

Streamlining the archiving of science blog posts

Martin Fenner 

Published September 19, 2023

Citation

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

Keywords

News, Rogue Scholar



Copyright

Copyright © Martin Fenner 2023. Distributed under the terms of the [Creative Commons Attribution 4.0 International License](https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Front Matter

The [Rogue Scholar science blog](#) archive is adding important functionality to existing science blogs. The first step after a blog has signed up with Rogue Scholar is archiving the content. This is not only needed for long-term preservation but also enables full-text search and DOI registration with meaningful metadata. Rogue Scholar uses the blog feed (in RSS, Atom, or JSON Feed format) for this, which is updated the moment a blog post is published or updated.

The elegant approach would be to notify Rogue Scholar when this happens so that Rogue Scholar can fetch the updated content, using a technology called [webhooks](#). But one important principle of Rogue Scholar is simplicity, not requiring any additional technical work for the participating blogs unless absolutely necessary. This means that at least for the time being regular checks of the blog feed are more appropriate for Rogue Scholar, and with a small update yesterday this workflow has been greatly improved.

Rogue Scholar now checks all participating blogs for new or updated content every 10 minutes. When new or updated content is found, it is processed and stored in the Rogue Scholar Postgres database within a minute. This in turn triggers an update of full-text search index running in [Typesense](#) which takes another minute. Rogue Scholar checks every 10 minutes whether a blog post is new or the metadata used for DOI registration have changed and triggers a DOI update. DOI registration consists of two parts:

- Registration of the DOI and URL in the DOI resolution service, so that <https://doi.org/10.53731/xszpd-6z265> redirects to <https://blog.front-matter.io/posts/releasing-commonmeta-py-v0-8/>. This happens within a few minutes.
- Registration of DOI metadata, so that blog post metadata can be found via Crossref services. This happens within a few hours.

While writing this blog post, I got two emails from Crossref telling me about new content registered with Crossref. There were two blog posts by Rogue Scholar blogs this morning published 4 and 11 minutes before the DOIs were successfully registered, compared to the delay of several hours (and in a few cases even longer) before this update.

With this new workflow in place, another bottleneck now becomes more visible. Rogue Scholar and Crossref (and all the services that use Crossref metadata) now know about the DOIs registered for blog posts, but how do the participating blogs learn about this? For the special case where the blog has an API and Rogue Scholar is allowed to write to it (currently this blog and the [Upstream](#) blog, which both use the Ghost blogging platform), this happens automatically as part of the DOI registration GitHub Action.

For all other blogs participating in Rogue Scholar that is still something I have to figure out, and the main challenge is again to come up with a workflow that doesn't require major technical work for the participating blogs. One strategy would be to come up with solutions for the individual blogging platforms, starting with Wordpress which is used by [42% of Rogue Scholar blogs](#). Obviously, a solution for this issue is more important for blogs that are updated frequently than blogs that are updated only a few times per month. Stay tuned.

References

Fenner, M. (2023). *Releasing commonmeta-py* v0.8. <https://doi.org/10.53731/xszpd-6z265>