Front Matter

DOI registration workflow for a science blog

Martin Fenner ⓘD

Published September 22, 2023

Citation

Fenner, M. (2023). DOI registration workflow for a science blog. *Front Matter*. https://doi.org/10.53731/w6nzs-jta75

Keywords

News, Rogue Scholar

Front Matter

In previous blog posts such as the one published earlier this week, I discussed the various elements involved in registering a DOI for a science blog post. Briefly, the Rogue Scholar service takes advantage of the fact that blogs

- use RSS feeds (or the Atom or JSON Feed format) to distribute content and metadata at the time of publication,
- these feeds contain the most important metadata needed for publication – such as title, authors, publication date, and
- addition metadata (such as abstract and references) can be automatically extracted from the full-text content included in the feed.

This basic workflow can be optimized in many ways, for example including funding information (watch out for a blog post next week), but one fundamental issue remains to be solved: how does the blog learn about the DOI registered for a new post, and automatically adds it to the blog?

Canonical URL

As much as possible Rogue Scholar takes advantage of technologies that have existed for a long time and are not specific to scholarly content. That's why the service works with existing blogs that use standard blogging software - currently eleven different platforms, the most popular being Wordpress, Ghost, and Hugo.

These platforms don't know about DOIs without extra work, but they all know about a similar concept: canonical URLs. Wikipedia explains:

> A **canonical link element** is an HTML element that helps webmasters prevent duplicate content issues in search engine optimization by specifying the "canonical" or "preferred" version of a web page. It is described in RFC 6596, which went live in April 2012.

The problem canonical URLs are addressing is duplicate content at different locations that can confuse search engines such as Google or Bing. This is related to the problem persistent identifiers such as DOIs are addressing for the scholarly community: accessing content over long periods of time that may change its location on the web (its URL), with two inter-related strategies:

- **URL redirection**. DOIs redirect to a target URL that can be changed by the publisher,
- **Persistence**. The publisher of scholarly content makes an extra effort to make sure content doesn't disappear (link rot), or significantly change (content drift).
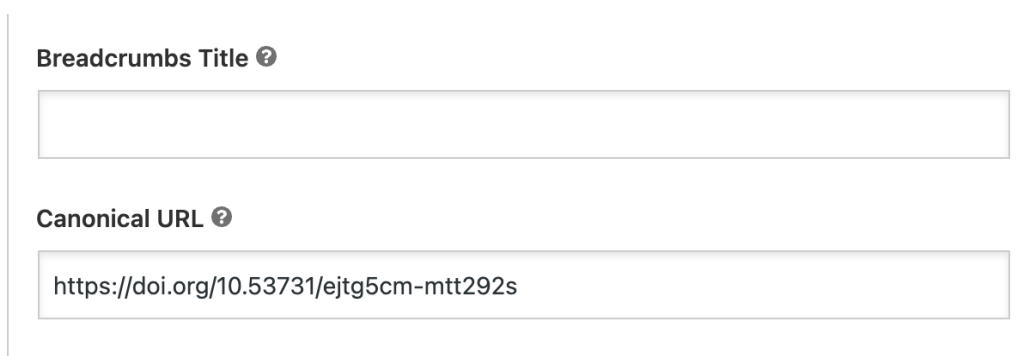
Obviously, canonical URLs are not DOIs, but they provide a standard way for a science blog to add a DOI to a post.

Backends

Science blogs provide a backend to store content and metadata, including the canonical URL. This can either be a database (as in the case of Wordpress or Ghost) or a file (as in the case of Hugo and many other static site generators).

Wordpress

Wordpress doesn't know about canonical URLs out of the box, but they can be added via a plugin, the most popular for this being Yoast SEO (which comes in free and paid versions). After installing and activating the plugin you can add a canonical URL in a new Yoast SEO section of the post editor:
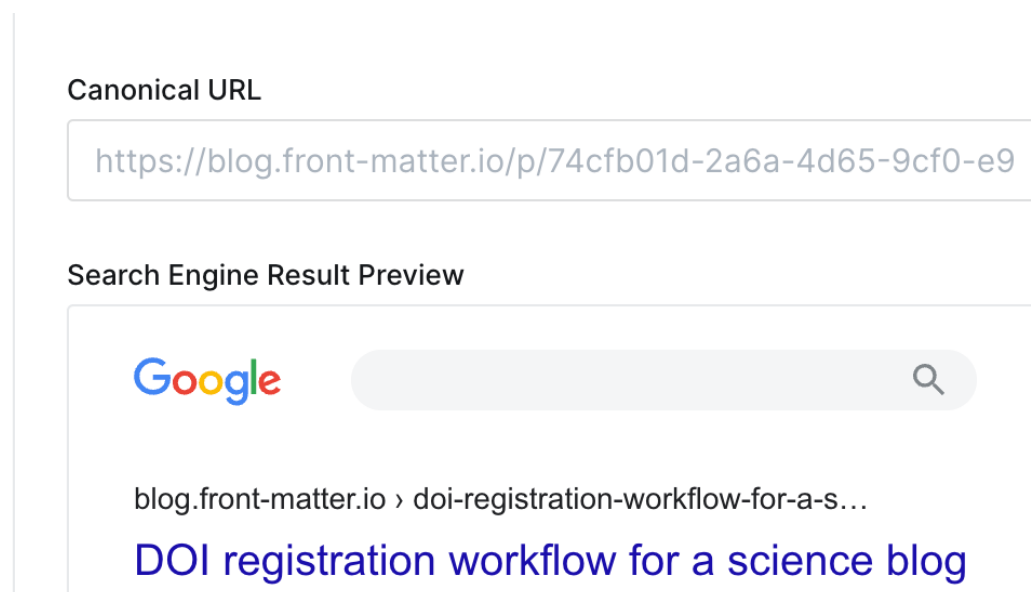

Alternatively, you can fiddle with your Wordpress configuration to add a custom field for the canonical URL.

Ghost

The Ghost blogging platform has a canonical URL field for every post, which you can access from the post settings sidebar:

Hugo

Hugo and other Open Source static site generators give you a lot of flexibility with metadata. If you add a `canonicalUrl` field to the blog post Front Matter, you can reuse it for the canonical URL (with some additional work).

The canonical URL or DOI is now stored with the blog post, but also exposed to web crawlers. The format is `<link rel="canonical" href="`https://doi.org/10.53731/gvb08-7kc16`">`.

Frontends

To display the canonical URL aka DOI on your blog frontend, you have to modify your blog theme, the popular themes for Wordpress, Ghost, and Hugo don't really support displaying the canonical URL out of the box, as they are primarily intended for web crawlers and not humans.

You should follow the Crossref DOI display guidelines, when thinking about how to display the DOI for your blog post, i.e. always be displayed as a clickable full URL link. Rogue Scholar displays DOIs like this:

Streamlining the archiving of science blog posts
Published September 19, 2023 in Front Matter
Author Martin Fenner (iD)

(doi) https://doi.org/10.53731/gvb08-7kc16   📑 BibTex   📄 Citation (APA)

This blog (using the Ghost platform) displays DOIs like this in a sidebar:

Sep 19, 2023 · 2 min

https://doi.org/10.53731/gvb08-7kc16

DOI registration workflow

The changes to the backend and frontend explained above are good enough for occasional blog posts or to get started with Rogue Scholar. After a blog post is published, Rogue Scholar will register a DOI within 30 minutes and show that DOI on the website or via API. You can then copy/paste that DOI into your new canonical URL field. A simple improvement would be notifications of new DOI registrations by email, similar to what Crossref is sending to Front Matter as the Crossref member:

Front Matter

```xml
<?xml version="1.0" encoding="UTF-8"?>
<doi_batch_diagnostic status="completed" sp="ds5">
    <submission_id>1590342900</submission_id>
    <batch_id>8a637b09-fda6-4980-baa1-147497683bd9</batch_id>
    <record_diagnostic status="Success">
        <doi>10.53731/w6nzs-jta75</doi>
        <msg>Successfully added</msg>
        <citations_diagnostic>
            <citation key="ref1" status="resolved_reference">10.53731/
            gvb08-7kc16</citation>
            <citation key="ref2" status="resolved_reference">Cite to nonCR doi:
            10.5281/zenodo.1324300</citation>
            <citation key="ref3" status="resolved_reference">10.1371/
            journal.pone.0115253</citation>
            <citation key="ref4" status="resolved_reference">10.59350/p000s-
            pth40</citation>
            <citation key="ref5" status="resolved_reference">10.53731/
            r79x921-97aq74v-ag5a2</citation>
        </citations_diagnostic>
    </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>
```

But maybe including a clickable link to the DOI just registered and some basic metadata that were registered (as it takes a few hours until the metadata show up in the Crossref REST API).

For blogs with a more frequent publication frequency (e.g. weekly or daily) this workflow should be automated. One important consideration is whether the blog should know the DOI that will be registered in advance, avoiding the round trip with Rogue Scholar and Crossref, and allowing customizations of the DOI name, such as `10.53731/front-matter.2023-09-19`. The biggest advantage would be that the DOI name can be shared in advance of publication, e.g. for press releases, or to reference in other content.

While these considerations are reasonable and not new for DOIs in general, for the science blog use case the workflow should be simple and I want to follow these principles:

- Rogue Scholar DOIs will be generated as a short random 10-character string upon DOI registration. Rogue Scholar users or staff can't modify the DOI names that will be generated. Rogue Scholar DOIs are cool DOIs.
- If you see a Rogue Scholar DOI, it can be used (immediately as a link, accessing the metadata after a few hours). Rogue Scholar is not offering DOIs that are not or not fully registered, i.e. DOIs for pending publications (Crossref) or draft DOIs (DataCite).

Front Matter

- DOI registration happens with the Rogue Scholar service talking to the Crossref API, participating blogs don't need to install or develop functionality to generate Crossref metadata and/or interact with the Crossref API.

The easiest architecture for automatically sending the registered DOI names back to the blog is using the blog API. I have this implemented for the blogs where Rogue Scholar has admin access to the blog API (this blog and two other blogs using the Ghost platform), and this updates the blog post immediately after DOI registration as part of the same GitHub Action. Most Rogue Scholar blogs have a write API, and in the case of static site generators, the underlying repository platform (typically GitHub or GitLab) has an API. If your blog is updated more than once a month and is hosted by Wordpress, Ghost, Hugo, or Jekyll, reach out to me if you want to participate in the DOI registration beta workflow.

References

Fenner, M. (2023). *Streamlining the archiving of science blog posts*. https://doi.org/10.53731/gvb08-7kc16

Wimalaratne, S., & Fenner, M. (2018). *D2.1 Pid Resolution Services Best Practices*. https://doi.org/10.5281/ZENODO.1324300

Klein, M., Sompel, H. V. de, Sanderson, R., Shankar, H., Balakireva, L., Zhou, K., & Tobin, R. (2014). Scholarly Context Not Found: One in Five Articles Suffers from Reference Rot. *PLOS ONE*, *9*(12), e115253. https://doi.org/10.1371/journal.pone.0115253

Deane-Pratt, A. (2017). *THOR's last hurrah*. https://doi.org/10.59350/p000s-pth40

Fenner, M. (2016). *Cool DOIs*. https://doi.org/10.53731/r79x921-97aq74v-ag5a2