Applying LLM to Library Metadata: Mapping Geography and Language in the Library of Congress Collection

Hongyu Zhou[1,2,†], Kai Li[3,**,†], Raf Guns[1], Brian Dobreski[3] and Tim C. E. Engels[1]

[1]University of Antwerp, Antwerp, Belgium
[2]University of Cambridge, Cambridge, United Kingdom
[3]University of Tennessee, Knoxville, United States

Extended Abstract

Books are among the most enduring forms of cultural and scholarly communication, yet they remain largely invisible in quantitative analyses of knowledge production. Unlike journal articles, which are easily captured in citation databases, books are embedded in library catalog systems whose rich metadata have rarely been exploited for large-scale research. In this paper, we demonstrate how large language models (LLMs) can enhance the research potential of such metadata and, in doing so, provide new empirical insights into the geography and language of global knowledge.

We assemble and process more than 6.4 million non-fiction bibliographic records from the U.S. Library of Congress (LC) catalog spanning 1970 to 2018. From the MARC-format metadata, we extract three analytical dimensions: (i) subject geography derived from Library of Congress Subject Headings (LCSH), (ii) publication locations, and (iii) languages of publication. A custom LLM-based normalization pipeline processes over 250,000 unique geographic strings, ranging from cities to historical regions, and maps them to contemporary countries and territories using ISO 3166 identifiers. Validation against the official MARC 043 geographic area codes achieves over 96% concordance, demonstrating the accuracy and scalability of LLM-assisted geographic classification for bibliometric research [1].

This integration of LLM-driven text normalization with curated library metadata highlights subtle but systematic differences between machine-inferred and human-assigned geographic classifications. The LLM-based approach can be instructed to align more closely with contemporary geopolitical boundaries and naming conventions, whereas librarian-curated metadata may exhibit inconsistencies or temporal lag in reflecting political or territorial shifts[2]. More broadly, this demonstrates how LLMs enable a new, adaptive form of knowledge organization that complements traditional cataloging by dynamically updating representations of place, language, and culture. Rather than replacing human curation, such systems can extend the bibliographic infrastructure of libraries into a continuously evolving framework for mapping global knowledge.

Using this LLM-enhanced dataset, we trace how the LC's global representation has evolved over the past five decades. Three empirical patterns emerge. First, the LC collection has undergone substantial geographic diversification: the share of books about North America declined from more than 30 percent in the 1970s to under 20 percent by the 2010s, while East Asia, Latin America, and Eastern Europe expanded rapidly, mirroring the globalization of publishing [3]. Second, geographic and linguistic dimensions are tightly aligned, with over 80 percent of books about a country published in its official language and 81 percent sharing the same publication and subject country, indicating that catalog metadata accurately reflect the spatial organization of book knowledge. Third, the linguistic composition of the collection has shifted markedly: English-language titles fell from over 50 percent to around one-third, while Chinese, Spanish, and other non-English languages rose steadily, signaling the emergence of a more multilingual and globally distributed archive.

Taken together, these results portray the Library of Congress not merely as a passive repository but as an active infrastructure of global knowledge representation. Its evolving catalog mirrors the United States' shifting intellectual engagement with the world and the diffusion of publishing capacity beyond the Western core. More broadly, our findings demonstrate how LLMs can transform long-standing bibliographic systems, originally designed for human catalogers, into computational data sources for mapping global information flows.

By linking geography, language, and publication metadata at scale, and by reconciling librarian-curated and LLM-derived classifications, this study contributes a methodological advance in the AI-assisted science of science. It shows how the combination of controlled vocabularies and generative models can illuminate hidden cultural and geopolitical dynamics in the global production of knowledge, offering new pathways for research on linguistic diversity, cultural equity, and the spatial organization of scholarship.

Keywords

Library Metadata, Large Language Models, Knowledge Geography, Cataloging, Cultural Representation

Declaration on Generative AI

The authors used large language models (LLMs) to assist with grammar and style editing, as well as text normalization within the research methodology. LLMs were employed to standardize geographic entities and harmonize metadata extracted from the Library of Congress catalog. No figures were generated using generative AI. All AI-assisted outputs were reviewed and verified by the authors, who take full responsibility for the content of this publication.

References

[1] F. A. Black, B. H. MacDonald, J. M. Black, Geographic information systems: A new research method for book history, Book History 1 (1998) 11–31.

[2] D. N. Joudrey, A. G. Taylor, D. P. Miller, Introduction to cataloging and classification, 11 ed., Bloomsbury Publishing USA, 2015.

[3] L. Leydesdorff, O. Persson, Mapping the geography of science: Distribution patterns and networks of relations among cities and institutes, Journal of the American Society for Information Science and Technology 61 (2010) 1622–1634.

AI4SciSci'25: Workshop on the Artificial Intelligence and the Science of Science, December 15, 2025, DeKalb, Illinois, USA

*Corresponding author: Kai Li (kli16@utk.edu)

†These authors contributed equally to this work.

kli16@utk.edu (K. Li)

0000-0002-8250-5875 (H. Zhou); 0000-0002-7264-365X (K. Li); 0000-0003-3129-0330 (R. Guns); 0000-0002-2448-3495 (B. Dobreski); 0000-0002-4869-7949 (T. C. E. Engels)