Mapping Metadata Inequities: Regional Disparities in Crossref Scholarly Records

Luis Montilla[1]
[1] *Crossref*

Purpose

The Crossref REST API provides the global community with access to over 165 million scholarly metadata records, serving as a foundational resource for discovery, citation, integrity assessment and provenance tracking (Hendricks et al. 2020). However, disparities in the completeness of this metadata can pose significant challenges to equitable access and utilization of scholarly knowledge. The diversity of barriers to metadata enrichment is complex and multifaceted, including but not limited to lack of institutional infrastructure, automation challenges, lack of awareness and/or training, language barriers or low availability of staff and resources to perform these tasks. Crossref schemas allow members to deposit rich metadata beyond basic bibliographic elements, namely abstracts, list of references, author identification via ORCIDs, institutional affiliation and also identification via ROR IDs, funding information, including funder and award IDs, journal update policies via the Crossmark service, license information, all of which contribute to the realize an open interconnected scholarly knowledge network that we aspire to as part of the Research Nexus. The regional differences in terms of journals included in the Crossref overall data have been previously described (Asubiaro & Onaolapo, 2023). Here, instead, we will explore regional disparities patterns in metadata completeness, focusing on highly relevant metadata fields such as references, author affiliations, ORCIDs, and funding information.

Methods

Crossref REST API provides unrestricted access to this metadata corpus, making it a de facto go-to tool for exploring potential variations in metadata completeness. The members API endpoint offers a set of percentage scores about the completeness of eleven recommended metadata fields faceted by content deposited in the last two years and historical content. Because the REST API is also a machine interface by definition, it allows automated and scripted routines to harvest and distil large amounts of metadata to perform low-level analyses. We will create a reproducible notebook to query the API via cURL or with specific libraries such as R httr2 or Python requests. This will allow us to automate a sequence of requests based on member ID data, including delays, to ensure that sequential and repeated calls never exceed a specified rate, which in turn can be passed to the appropriate formatting and cleaning steps to perform the exploratory data analysis.

Value and expected output

Identifying regions with less coverage will allow us to organize more targeted local events and health checks. Having this updated baseline is fundamental to Crossref's recent efforts to

provide support to the community to understand the importance of enriching their metadata records and reinforces the role of their members as metadata stewards. To address some of the aforementioned challenges, we started a regular series of metadata health check webinars, in English, Spanish, and Indonesian; additionally, the frequently used interface Participation Reports, has been expanded to include additional metadata that the community and consult at any moment. Similar efforts have also explored metadata completeness patterns across search engines and scholarly databases (Delgado-Quirós & Ortega, 2024; Céspedes et al. 2025), which often rely on Crossref metadata to support their services; other studies have relied on Crossref's reports to make similar reviews (Ermakov, 2021), however this would represent an updated in-house assessment that can inform of the success of current and future efforts to continue supporting the global community and improve the quality of the scholarly record.

References

Asubiaro, T. V., & Onaolapo, S. (2023). A comparative study of the coverage of African journals in Web of Science, Scopus, and CrossRef. In Journal of the Association for Information Science and Technology (Vol. 74, Issue 7, pp. 745–758). Wiley. https://doi.org/10.1002/asi.24758

Céspedes, L., Kozlowski, D., Pradier, C., Sainte-Marie, M. H., Shokida, N. S., Benz, P., Poitras, C., Ninkov, A. B., Ebrahimy, S., Ayeni, P., Filali, S., Li, B., & Larivière, V. (2025). Evaluating the linguistic coverage of <scp>OpenAlex</scp>: An assessment of metadata accuracy and completeness. In Journal of the Association for Information Science and Technology. Wiley. https://doi.org/10.1002/asi.24979

Delgado-Quirós, L., & Ortega, J. L. (2024). Completeness degree of publication metadata in eight free-access scholarly databases. In Quantitative Science Studies (Vol. 5, Issue 1, pp. 31–49). MIT Press. https://doi.org/10.1162/qss_a_00286

Ermakov, A. V. (2021). Analysis of Crossref Reports in Order to Improve the Quality of Metadata of Scientific Publications. In Scientific Conference "Scientific Services & Internet." 23rd Scientific Conference "Scientific Services & Internet – 2021." Keldysh Institute of Applied Mathematics. https://doi.org/10.20948/abrau-2021-4-ceur

Hendricks, G., Tkaczyk, D., Lin, J., & Feeney, P. (2020). Crossref: The sustainable source of community-owned scholarly metadata. In Quantitative Science Studies (Vol. 1, Issue 1, pp. 414–427). MIT Press - Journals. https://doi.org/10.1162/qss_a_00022