Collecting metadata for science blog posts

Martin Fenner

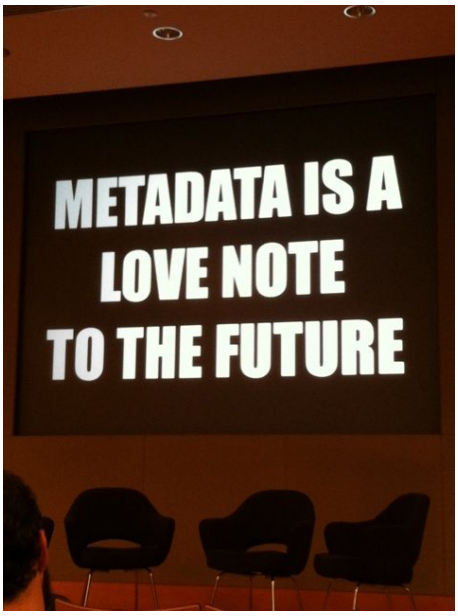
Published October 2, 2023

Citation

Fenner, M. (2023). Collecting metadata for science blog posts. *Front Matter*. https://doi.org/10.53731/yaws9-eyt23

Keywords

Feature, Rogue Scholar



Metadata are an important feature of every scholarly resource. For science blog posts – which are published with far fewer resources than for example journal articles or books – there is an additional requirement: make the metadata collection as painless as possible. In this post, I describe the lessons learned over the years, including recent work on the Rogue Scholar science blog archive.

Google Scholar and HTML meta tags

Scholarly resources published on the web want to expose their metadata to make it easier to find them. One important driver is Google Scholar and their Inclusion Guidelines for Webmasters are followed by most scholarly publishers and repositories. The guidelines rely heavily on HTML meta tags, in particular the Highwire Press and Dublin Core subsets. One problem with HTML meta tags is that they can't easily describe structured content, such as multiple authors, each with a name and ORCID ID. The bigger problem for science blogs is that generating these tags is a lot of work, and not really supported by standard blogging platforms.

Schema.org and Google Dataset Search

Schema.org describes structured content on the internet. It overcomes the limitations of HTML meta tags and can easily describe structured content such as multiple authors, each having a given name, family name, identifier, and affiliation. And schema.org is heavily used by Google and other search engines. But while schema.org is essential for Google Dataset Search and discovery of datasets published on the internet, it has seen little adoption to describe textual scholarly publications such as journal articles, conference proceedings, or books. One reason is that schema.org uses a very different approach from Google Scholar, another reason is that it is more difficult to implement.

I have used schema.org for many years to register DOIs with metadata for the DataCite blog. This approach worked well but was probably too complex to be adopted by a larger number of science blogs. One improvement was the combination of schema.org with HTML meta tags, but still required a lot of customization of each blog.

RSS and Atom Feeds

Blogs provide a unique mechanism to distribute metadata and content: RSS feeds, and the related Atom and JSON Feed formats. These formats provide all the required and some of the recommended metadata needed for scholarly content, including title, authors, publication date, abstract, location (URL), and language. RSS and Atom have been around for more than 15 years (JSON Feed was announced in 2017) and all blogging platforms support at least one of these formats.

RSS and related formats basically solved the challenge of metadata collection for science blogs, and therefore the Rogue Scholar science blog archive relies heavily on them. The experience

collecting metadata and content from more than 50 different science blogs (using 11 different blogging platforms) over the past several months has been very positive.

One major challenge with RSS and related formats is that they are not very good at providing archival content as they focus on the most recent blog posts. Several popular blogging platforms (e.g. WordPress and Blogger) provide pagination to access older content, but other platforms (e.g. the static site generators Hugo and Jekyll) provide no built-in pagination of RSS feeds.

Blogging platform APIs

RSS feeds are "poor man's APIs", e.g. they use the older XML serialization (RSS and Atom) instead of JSON serialization that dominates APIs today, have trouble accessing older content, and are read-only.

The next step in the evolution of metadata and content collection for science blogs is therefore to use existing JSON APIs, and in the last few weeks, the Rogue Scholar backend has been refactored to use these APIs if available. Ghost, WordPress (both self-hosted and WordPress.com), and Substack all provide nice JSON APIs so that the majority of Rogue Scholar blogs and blog posts are now retrieved via REST API calls. The early experience using these JSON APIs has been very encouraging and follows the fundamental principle of Rogue Scholar to not put a burden (technical, financial, or otherwise) on participating science blogs.

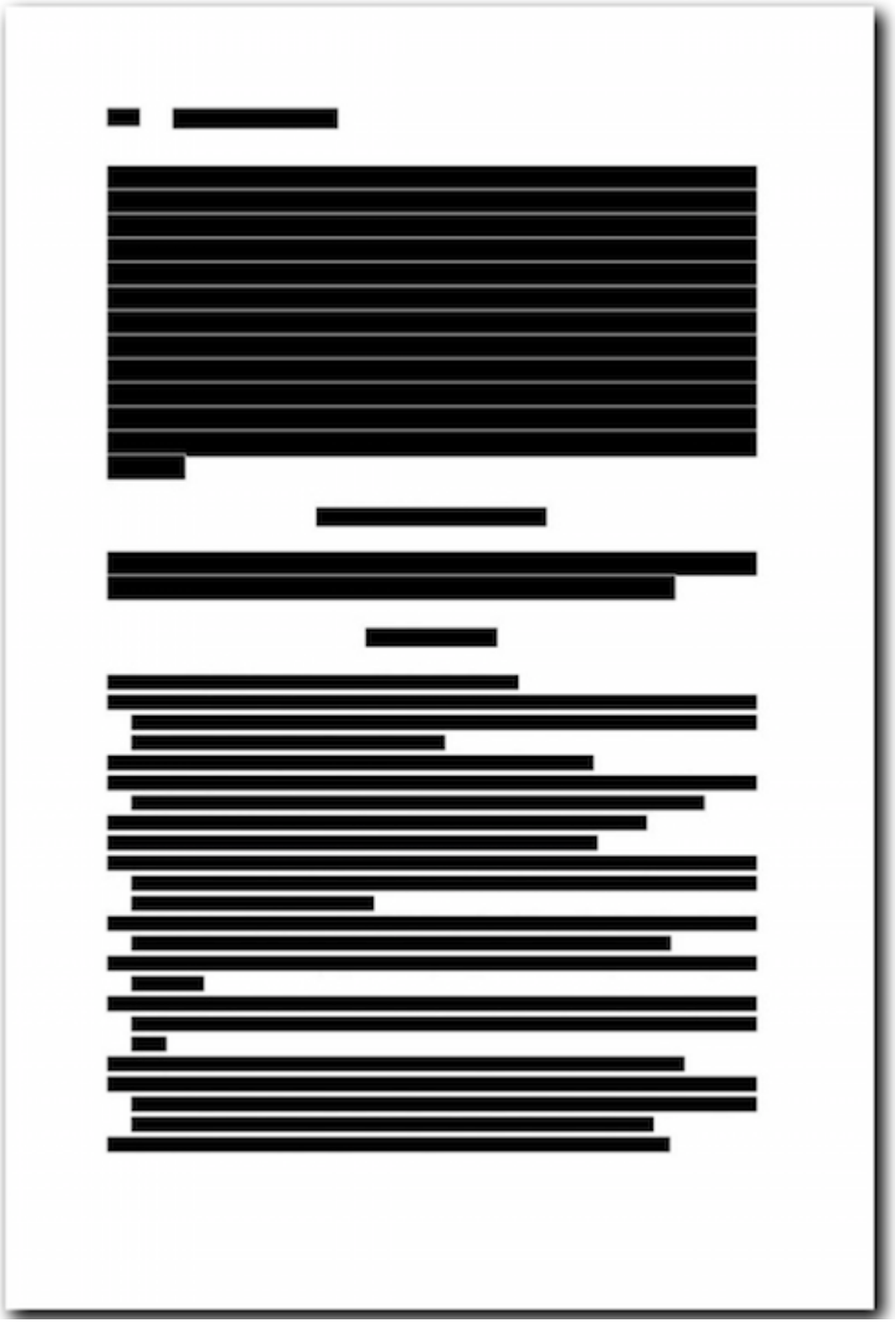
The use of blog APIs addresses another important problem: how are the DOIs registered by the Rogue Scholar service automatically added to the blogging platform? This issue is solved for several participating blogs using the Ghost platform and I am currently working on implementing this for blogs using WordPress, the most popular blogging platform on Rogue Scholar.

Extracting metadata from full-text content

One important limitation of using RSS feeds or blogging platform APIs is that they will only provide standard metadata, which sometimes might not be enough for scholarly content. I am currently exploring with one science blog extending the JSON Feed format with extensions, but that requires technical work by participating blogs and will probably scale poorly

A different approach that Rogue Scholar has followed for a few months takes advantage of the fact that all Rogue Scholar blog posts are available as full-text content with an open license (CC-BY) that allows reuse. A good example of important metadata that are not part of RSS or REST API metadata but included in the full-text content is references. As they typically follow a well-known pattern (a *References* or *Bibliography* section followed by a list of references formatted with various citation styles and including a link), it is not too difficult to extract them and include them in the metadata registered with a DOI.

Front Matter



Rogue Scholar implemented this approach in June. About 10% of Rogue Scholar blog posts now have their references registered with Crossref.

Other metadata that can be extracted from the full-text content describe funding information and relationships, for example, these common use cases for science blogs:

- **IsIdenticalTo**: The same content is cross-posted elsewhere, either on another blog or as a PDF in a repository,
- **IsTranslationOf**: The same content has been posted translated into another language,
- **IsPreprintOf**: The content has been published as a peer-reviewed paper in a journal,
- **HasAward**: The content has been funded as part of work on a research grant.

Front Matter

I recently started piloting this approach with two other blogs, including relationship links in an *Acknowledgments* section, and converting them to Crossref DOI metadata. About 1% of Rogue Scholar blog posts now include funding information and/or relationships.

Acknowledgments

This blog post was originally published on the DataCite Blog. This work was funded by the European Union's Horizon 2020 research and innovation programme under grant agreement No. 654039.

References

Fenner M, Demeranville T, Kotarski R, et al. *D2.1: Artefact, Contributor, And Organisation Relationship Data Schema*. Zenodo; 2015. doi:10.5281/ZENODO.30799

Acknowledgments and References sections used to extract metadata.

Conclusions

Collecting metadata for science blogs can be challenging, and this blog post summarizes some of the important lessons learned. However, the experience is also encouraging, as collecting metadata does not have to be a frustrating and time-consuming experience. We have a number of complementary approaches at our disposal. Rogue Scholar is currently mostly processing RSS feeds and blog platform APIs, but we can also complement this approach with metadata from HTML meta tags and/or schema.org.

I am sure the last chapter of this story hasn't been written yet. At least two challenges remain: collecting metadata from blogs that are no longer active and only persist in archived form, and updating blog posts with registered DOIs that are written using static site generators hosted on platforms such as GitHub and GitLab. Twenty-five percent of Rogue Scholar blogs fall into the latter category and the solution probably involves some form of GitHub Action (or Gitlab CI/CD) that triggers a pull request.

References

Fenner, M., Crosas, M., Grethe, J. S., Kennedy, D., Hermjakob, H., Rocca-Serra, P., Durand, G., Berjon, R., Karcher, S., Martone, M., & Clark, T. (2019). A data citation roadmap for scholarly data repositories. *Scientific Data*, *6*(1), Article 1. https://doi.org/10.1038/s41597-019-0031-8

Fenner, M. (2023). *Starting to include references in DOI metadata for blog posts*. https://doi.org/10.53731/6mkrk-dzh02