

Optimizing metadata enhancement workflow of current research information system via CrossRef – pilot study for a real-life project

Kin Lok Rocky Mak* and Richard Parish

Anglia Ruskin University, Chelmsford, United Kingdom

Abstract

This pilot study aims to explore the possibility of reducing human intervention in assuring metadata quality of a current research information system via CrossRef metadata ingestion. 5 key Dublin Core elements have been selected for review and a significant discrepancy in metadata accuracy between fields is observed. Further study is needed to confirm possibility of automation.

Keywords

Current Research Information System, Metadata Quality, Automation

1. Introduction

To comply with relevant funder and reporting requirements on open research, current research information systems (CRIS) are frequently procured and operated by UK universities to capture and store metadata of research-related activities. Maintaining a CRIS can be a labour-intensive process as much work has to be done to harvest, clean and validate data ingested into the CRIS¹. While a range of data sources are available for CRIS administrators to harvest from, it is only discovery that has been solved. Metadata accuracy from these data sources are often overlooked both as a practical and research issue for repository managers and researchers alike, evident from the lack of extant literature on the topic. As a pilot study for a real-life project aimed at optimizing workflow on metadata enhancement in the CRIS and institutional repository (IR) of a medium-sized UK university, this project aims to evaluate the data accuracy of CrossRef metadata ingestion to Symplectic Elements, a CRIS in use by the author's institution. In turn, this pilot study hopes to examine the possibility in reducing human intervention for parts of the metadata quality assurance process.

2. Methodology

A list of research outputs affiliated with the authors' institution during the period 1 October 2023 to 1 Mar 2024 were retrieved from Scopus. 100 records were then chosen with the use of a randomizer for metadata review. 5 Dublin Core elements (title, creator, description, publisher and date) are chosen given their importance for findability, interoperability and accessibility for

* Corresponding author.

✉ rocky.mak@aru.ac.uk (K.L.R. Mak); richard.parish@aru.ac.uk (R. Parish)

ORCID 0000-0003-2400-6063 (K.L.R. Mak); 0000-0002-5385-5761 (R. Parish)



© 2024 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

records in the CRIS. The first author, who is an experienced repository manager, compared between the original metadata available on the hosting site of publication and the metadata being ingested to Symplectic Elements via CrossRef. Exact match is desired due to the need of metadata consistency as CRIS acts as the source of truth for internal and external reporting and research evaluation processes².

3. Findings

In the 100 records chosen, 76 records are available on CRIS via CrossRef for metadata review. A number of key issues have been identified from this review process.

Accuracy of title and creator field ingestion is high, with only 4 and 3 discrepant records respectively. All 4 discrepancies in the title field relates to book/book chapters where title/sub-titles are not clearly defined and not supported by Dublin Core vocabularies. 2 out of 3 discrepancies in the creator field related to group collaborators as authors, where its popularity has grown due to advent of Big Science but not well supported by Dublin Core.

The description field, however, is empty for 34 records and a further 42 records have issues impeding human accessibility but are accurate in content. The latter are affected by the lack of support of JATS schema, the native metadata format for many publishers. As such the labels are not removed when ingested to the CRIS³. While machine readability is not affected, it has effects downstream when records in the CRIS are exported to IR for showcasing purposes.

A difference in practice between CRIS administrators and publishers in metadata creation is revealed in the publisher field. 18 are marked with the company name instead of imprints. This may cause issues as different imprints with the same publishing house might have different Open Access policies. This might in turn lead to erroneous open access reporting in the system, which is crucial for UK universities in the Research Excellence Framework (REF) processes.

Significant inconsistencies can be found in the date field, where 32 records are deemed problematic in the review process. The most common issue found is the absence of online publication date, which is important for REF and funder open access compliance. Some others lack a precise date of publication in CrossRef ingestion but is available on publishers' site. While some publishers provide rich metadata on timeline of the whole scholarly communication process (acceptance, online available and issue publication), this is not a sector-wide practice.

The 24 records unavailable on CrossRef are due to 2 reasons. 20 records are not harvested despite correct affiliation information and is recognized by Scopus. While this may be a CRIS harvesting issue, it also reflects potential advantages of incorporating a multitude of third-party platforms in CRIS metadata harvesting due to differences in breadth and depth of metadata richness as illustrated in previous studies⁴. 4 records lack a DOI and are therefore not available on CrossRef. Harvesting these remains a constant issue for repository managers, leading reliance to third-party data sources that are not reliant on DOI for harvesting.

4. Conclusion and future development

There is still much to explore in fully automating metadata ingestion to a CRIS system via different data sources without human intervention. The pilot has indicated possibility of foregoing checking of 2 key fields given the high degree of accuracy. Human enhancement of other fields remains necessary due to the inconsistencies between publishers in metadata practices, incompatibility of publisher and CrossRef schema and the specific demands of UK universities in research evaluation and reporting. The next step from this pilot project is to examine metadata quality of third-party data sources such as Scopus, Web of Science and Dimensions so to test possibility of reducing human intervention in the quality assurance process.

¹ Azeroual, Otmane, and Joachim Schöpfel. "Quality issues of CRIS data: An exploratory investigation with universities from twelve countries." *Publications* 7.1 (2019): 14.

² Kumar, Vinit, Chandrappa, and N. S. Harinarayana. "Exploring dimensions of metadata quality assessment: A scoping review." *Journal of Librarianship and Information Science* (2024): 09610006241239080.

³ Hendricks, Ginny, et al. "Crossref: The sustainable source of community-owned scholarly metadata." *Quantitative Science Studies* 1.1 (2020): 414-427.

⁴ Harzing, Anne-Wil. "Two new kids on the block: How do Crossref and Dimensions compare with Google Scholar, Microsoft Academic, Scopus and the Web of Science?." *Scientometrics* 120.1 (2019): 341-349.