auto-updating ORCID records. ORCID identifiers are included in Crossref deposits along with author information, for example:

```
<contributors>
    <person_name sequence="first" contributor_role="author">
        <given_name>Josiah</given_name>
        <surname>Carberry</surname>
        <ORCID>http://orcid.org/0000-0002-1825-0097</ORCID>
    </person_name>
 </contributors>
```

An ORCID identifier may only be deposited as part of a Crossref metadata deposit, using their web deposit form. Publishers are encouraged to redeposit previous content to include ORCID identifiers if they're collecting these retrospectively.

Gaps and next steps

In general, processes for collecting ORCID iDs have been identified and implemented. Integration into accepted manuscripts and conversion into XML-based production and syndication have been defined. A description of best practice has been defined as a component of the recent publisher initiative to require ORCID identifiers in the manuscript submission and review workflow, which includes guidelines for collecting identifiers, display in published manuscripts, assertion of authorship by the publisher, and connection of this assertion with the author's ORCID record.

There remain some challenges. One, as mentioned above, is the relatively low use of ORCID iDs by authors in the manuscript submission process. This is due in part to the need to increase awareness and understanding of ORCID in the researcher community. With some high-profile journals starting to require ORCID iDs and the launch of Crossref-mediated auto-update processes, as well as clear and concise messaging about the benefits of use to authors, adoption is increasing. Coupled with these outreach efforts is educating authors and streamlining their experience, so that they know what to expect during the submission process: what metadata to include and why-whether that is their ORCID iD, funding information, datasets, or identifiers for their research reagents-and when and where to include the information.

Another challenge is the continued use of type-in fields and fetch-searches of the ORCID registry to populate submission forms. With many manuscript tracking system vendors supporting authenticated collection of identifiers, and ORCID providing authentication as a component of its public API, it is becoming easier for journals to do the right thing. Further, the use of ORCID single-sign-on (SSO) in manuscript systems is providing an easy way to both show benefit to authors and encourage use of authentication. There remains uncertainty over use of the "authenticate" flag when using ORCID iDs, and there is a lack of specification for this flag in PubMed syndication.

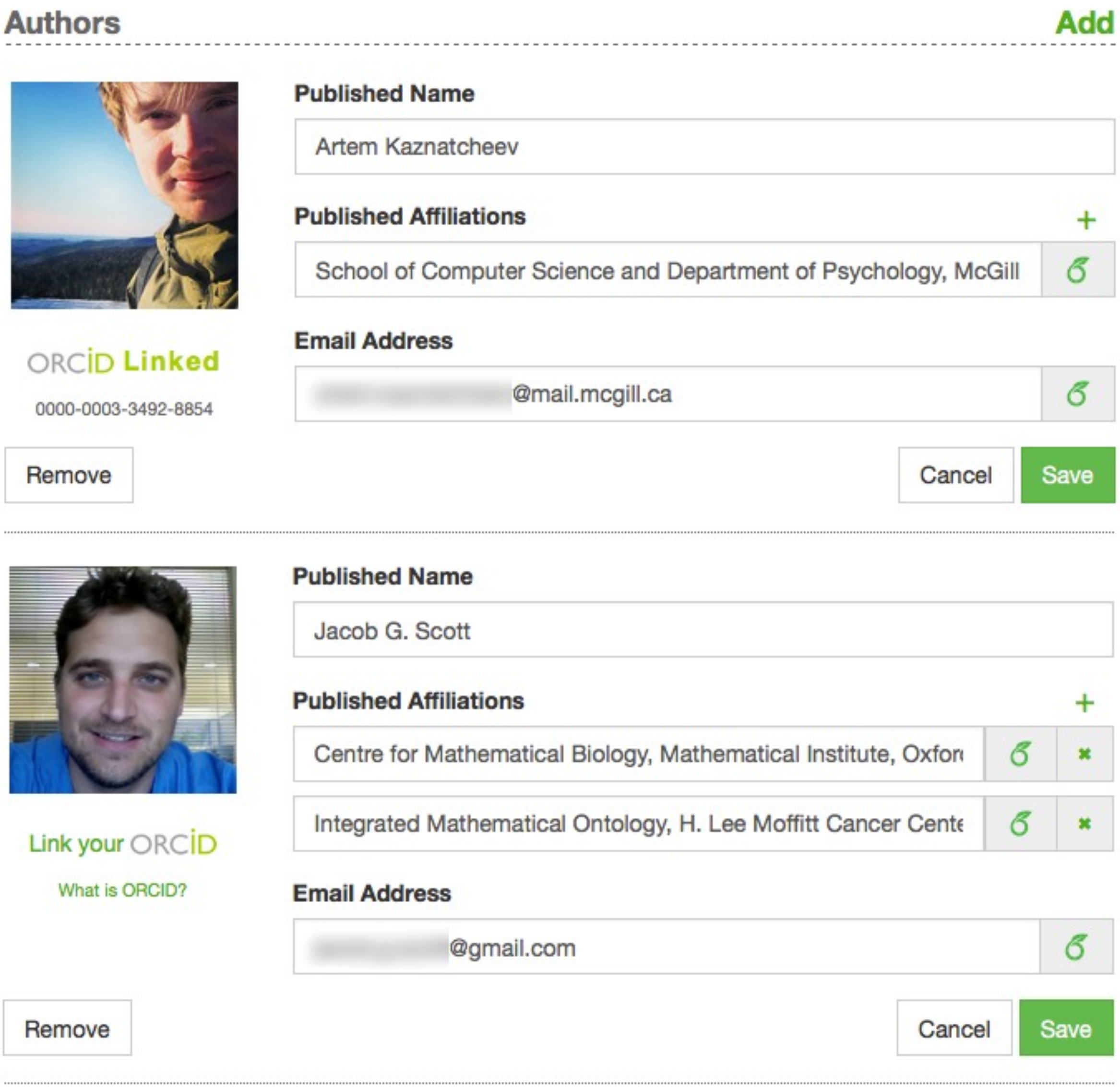
In addition, there are the challenges of co-authors. When should their data be collected? From one perspective, it makes sense to collect information at the time the manuscript is submitted, to ensure that all authors agree to be responsible for the paper, but some see this as adding undue burden on authors. Others argue that the best time to gather information -including identifiers for the co-authors, affiliation, and contributor role-is at the time of manuscript acceptance. Either scenario has implications for how to handle authorship order changes. This also points to the need to connect authors and identifiers as discrete fields in an XML document tied to a back-end database, rather than flat text fields in a PDF or word document. Conversion from flat file to XML in handoff between acceptance and production can cause numerous problems, including improper assignment of identifiers.

All of that aside, a broader concern is how these identifiers - which can help streamline the publication process, improve discoverability, and be leveraged for processes including rights management and open access compliance - can be effectively incorporated into the publication process. At present, the focus has been collection at the point of manuscript submission. This stacks up the burden on authors at one critical point, rather than at the point an item is used or involved in the research process (Figure 3).

Identifier collection could be enabled through XML workflows as research is carried out, with e-lab notebooks, for example, and XML-enabled manuscript editors (such as the Overleaf case study in this paper). This would reduce the single-point collection burden on researchers, and might improve adoption and use by connecting the collection process into the research process itself. It would also better enable linkages with funding and co-authors.

Identifiers are a package deal: they need to be integrated into standard research workflows in a way that reinforces a familiar user experience, does not add to researcher burden, uses best practices including authenticated collection, and provides benefits to all parties in the form of increased data processing automation, reduced burden, and improved search accuracy and discoverability. Together, these steps would improve author attribution, trust, discoverability, and also support the unambiguous connections between researchers and their research, affiliations, and funding desired and needed by the community.

Figures



0000-0003-3492-8854

Published Name
Artem Kaznatcheev

Published Affiliations
School of Computer Science and Department of Psychology, McGill

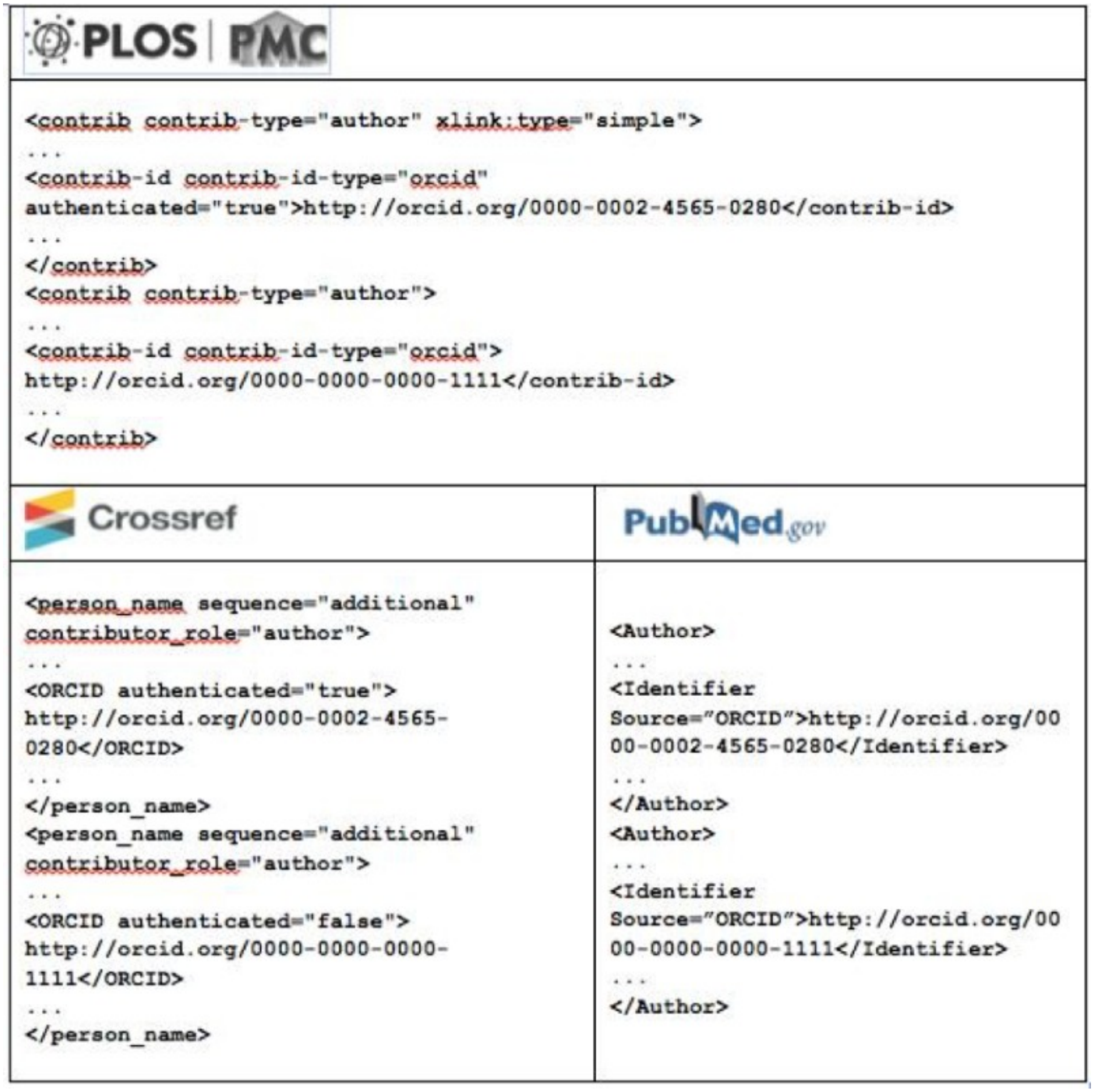
Email Address
[redacted]@mail.mcgill.ca

Published Name
Jacob G. Scott

Published Affiliations
Centre for Mathematical Biology, Mathematical Institute, Oxfor
Integrated Mathematical Ontology, H. Lee Moffitt Cancer Cente

Email Address
[redacted]@gmail.com

Fig. 1 An author can link their ORCID IDs to their Overleaf account using Overleaf's Author Overlay



PLOS | PMC

```
<contrib contrib-type="author" xlink:type="simple">
...
<contrib-id contrib-id-type="orcid"
authenticated="true">http://orcid.org/0000-0002-4565-0280</contrib-id>
...
</contrib>
<contrib contrib-type="author">
...
<contrib-id contrib-id-type="orcid">
http://orcid.org/0000-0000-0000-1111</contrib-id>
...
</contrib>
```

Crossref

```
<person_name sequence="additional"
contributor_role="author">
...
<ORCID authenticated="true">
http://orcid.org/0000-0002-4565-
0280</ORCID>
...
</person_name>
<person_name sequence="additional"
contributor_role="author">
...
<ORCID authenticated="false">
http://orcid.org/0000-0000-0000-
1111</ORCID>
...
</person_name>
```

PubMed.gov

```
<Author>
...
<Identifier
Source="ORCID">http://orcid.org/00
00-0002-4565-0280</Identifier>
...
</Author>
<Author>
...
<Identifier
Source="ORCID">http://orcid.org/00
00-0000-0000-1111</Identifier>
...
</Author>
```

Fig. 2 ORCID markup from composition to syndication

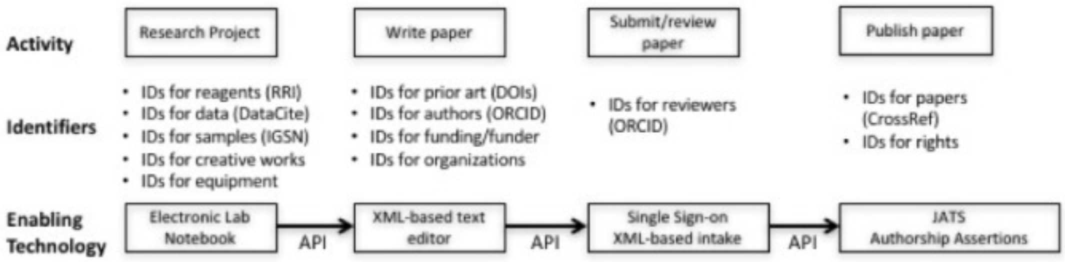


Figure 3. Identifiers and enabling technologies in research publication workflows. Collecting information on the components of a research project can be streamlined if the collection point is pushed closer to the act of creation or use, shown **above**. This is very different than current practice shown **below**, in which identifiers are collected at the point of manuscript submission, which may be months to years after the described activities have taken place.

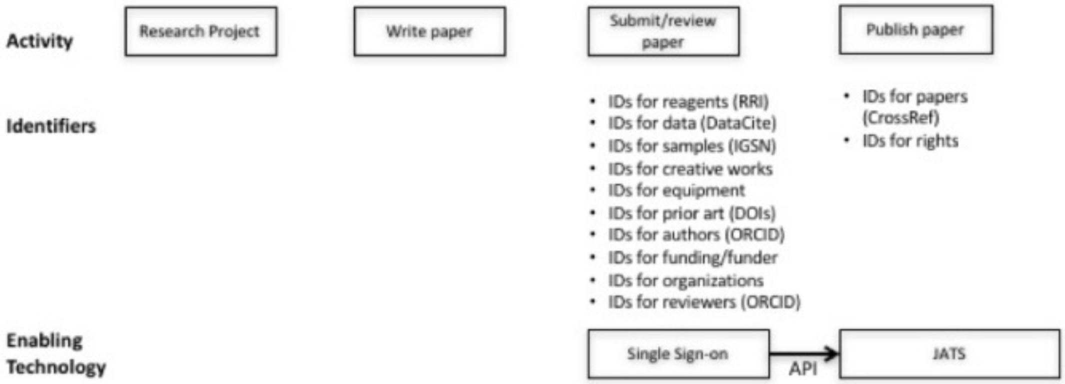


Fig. 3 Identifiers and enabling technologies in research publication workflows.

Bookshelf ID: NBK350150