https://doi.org/10.51408/issi2025_107

Exploring Google Books, Open Library, and Wikipedia as Sources for Book Metadata: The UK and Lithuanian Cases

Eleonora Dagienė

eleonora.dagiene@mruni.eu
Institute of Communication, Mykolas Romeris University,
Ateities g. 20, LT-08303 Vilnius (Lithuania)

Abstract

This paper investigates the potential of Google Books, Open Library, and Wikipedia as sources of metadata for scholarly books, focusing on publications from the UK and Lithuania. Utilising ISBNs as unique identifiers, the study analyses the availability, accuracy, and completeness of metadata across these platforms. Initial findings reveal significant disparities between UK and Lithuanian book metadata, with UK publications exhibiting higher coverage and consistency. The research highlights the limitations of these sources, particularly for non-English language publications, and underscores the need for further investigation to develop a more comprehensive and reliable book metadata ecosystem. This research contributes to the ongoing discussion about improving book metrics and enhancing the evaluation of scholarly outputs.

Introduction

Bibliometric research relies heavily on comprehensive and reliable data sources. However, existing research often faces limitations in capturing the full spectrum of scholarly publications, particularly scholarly books, which remain the most challenging and therefore least researched outputs (Borgman & Furner, 2005). Previous studies have explored book citation metrics using various sources, including traditional journal-oriented citation indexes (Halevi et al., 2016; Zuccala & Robinson-García, 2019), online platforms such as Google Books, Google Scholar, and Wikipedia (Kousha et al., 2011; Kousha & Thelwall, 2017), and the WorldCat library catalogue (Torres-Salinas et al., 2021). These studies, however, often focus on books already included in established databases, which may exhibit geographic or linguistic biases.

This limitation becomes evident when examining national research outputs. For instance, a significant proportion of books from countries such as Lithuania (Dagienė, 2024a) and Croatia (Sīle et al., 2021) are entirely missing metadata in the internationally recognised WorldCat catalogue, rendering them practically invisible to international readers. This lack of comprehensive metadata hinders the development of intelligent research metrics (Moed, 2007), and ultimately limits our ability to accurately evaluate and recognise the contributions of scholarly books.

This research-in-progress explores the availability and quality of book metadata in Google Books, Open Library, and Wikipedia for scholarly books submitted as research outputs in the UK and Lithuania, using ISBNs as the primary book identifier. The study aims to answer the following research questions:

1. What is the availability of book metadata in these sources for these particular books (from the UK and Lithuania)?

2. How consistent and accurate is the metadata across these chosen sources?
3. What are the challenges and opportunities in using these data sources for developing book metrics?

By addressing these research questions, this study aims to contribute to a better understanding of the challenges and opportunities in leveraging book metadata for research evaluation and knowledge discovery. Specifically, it explores the possibilities and sources for creating intelligent research metrics. The primary goal is to identify potential data sources and approaches for developing comprehensive intelligent book metrics that can reveal the merit of every book, contributing to a more nuanced, fair, and effective research evaluation system (European Commission, 2021; UNESCO, 2021). Combined with transparent peer review, such intelligent research metrics hold the potential to transform research evaluation practices and address the needs of the future of scholarly communication (Kraker et al., 2016).

Methodology

This research employs a mixed-methods approach, combining quantitative and qualitative analysis of book metadata from three globally recognised sources: *Google Books*[1], *the Open Library*[2], and *Wikipedia*[3]. The empirical analysis uses datasets of books submitted as research outputs for research evaluation between 2008 and 2020, comprising 38,050 ISBNs in the UK (Dagienė, 2023c) and 5,199 ISBNs in the Lithuanian datasets (Dagienė, 2023b). These datasets provided publication years (as provided by submitting institutions). Additionally, they provided book type (authored book or edited volume from the national submission systems), country of ISBN issuance, publisher name, and primary publisher occupation (obtained from the Global Register of Publishers) (Dagienė, 2024b).

Primary metadata (authors, titles, publishers, and publication years) for each ISBN was collected from the three sources using the Python package *isbnlib*[4], which provides functions for retrieving metadata via application processing interfaces (APIs). To track changes in metadata availability, data for all ISBNs from Google Books, Open Library, and Wikipedia were collected in both January 2023 and January 2025.

The analysis focused on determining the number of books from the UK and Lithuanian datasets present in each source and assessing the completeness of their metadata. Books were categorised based on the level of agreement between the three sources. "Matched" shows an exact match in at least two sources, ideally three, suggesting accurate data. "Partial match" signifies potential data availability where at least one author or key title words matched across at least two sources. "One source only" shows data available in only one of the three sources. "No exact match"

[1] Google for Developers. https://developers.google.com/books/docs/v1/libraries accessed 2 January 2025

[2] Open Library. Developer Center / APIs / Books API https://openlibrary.org/dev/docs/api/books accessed 2 January 2025

[3] Wikimedia REST API https://en.wikipedia.org/api/rest_v1/#/Citation accessed 2 January 2025

[4] isbnlib – a python library to validate, clean, transform and get metadata of ISBN strings (for devs) https://github.com/xlcnd/isbnlib; accessed 2 January 2025

highlights discrepancies in metadata elements requiring further review. "No data" means no metadata is available in any of the sources explored.

The figures presented in the following sections illustrate the availability of metadata (author, title, publisher, year, language) for the UK and Lithuanian book ISBNs from 2008 to 2020. The findings suggest that if books have metadata that fall into the first three categories, the sources might be suitable for compiling ISBN metadata for various purposes.

Availability of book metadata

This section presents initial findings on the availability of book metadata in Google Books, Open Library, and Wikipedia. The analysis examines datasets of the UK and Lithuanian book ISBNs, comparing the metadata gathered in January 2023 and January 2025.

Overall, metadata availability for UK books is high across all sources. Google Books' coverage remained consistently high, with 84.3% (32,076 ISBNs) in 2023 and 81.4% (30,972 ISBNs) in 2025. Open Library's coverage increased slightly from 92.1% (35,028 ISBNs) in 2023 to 93.6% (35,601 ISBNs) in 2025. Wikipedia, while initially the highest in 2023 at 99.1% (37,695 ISBNs), saw a decrease to 93.7% (35,655 ISBNs) in 2025, warranting further investigation.

In contrast, Lithuanian book metadata is less readily available. Google Books' coverage increased from 27.6% (1,436 ISBNs) in 2023 to 42.0% (2,181 ISBNs) in 2025, but a significant proportion (58%) still lacks records. Open Library showed minimal change, with coverage around 31-32%. Wikipedia's coverage also decreased, from 70.9% (3,688 ISBNs) in 2023 to 55.4% (2,881 ISBNs) in 2025, possibly because of fewer Lithuanian books being cited/mentioned on Wikipedia.

Combining the 2025 data from all three sources, only 1.5% of UK ISBNs lack metadata. Wikipedia and Open Library are the best choices for UK ISBN metadata, with Google Books also providing sufficient coverage. The picture is less promising for Lithuanian ISBNs, with over a third lacking records even after combining data from all three sources. Even when records exist, they are often incomplete. The next section will explore the consistency and accuracy of the metadata. Further research is needed to understand the decrease in Wikipedia coverage for both UK and Lithuanian books and to identify additional data sources to improve metadata availability for Lithuanian books.

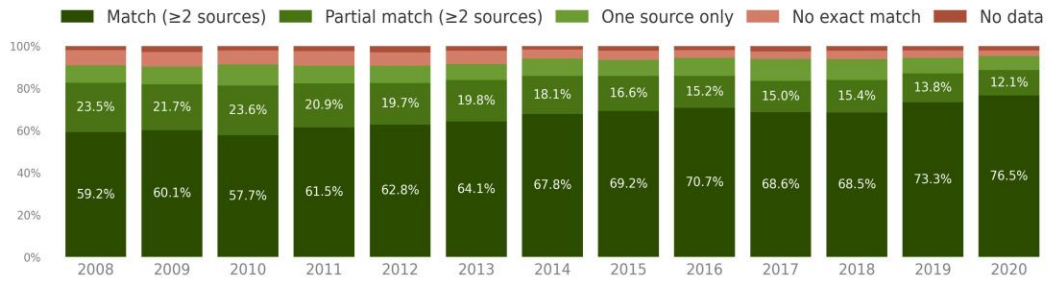
Accuracy and completeness of the metadata

This section examines the accuracy and completeness of key metadata fields (author names, titles, publishers, years, and languages) across Google Books, Open Library, and Wikipedia. Because of space constraints, this paper presents combined results from the three sources. A detailed analysis of each source will be provided in the full paper.

Figure 1 shows how author information is represented, combining data from all three sources. UK ISBNs consistently show higher author information availability than Lithuanian ISBNs, ranging from 90% to 95% for UK books compared to 49% to 57% for Lithuanian books between 2008 and 2020. Furthermore, almost 90% of UK

ISBNs have author information that is consistent in at least two sources in recent years, indicating higher data quality. In contrast, only slightly over 20% of Lithuanian ISBNs have matching author information, with almost half lacking author data across all three sources. This highlights a significant gap in author information for Lithuanian books.

a)



b)

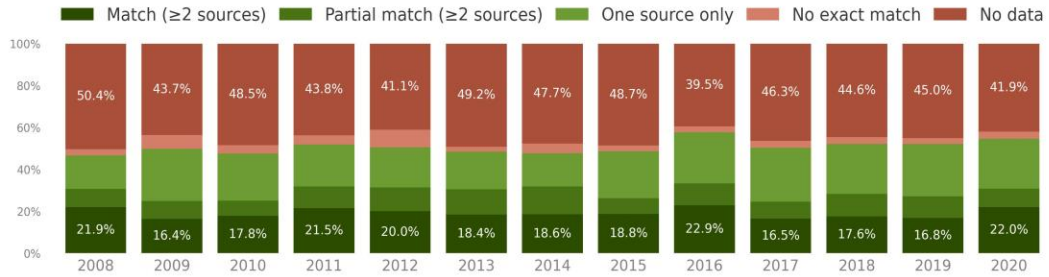
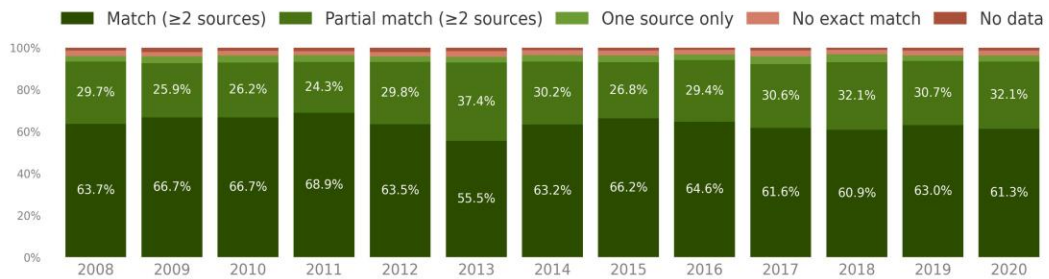


Figure 1. Author information availability for: a) UK and b) Lithuanian ISBNs.

Turning to ***title information***, the sources provide more complete data for titles than for authors (combining both matched and partially matched records). As with author information, UK ISBNs show higher data quality with at least 90% of UK titles have matches in two or three sources. While Lithuanian ISBNs have less missing title data than author data, they still face challenges. Only around 20% of titles are consistently represented in recent years, and over 30% of titles have no data across any of the sources. For Lithuanian ISBNs, information available in only one source is similar to the amount of matched or partially matched records across multiple sources. This suggests potential inconsistencies in title information for Lithuanian books that are already available not speaking about missing records.

a)



b)

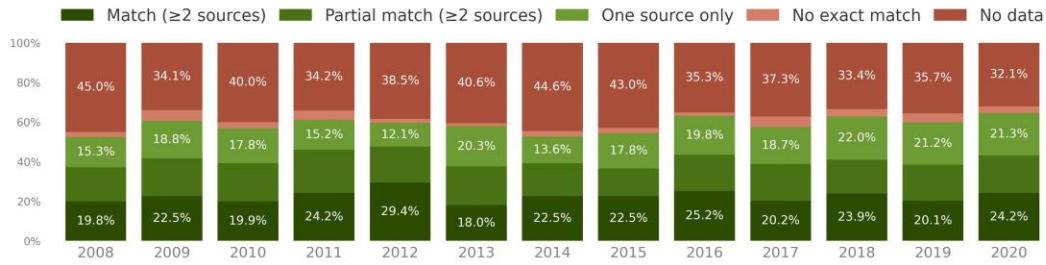
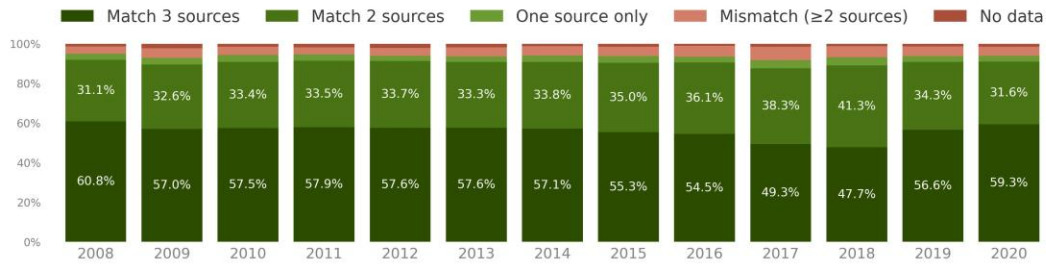


Figure 2. Title information availability for: a) UK and b) Lithuanian ISBNs.

Figure 3 illustrates the availability of year information. Despite more reliable year information from UK ISBNs, we found some inconsistencies in the gathered data; sometimes, the years extracted from the sources were clearly inaccurate (e.g., 2028, 1958, or 1810).

a)



b)

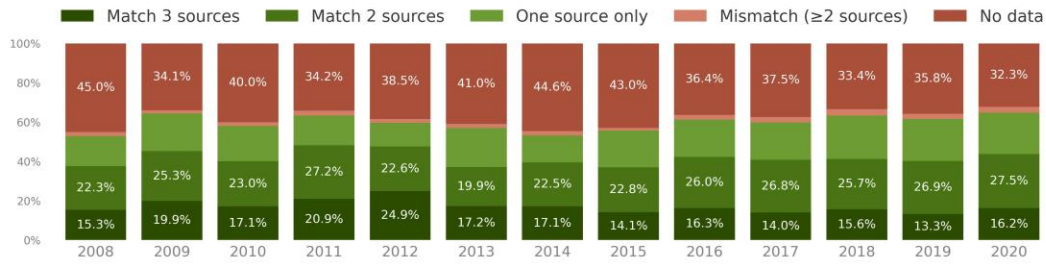


Figure 3. Year information availability for: a) UK and b) Lithuanian ISBNs.

Additionally, some years mismatched between those gathered from three sources and those reported by the UK and Lithuanian institutions at the submission stage. These mismatches were more frequent for UK ISBNs than Lithuanian ISBNs. Despite these issues, over half of the year records for UK ISBNs match across all three sources, and when combined with those that match in two sources, almost 90% of UK records have consistent information on the years. In contrast, over a third of Lithuanian records are missing year information entirely, with very few records having years that match across all three sources.

Publishers. Only 2.1% of UK books lack publisher information in their records, compared to 42.0% of Lithuanian books. Figure 4 shows the overall percentages of

available and missing publisher data, with numbers very similar to those seen for author and title information. A closer look at 'Partial match' records revealed cases where publishers operate under multiple imprints. In such cases, one source may list the imprint's name, while others provide the parent publisher's name, leading to discrepancies. Similar inconsistencies arise after publishers merged, acquired, or reorganised their imprints. Considering that so many ISBNs have no metadata even with all three sources combined, if someone is looking for the reliable publisher data, they can get data from the Global Register of Publishers (Dagienė, 2024b). This registry metadata contains publishers' information for every ISBN issued in any of the over two hundred countries that have joined the ISBN system.

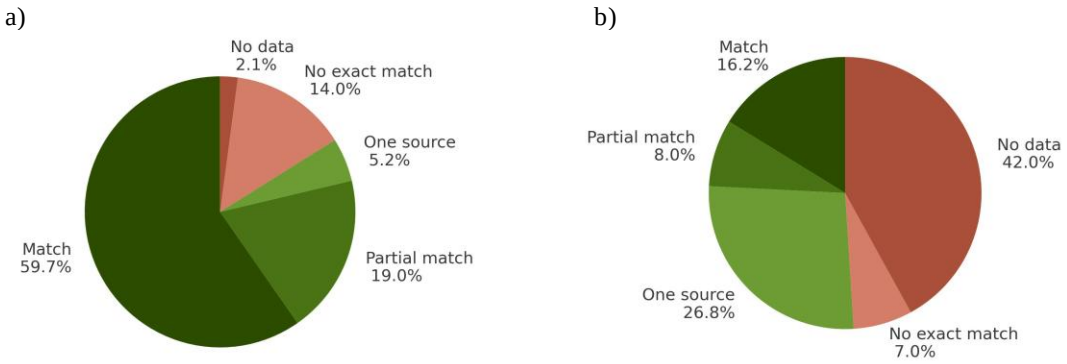


Figure 4. Publisher information availability for: a) UK and b) Lithuanian ISBNs.

Book ***language information*** is available only in Google Books, and its coverage is lower than that of authors, titles, or publishers for both the UK and Lithuanian datasets. The results show that UK books were issued in 29 languages, while Lithuanian books were issued in 17 languages. In the UK ISBN results, language information is missing for 18.6% of books. Of the entire UK dataset, 78.1% are in English, 1.8% are in German, 0.6% are in French, and other languages have even smaller numbers of ISBNs (Figure 5a).

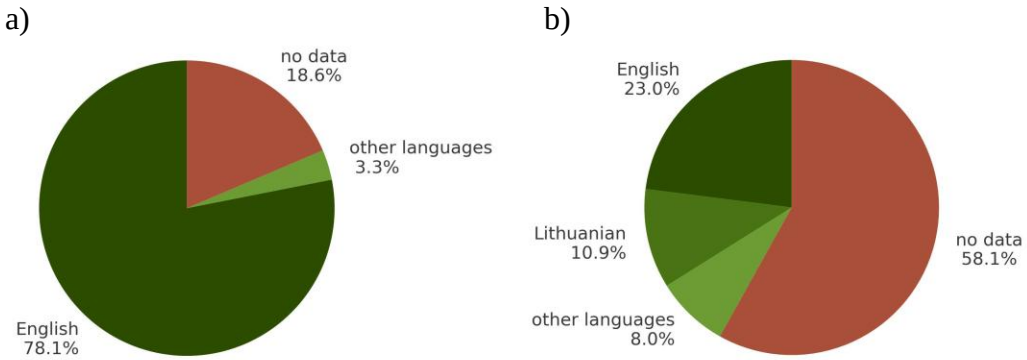


Figure 5. Language information availability in Google Books for: a) UK and b) Lithuanian ISBNs.

Interestingly, Figure 5b shows that more Lithuanian ISBNs are assigned to English-language books than to Lithuanian ones, which contradicts previous findings that if half of the Lithuanian books were issued in Lithuania, they would likely be in Lithuanian, not in other languages. Presumably, the significant of metadata of domestically published books is missing in the international sources analysed in this research study.

This analysis of the accuracy and completeness of metadata across Google Books, Open Library, and Wikipedia reveals significant variations in data quality for UK and Lithuanian books. The UK's ISBNs consistently show higher data quality and completeness across all metadata fields. In contrast, Lithuanian ISBNs exhibit lower data quality, with a significant proportion missing primary metadata elements as author names and titles. These findings highlight the challenges in relying solely on these sources for comprehensive book metadata, particularly for research evaluation purposes. The full paper will provide a more detailed analysis of each source and explore strategies to address these limitations.

Challenges, opportunities, and initial conclusions of using three data sources

This research investigates the potential of Google Books, Open Library, and Wikipedia as sources for book metadata, empirically testing their efficacy using UK and Lithuanian scholarly publications. While initial results show these platforms providing a significant amount of ISBN metadata, they also reveal notable limitations, particularly for publications from non-English-speaking countries as Lithuania. This disparity is clear in the higher coverage and accuracy of metadata for UK books compared to their Lithuanian counterparts. Moreover, inconsistencies in author names, publication years, and publisher information frequently occur across these sources, even for already represented books. These initial findings underscore the need for further investigation to identify the missing data and find out the reasons behind these gaps.

The full paper will, therefore, focus on three key areas. First, a deeper analysis will be conducted to investigate the influence of publisher type and size on metadata availability and quality, exploring potential avenues for improving the representation of books and their ISBNs through publisher engagement. Second, the research will identify the underlying sources that Google Books, Open Library, and Wikipedia leverage to compile their extensive book databases and potentials for covering underrepresented book metadata. Third, a thorough examination of the nature and the extent of discrepancies across these platforms will be undertaken. This will provide crucial insights into whether these inconsistencies can be rectified, paving the way for a unified and more reliable book metadata universe.

Ultimately, the overarching goal of this project is to identify and analyse a comprehensive range of book metadata within three platforms explored here, representing only the initial steps. By researching the ways for integrating data from the diverse sources, we aim to achieve maximum metadata coverage for books within any dataset, moving beyond the current limitations of a Western, English-language-centric focus and embracing a truly global perspective. This will significantly

enhance the reliability and comprehensiveness of book metadata, improving its value for research evaluation (Dagienė, 2023a) and the development of robust book metrics that assist book peer-review assessment.

References

Borgman, C. L., & Furner, J. (2005). Scholarly communication and bibliometrics. *Annual Review of Information Science and Technology*, *36*(1), 2–72. https://doi.org/10.1002/aris.1440360102

Dagienė, E. (2023a). Prestige of scholarly book publishers—An investigation into criteria, processes, and practices across countries. *Research Evaluation*, *32*(2), 356–370. https://doi.org/10.1093/reseval/rvac044

Dagienė, E. (2023b). *The metadata of books submitted as research outputs to annual Lithuanian research assessments from 2008 to 2020 [Data set]* [Dataset]. https://doi.org/10.5281/zenodo.10070933

Dagienė, E. (2023c). *The metadata of books submitted as research outputs to REF 2014 and REF 2021 [dataset]* [Csv]. Zenodo. https://doi.org/10.5281/zenodo.10071003

Dagienė, E. (2024a). Mapping scholarly books: Library metadata and research assessment. *Scientometrics*, *129*, 5689–5714. https://doi.org/10.1007/s11192-024-05120-1

Dagienė, E. (2024b). The challenge of assessing academic books: The UK and Lithuanian cases through the ISBN lens. *Quantitative Science Studies*, *5*(1), 98–127. https://doi.org/10.1162/qss_a_00284

European Commission. (2021). *Towards a reform of the research assessment system* (p. 21). European Union. https://doi.org/10.2777/707440

Halevi, G., Nicolas, B., & Bar-Ilan, J. (2016). The complexity of measuring the impact of books. *Publishing Research Quarterly*, *32*(3), 187–200. https://doi.org/10.1007/s12109-016-9464-5

Kousha, K., & Thelwall, M. (2017). Are Wikipedia citations important evidence of the impact of scholarly articles and books? *Journal of the Association for Information Science and Technology*, *68*(3), 762–779. https://doi.org/10.1002/asi.23694

Kousha, K., Thelwall, M., & Rezaie, S. (2011). Assessing the citation impact of books: The role of Google Books, Google Scholar, and Scopus. *Journal of the American Society for Information Science and Technology*, *62*(11), 2147–2164. https://doi.org/10.1002/asi.21608

Kraker, P., Dörler, D., Ferus, A., Gutounig, R., Heigl, F., Kaier, C., Rieck, K., Šimukovič, E., & Vignoli, M. (2016, December 30). *The Vienna Principles: A Vision for Scholarly Communication in the 21st Century*. https://doi.org/10.5281/zenodo.55597

Moed, H. F. (2007). The future of research evaluation rests with an intelligent combination of advanced metrics and transparent peer review. *Science and Public Policy*, *34*(8), 575–583. https://doi.org/10.3152/030234207X255179

Sīle, L., Guns, R., Zuccala, A., & Engels, T. (2021). Towards complexity-sensitive book metrics for scholarly monographs in national databases for research output. *Journal of Documentation*, *77*(5), 1173–1195. https://doi.org/10.1108/JD-06-2020-0107

Torres-Salinas, D., Arroyo-Machado, W., & Thelwall, M. (2021). Exploring WorldCat identities as an altmetric information source: A library catalog analysis experiment in the field of Scientometrics. *Scientometrics*, *126*(2), 1725–1743.

https://doi.org/10.1007/s11192-020-03814-w

UNESCO. (2021). *Recommendation on Open Science* (p. 34). UNESCO.

Zuccala, A., & Robinson-García, N. (2019). Reviewing, Indicating, and Counting Books for Modern Research Evaluation Systems. In *Springer Handbook of Science and Technology Indicators* (pp. 715–728). https://doi.org/10.1007/978-3-030-02511-3_27