



Universiteit
Leiden
The Netherlands

Quantitative research assessment and its unintended consequences

Dagiene, E.

Citation

Dagiene, E. (2025, October 30). *Quantitative research assessment and its unintended consequences*. Retrieved from <https://hdl.handle.net/1887/4281943>

Version: Publisher's Version

License: [Licence agreement concerning inclusion of doctoral thesis in the Institutional Repository of the University of Leiden](#)

Downloaded from: <https://hdl.handle.net/1887/4281943>

Note: To cite this publication please use the final published version (if applicable).

Chapter 6.

Mapping scholarly books: Library metadata and research assessment

This chapter is based on:

Dagienė, E. (2024a). Mapping Scholarly Books: Library Metadata and Research Assessment. *Scientometrics* 129, 5689–5714. <https://doi.org/10.1007/s11192-024-05120-1>

6.1. Introduction

Current research assessment policies in Lithuania prioritise book publication with prestigious publishers to maximise visibility and internationalisation of national research. However, this “prestige economy”, prevalent not just in Lithuania but worldwide, goes beyond inconsistencies in evaluation, neglecting the merits of individual books (Dagienė 2023a). New approaches, such as comprehensive book metrics, are needed to track individual books throughout their lifecycle.

This paper advocates for individual book assessment, aligning with the open science and responsible research evaluation principles outlined by UNESCO (UNESCO 2021), the European Commission (2019), and various declarations and guides (Collins et al. 2015; European Commission 2019a; Federation of Finnish Learned Societies et al. 2019; Kraker et al. 2016; Universities UK Open Access and Monographs Group 2019). While these principles are widely supported, they are often overlooked in research assessment practices, particularly with regard to books.

To adapt policies, policymakers require an awareness of the evolving book publishing landscape. Examining book metadata in library catalogues can offer valuable insights. Existing studies primarily focus on book citations; less is known about the broader impact of books. Yet to evaluate scholarly books comprehensively and responsibly, we need metrics that capture a book’s full journey, from creation and dissemination to digital preservation. Existing metrics often lack this detail, highlighting the need for new approaches.

Many researchers have used data from WorldCat⁶⁶, the world’s most comprehensive library catalogue, to explore book metrics. WorldCat holds extensive metadata for millions of books across thousands of libraries worldwide. This metadata is stored in MARC21 XML format, allowing researchers to analyse and utilise it efficiently. Examining this accumulated metadata can help identify practical issues such as collaborations in the book publishing industry and book metadata supply at both national and individual levels.

This study investigates the suitability of book metadata for evaluating individual books and analyses the visibility of nationally assessed books in WorldCat. Specifically, I address the following research questions:

RQ 1: Which book metadata elements required for research assessment systems are present in WorldCat and are suitable for evaluating individual books?

RQ 2: What is the level of visibility in WorldCat of books submitted to national research assessments, and who are the primary metadata suppliers?

In this empirical research, I used two datasets of ISBNs of books submitted as research outputs in the UK (Dagienė 2023c) and Lithuania (Dagienė 2023b) derived from a previous study on book evaluation (Dagienė 2024b). However, these questions might be tested using any other set of valid ISBNs and using this research methodology regardless of specific book assessment practices.

⁶⁶ Inside WorldCat <<https://www.oclc.org/en/worldcat/inside-worldcat.html>> accessed on 22 July 2024.

This paper is structured as follows: Section 6.2 presents a literature review on the development of book metrics and their main data sources. Section 6.3 describes the research methodology and data sources for the empirical study. Section 6.4 addresses the first research question, providing insights into the most relevant book metadata required for research assessment system and their equivalents in MARC21 XML format. Section 6.5 delves into book visibility in library catalogues and the main metadata suppliers for both UK- and Lithuanian-authored books. Finally, Section 6.6 summarises the findings, draws conclusions, and discusses implications for research assessment practices.

6.2. Literature review

Evaluating the impact of scholarly books presents a unique challenge compared to journal articles. In this section I review the literature on metrics and approaches for assessing the impact of scholarly books, moving beyond sole reliance on citation counts to consider diverse indicators such as book reviews, library holdings, online platforms, and altmetrics. By examining these diverse perspectives, I aim to provide a holistic understanding of the current state of book impact assessment and identify potential future directions for research.

6.2.1. Book citation metrics

At the beginning of book metric exploration, researchers attempted to assess book impact through *book citations*, as they did for peer-reviewed journals. Perhaps the first and most frequently used databases to explore book citation impact were the Science Citation Index and the Social Sciences Citation Index (Butler and Visser 2006; Cronin, Snyder, and Atkins 1997).

Further research emerged after the Book Citation Index (BKCI) was added to the Web of Science Core Collection (Gingras and Khelfaoui 2019; Gorraiz, Gumpenberger, and Purnell 2014; Gorraiz, Purnell, and Glänzel 2013; Zuccala et al. 2018). The launch of the BKCI has led scholars to identify differences between monographs and edited volumes, and between book series and annual series (Leydesdorff and Felt 2012).

Since then, research on book citation impact has become more complex in the range of sources it considers (Halevi et al. 2016; Linmans 2010; Zhou and Zhang 2021; Zuccala and Robinson-García 2019). In addition to well-established data sources, researchers have examined whether *Google Books*, *Google Scholar*, and *Scopus* (Kousha, Thelwall, and Rezaie 2011) or book mentions and citations in *Wikipedia* (Kousha and Thelwall 2017) can potentially be considered as a valuable measure of citation impact for book-based disciplines.

6.2.2. Book reviews

As complementary metrics to book citations, researchers have gradually explored *book reviews* to measure the impact of scholarly books. Among those studied to date are reviews indexed in the *Web of Science databases* (Gorraiz et al. 2014), book reviews published in a *particular journal* (Zuccala, van Someren, and van Bellen 2014), and book reviews sourced from *Choice Reviews Online* (Kousha and Thelwall 2015; Zhou and Zhang 2020).

Researchers have also examined online reviews from *Amazon* (Kousha and Thelwall 2016) and *Goodreads*, together with their ratings, for impact assessment (Kousha, Thelwall, and Abdoli 2017). They found that both can be used as evidence of the impact of popular academic books in the arts, the humanities, and, to some degree, the social sciences, although they cautioned that qualitative research is needed to verify their quantitative findings.

6.2.3. Library holdings and metrics

Library holdings have served as yet another way for scientometricians to measure book impact. Often referred to as catalogue inclusions (Torres-Salinas and Moed 2009) or lib citations (White et al. 2009), they can potentially address some of the most glaring shortcomings of citation analysis.

Researchers have compared library holdings with citations across various databases Linmans (2010), while testing new indicators for humanities, introduced library holdings from WorldCat. He derived citations from Web of Science, the library holdings from WorldCat, and the productivity data from METIS, a Dutch database covering research output. Linmans suggested that “the citation counts for books supplied by Google Scholar will be of great value”.

Building on this, Cabezas-Clavijo et al. (2013) examined correlations between library loans in university libraries and citations in Google Scholar and Web of Science databases. Zuccala & White (2015) investigated whether lib citations in WorldCat correlate with citations in Scopus. After further research, they concluded that numerous lib citations may not necessarily signal high-quality books (White and Zuccala 2018).

6.2.4. Broader impacts of books

Researchers have also combined information on library holdings with additional book metadata. For instance, in studies aimed at ranking scholarly book publishers, researchers first pinpointed book titles cited in Scopus history journals, then juxtaposed the extracted book metadata with that in WorldCat, which includes library holdings (Zuccala and White 2015). In a recent study, the researchers utilised both library holdings in WorldCat as a visibility indicator and citations in Google Scholar to measure the academic impact of books (Zuccala et al. 2021).

Many studies have analysed the broader impacts of books. In one investigation, researchers evaluated impact via meticulous analysis of citation literature. They employed three indicators: sales on amazon.co.uk; at least one citation on Google Books, Google Scholar, or Scopus; and library holdings on WorldCat (Zhou and Zhang 2021). In a separate study, Maleki (2022) modelled the relationships between print and electronic book format holdings, book citations, and altmetrics. This helped to draw out various aspects of impact made by printed and electronic books.

6.2.5. WorldCat and OCLC data

Among bibliographic databases, WorldCat, maintained by OCLC, has become the main source for data on library holdings and other book-related research. In addition to collecting library holdings, researchers (Halevi et al. 2016) have employed OCLC's book classification system to allocate books to their appropriate subject areas.

To examine whether Lotka's law for literary author productivity holds true, Friedman & Bernstein (2017) analysed the number of bibliographic records in WorldCat associated with famous authors. Other studies (Tausch 2023; Torres-Salinas, Arroyo-Machado, and Thelwall 2021; Zavalin 2023) analysed author productivity using publication records, investigated subject classifications, and explored research trends using various data points. Although WorldCat Identities is now discontinued, users can continue exploring similar data through WorldCat Entities⁶⁷.

6.3. Research design

I took a mixed-methods approach to explore book metadata associated with research outputs in the UK and Lithuania. For the empirical analysis, I used ISBNs sourced from national research assessments: 38,050 ISBNs from the UK's REF 2014⁶⁸ and REF 2021⁶⁹ (Dagienė 2023c) and 5,199 ISBNs from Lithuania's annual assessments (Dagienė 2023b). The years 2008–20 were aligned to ensure comparable representation of the findings from Lithuanian and UK data.

First, I manually reviewed randomly selected ISBNs from UK and Lithuanian (LT) books in WorldCat. Later, I quantitatively analysed metadata of all ISBNs in both national datasets, extracting the available metadata from particular MARC21 fields in the datasets provided by OCLC.

6.3.1. Evaluating WorldCat book metadata elements for individual books

To answer the first research question and determine what types of book metadata are required for research assessment systems, are suitable for evaluating individual books, and are present in WorldCat, I manually searched through book metadata using 180 randomly selected ISBNs: 90 each from the 37,641 records available in the OCLC databases for REF books and the 3,666 records for LT books. These ISBNs were used to screen the publicly accessible WorldCat bibliographic catalogue⁷⁰. The focus was on evaluating the availability of book metadata mandated for research assessments, which typically require institutions to submit data to research evaluation systems.

Primary book metadata usually encompasses elements such as author, title, publisher, language, year of publication, and ISBNs. Although ISBN codes are machine-readable, have

⁶⁷ WorldCat Entities <<https://entities.oclc.org/worldcat/entity>> accessed on 22 July 2024.

⁶⁸ REF2014 <<https://www.ref.ac.uk/2014>> accessed on 22 July 2024.

⁶⁹ REF2021 <<https://results2021.ref.ac.uk>> accessed on 22 July 2024.

⁷⁰ WorldCat catalogue <<https://www.worldcat.org/>> accessed on 22 July 2024.

the potential for automated validation, and can be used to obtain book metadata from various sources, some research evaluation systems do not give them enough credit (Dagienė and Li 2021). The UK's REF, which already mandates and automatically verifies ISBN codes, stands as an exception.

Beyond primary metadata, such as title, author, or ISBN, WorldCat also offers additional metadata. In this study, the exploration of this additional metadata commenced with advanced searches, entering the selected ISBNs, and specifying the intention to locate a particular record. The first title presented in the search results was chosen, and its associated metadata fields on the opened webpage were meticulously examined.

Subsequent exploration of available metadata types began by selecting 'Edition: View all formats and editions' on the newly opened webpage. The investigation aimed to determine whether WorldCat contains any additional metadata pertinent to current research assessment requirements, including open access status, or that could shed light on prevailing practices of scholarly knowledge production, review, organisation, dissemination, and preservation—stages identified as the most important for the future of scholarly communication (Kraker et al. 2016).

It is important to acknowledge that the WorldCat website has been subject to changes since the time of this research project. For instance, after a recent update, the website lacks the 'Responsibility' field, formerly a valuable asset due to its relevance to research assessments that necessitate peer review information for scholarly books.

6.3.2. Analysing OCLC MARC21 metadata for completeness of national datasets

To address the second research question regarding the level of visibility in WorldCat of books submitted to national research assessments and the identities of primary metadata suppliers, I delved into book metadata provided by OCLC in the MARC21 format. This allowed me to analyse the quantities of book metadata elements in national datasets.

I obtained the relevant metadata by submitting ISBNs to the OCLC team and requesting all available metadata for those ISBNs. The provided data arrived in two instalments: December 2021 (covering years 2008–14) and December 2022 (covering years 2015–20). Selection of ISBNs for LT books aligned with the corresponding years.

While OCLC processed both sets of MARC21 XML files as a combined entity, prior research has documented a noticeable English language bias in WorldCat's coverage (Torres-Salinas et al. 2021; Wakeling et al. 2017). Therefore, I conducted separate analyses examining metadata availability and completeness for books authored by researchers from the UK and Lithuania.

Book metadata in MARC21 format facilitates machine-readable representation and communication of bibliographic and related information. The specific OCLC MARC21 XML structure employs numerous fields and subfields, each corresponding to distinct book metadata elements. OCLC provides a comprehensive explanation of MARC21 fields in its

online training and support documentation⁷¹; a further series of OCLC web pages illustrates the usage of the individual MARC tags⁷².

I began investigating the empirical MARC21 data provided by OCLC identifying the fields that have every record in the datasets. Notably, a consistent subset of five fields (001, 008, 020, 040, and 245) appeared across all REF and LT records within the OCLC databases. These fields play a critical role in book identification and description, while others may be absent from specific entries or lack subfields containing the desired metadata elements.

Fields 001 and 008 are mandatory for all OCLC MARC21 records. Field 001 acts as a control field, while field 008 is a fixed-length control field containing elements common to all MARC formats, often referred to as the “leader”. For instance, leader positions 35–9 hold the language code, indicating the primary language of the catalogued item (see subsection 6.4.2). Additionally, leader positions 7–10 contain two dates, with the first representing the publication year. Notably, OCLC employs a structured field system with fixed width, allocating four positions for the year in YYYY format within the MARC21 leader (see subsection 6.4.3).

Field 020 is specifically designed to record the ISBN for the particular book. In some cases, this field may contain subfields indicating invalid or cancelled ISBNs, such as those with incorrect check digits or mismatches with the catalogued item. Furthermore, field 020 may be repeated to capture all ISBNs associated with the specific ISBN, potentially identifying different editions or parts linked to the book. While I do not explore this aspect in detail here, it may hold significance for research assessment system developers.

Field 040 serves to identify the organisations responsible for creating, maintaining, or modifying each bibliographic record. This field and its subfields provide valuable information regarding the provenance of the record, aiding in understanding its context and credibility. Multiple codes may be present within the field, each representing an organisation involved in the development of the record (see subsection 6.5.2).

Field 245 functions as the primary container for title information, playing a pivotal role in accurately representing the title and authorship of a book within the MARC21 XML metadata. Field 260 primarily deals with aspects related to the publication, distribution, acquisition, etc. of the book, complementing the title information found in field 245 (see subsection 6.4.1).

It is important to reiterate that MARC21 formats encompass a vast array of elements potentially applicable to research assessment systems or monitoring open access policy implementation. However, the subsequent sections will focus solely on the fields that are most clearly relevant to research assessment policies.

71 Bibliographic Formats and Standards. OCLC <<https://www.oclc.org/bibformats/en.html>> accessed on 22 July 2024.

72 MARC tags details. OCLC <https://help.oclc.org/Library_Management/OLIB/Data_import_and_export/MARC_tags_details> accessed on 22 July 2024.

6.4. Key book metadata elements for research assessment

Research assessment requirements vary across countries, prompting diverse interests among experts evaluating book outputs. These interests depend on the specific metadata that open access and national research evaluation policies demand.

Simulating the experiences of experts evaluating individual books or analysing national achievements, this section explores key metadata elements within WorldCat and MARC21 XML data, answering the first research question: Which book metadata elements required for research assessment systems are present in WorldCat and are suitable for evaluating individual books?

6.4.1. Book titles, authors, and variety of contributors

WorldCat draws upon MARC21 data fields to present key book metadata elements such as titles and contributors. While these elements might appear unique, experts navigating WorldCat may encounter slight variations (Figures 1–2).

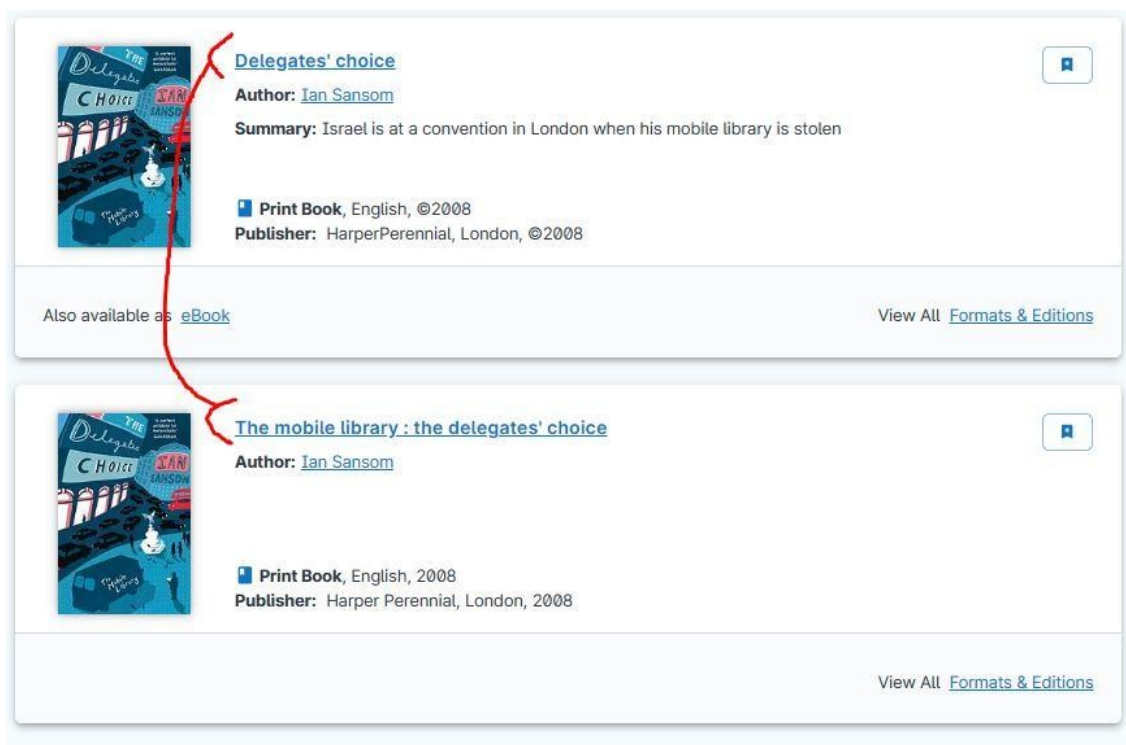


Figure 1. An example of different titles for the same book (ISBN 9780007255344).

Notably, the “Author” field can offer valuable insights, showcasing not only the book’s writer(s) but also individuals who played contributing roles such as “Translator”, “Writer of Introduction”, or “Degree Supervisor”. The composition of the “Author” field can also vary depending on the book’s genre, edition, or format, as illustrated in Figure 2.



Figure 2. An example of different compositions of the “Authors” field depending on the book genre or format: (a) one author when selecting the printed book and (b) more contributors when selecting “Thesis, Dissertation” format (ISBN 9783319706658).

Figure 3 presents an intriguing example from the Lithuanian sample, highlighting an academic who serves as both author and publisher. WorldCat identifies this researcher in both roles. Further verification through the Global Register of Publishers confirms her registered status as a Polish publisher. However, despite the listed affiliation, search engines revealed neither a dedicated website for the publishing company nor concrete evidence of the researcher’s ownership, though they did surface a published interview (Lamanauskas 2017). This interview reveals the publisher’s affiliation with the university that is mentioned among the contributors without a specific role designation. This case underscores the comprehensiveness of WorldCat in displaying all contributors identified in MARC21 data.



Figure 3. An example of contributor metadata revealing a self-published book (ISBN 9788394810429).

Analysing the underlying MARC21 data, I found that *book titles* for both REF and LT books primarily originated from field 245a, with only 0.06% of all books missing title metadata. Notably, these missing titles were primarily in languages other than English, subsequently incorporated into WorldCat manually.

However, the picture for *author* and *contributor* data, primarily sourced from field 245c, is slightly different. Here, a more substantial data gap exists, with missing information for 11.9% of REF books and 10.1% of LT books. To address these gaps, I employed manual checking and merged data from various designated fields (100, 110, 700, etc.). While this process improved completeness, the extracted data often required further cleaning and standardisation for accurate author-specific analyses.

Overall, while WorldCat offers valuable information about book titles, authors, and contributors, users should be mindful of potential inconsistencies and limitations in the data. Manual intervention and further cleaning may be necessary for specific analyses, particularly those focusing on individual authors or contributors.

6.4.2. Book language: Representation and trends

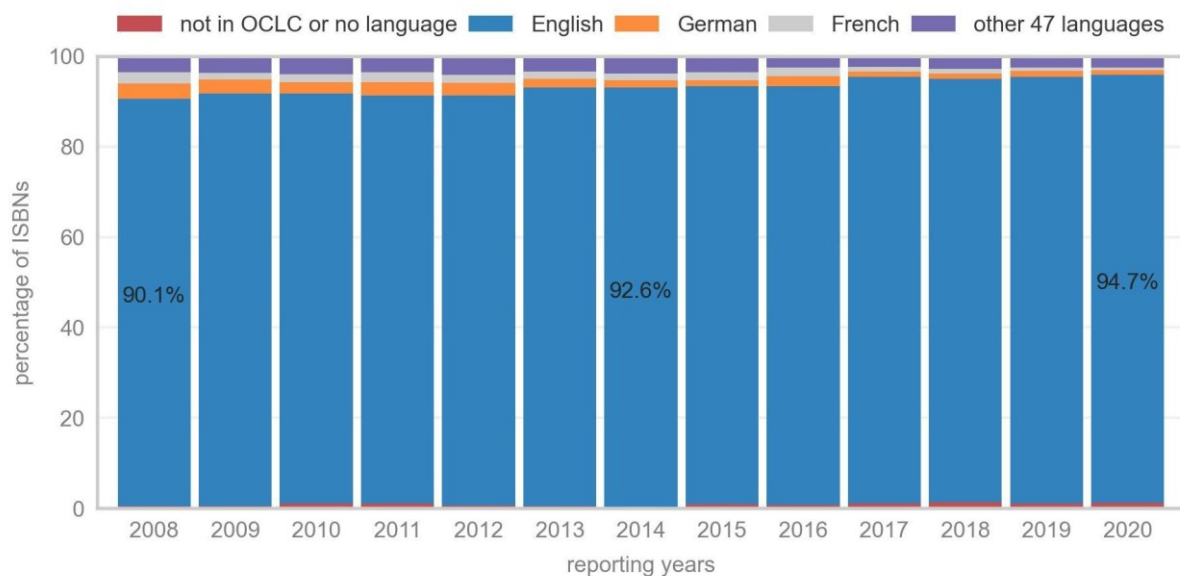
WorldCat prominently displays book languages, allowing users to quickly identify the primary language of a work. For multilingual books, separate records may exist (Figure 4), highlighting diverse contributions and intellectual origins as well. The information on languages is further encoded within the mandatory MARC21 “leader” field (positions 35–7) using three-letter language codes.



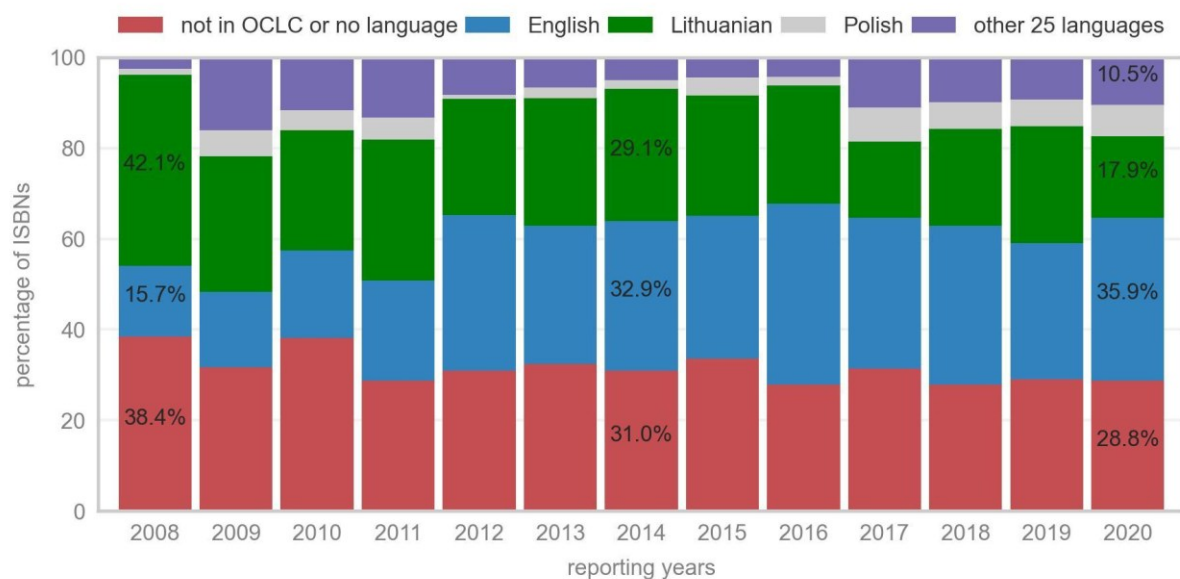
Figure 4. Split records for the same book highlight multilingualism (ISBN 9789955185611).

While REF books were written in 48 languages and LT books in 26, a relatively small portion lacked clear language identification: 0.83% of REF records (313 of 37,641) and 0.76% of LT

records (28 of 3,666). These ambiguities stem from the limitations of the three-letter code system, which can represent various meanings such as “Undetermined” or “Multiple languages”.



a)



b)

Figure 5. Top three vs. other languages with annual trends for (a) REF and (b) LT books.

Additionally, collective codes indicate broader language groups (e.g. Slavic or Romance), making precise identification difficult if the particular languages are not further specified. Fortunately, MARC21 field 041 offers more detailed language information, though my investigation found that many records with unclear leader codes lacked corresponding details in this field. This highlights the potential inconsistencies and gaps within available metadata, but only for less than 1% of all national records. Figure 5a illustrates that over 90% of REF books have been published in English, while the percentage of books in German and French is declining. Figure 5b reveals a significant decline in the number of books in Lithuanian from

42.1% to 17.9% between 2008 and 2020, while English publications doubled during the same period. This trend might be partly explained by national research assessment policies that incentivise publishing with “prestigious” foreign publishers, who often publish in English.

Altogether, this analysis of book languages in WorldCat and MARC21 metadata yielded valuable insights into language trends within research publications, despite some missing data. The data confirms the great predominance of English in REF books while illustrating the growing tendency towards bilingual publishing among Lithuanian institutions, where English is gaining ground alongside Lithuanian. This shift warrants further exploration to understand the driving factors, such as national research assessment policies, and their potential impact on the visibility and accessibility of research outputs.

6.4.3. Publication year: Completeness and challenges

While the content of a book holds more weight than its publication date in expert evaluations, accurate publication years remain crucial for bibliographic research and identifying trends across specific timeframes. Like in an earlier study (Dagienė 2024b), in this paper I rely on publication years reported by institutions during research assessment submissions. However, the results show that reporting years are not always the same as the publication years found in library catalogues.

Fortunately, union catalogues such as WorldCat (MARC21 metadata) offer a reliable alternative source for publication year information. The mandatory MARC21 fields hold publication year data, making them ideal for retrieval. The MARC21 “leader” field emerges as the primary and most comprehensive source for publication years, offering an impressive coverage of 99.2% for REF books and 97.5% for LT books available in the OCLC databases. I identified only minor discrepancies due to variations in data format and the presence of manually entered mistaken information. This highlights the crucial role of MARC21 metadata in ensuring data completeness and accuracy for bibliometric analyses.

Field 260c, serving as an additional source for languages other than English, contained publication years for a third of REF books and two-thirds of LT books. Unstandardised manually recorded publishing and copyright dates made the data in this field inconsistent and difficult to process automatically, though it remains helpful in cleaning and complementing the years obtained from the “leader” field.

With up to 1% of OCLC entries missing year data overall, comparing the reported years with those retrieved from OCLC reveals a substantial level of agreement: 78.0% for REF books and 79.9% for LT books. Nevertheless, some discrepancies exist, involving publications that occurred earlier or later than the reported dates. 18.8% of REF and 17.2% of LT books were published one to three years earlier than the dates reported by institutions, while 3.0% of REF and 1.7% of LT books were published one to three years later than the dates reported by institutions (delayed publications).

These discrepancies notwithstanding, the analysis of MARC21 metadata within library catalogues holds valuable potential for retrieving more accurate publication years for almost

all the REF and LT books available in OCLC. However, exploring alternative data sources remains necessary for the one-third of LT books not included within the OCLC databases.

6.4.4. Book genres: Challenges in standardisation and consistency

The UK's REF policies define books as long-form research outputs, but in some countries, including Lithuania, diverse types of books play a significant role in research assessments. Policies may prioritise specific genres, often without unambiguous definitions or standardised application across contexts. Bibliometricians have investigated these practices, highlighting the inconsistent interpretation of terms such as “(scholarly) monograph” by different stakeholders, including policymakers, academics, publishers, and librarians (Clemens et al. 2010; Dagienė 2024b; Sile et al. 2021; Zuccala et al. 2018; Zuccala and Cornacchia 2016).

Research evaluation requirements often diverge from the publishing industry standard: the ISBN Manual (International ISBN Agency 2017) lacks definitions for “book”, “book genre”, or even “scientific monograph.” Conversely, library cataloguing rules (International Federation of Library Associations and Institutions (IFLA) 2009) define a book as a “work” representing an “intellectual or artistic creation”, with variations such as translation considered a separate “expression”. The physical manifestation of an “expression” can be diverse, encompassing print, electronic formats, and other media. This suggests that publishers and library catalogues may not readily capture book genres in the way book assessment protocols perceive them.

Analysing REF and LT book samples in WorldCat, I found a wide variety of terms in the “Genres” field: “academic dissertation”, “novels”, “conference papers”, “biographies”, “poetry”, and many others in different languages. Exploring the completeness of genre metadata for these books, I extracted data from MARC21 field 655a, “Genre/Form”, which is optional; consequently, 25.2% of REF and 59.0% of LT books had no such field in their MARC21 records. Even where data existed, hundreds of unique words in multiple languages described not only typical genres but a diversity of forms, characteristics, and indeed categories, though the latter were not necessarily genres eligible for Lithuanian research assessment. As recognised earlier, the REF policies require long-form research outputs but do not stipulate specific genres for submission.

Figure 6 shows the 7 most prevalent genres among the REF and LT books identified in MARC21 field 655a. In the Lithuanian data, we see a mismatch between the top identified genres and those required by national research evaluation policy. Formats such as “electronic books” and “edited volumes” dominate, while “monographs” are rare.

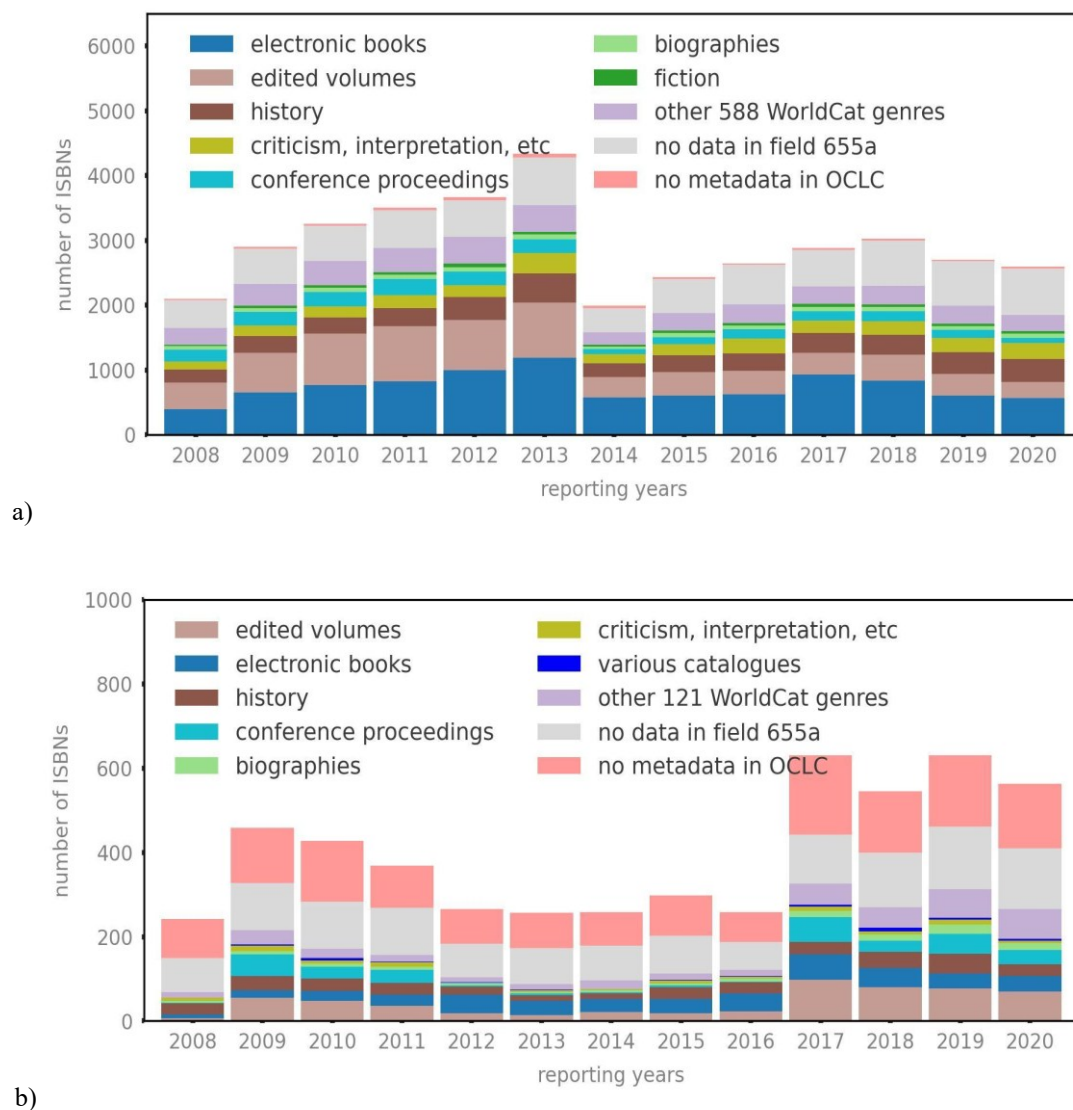


Figure 6. Top 7 vs. remaining genres for (a) REF and (b) LT books identified in MARC21 metadata field 655a.

Prior to this research, I expected large, globally recognised publishers to handle book metadata and provide it to library catalogues. Based on this belief and my previous research, showing that UK and US publishers issued about 80% (and the top 10 publishers roughly 60%) of all REF-assessed books (Dagienė 2024b), I expected that the genres of REF books I obtained from MARC21 would be predominantly in English. Surprisingly, the REF data revealed a variety of genres across multiple languages. Figure 7 shows that identifying categories such as “electronic books” or “conference proceedings” required aggregating records with various terms in German, Polish, French, and 26 other languages in which book metadata was supplied (see also subsection 6.5.2).

```

electronic_books = ['e-book download', 'e-book online only', 'e-books', 'ebook', 'ebooks', 'electronic books', 'eleconic books',
                    'electronic book', 'electronic books', 'elektronische publikation', 'elektronisches buch',
                    'elektronische publikation, publications électroniques', 'libros electronicos', 'libros electrónicos',
                    'livre électronique (descripteur de forme)', 'livres électronique', 'livres électroniques', 'livres numériques',
                    'livres num@ériques', 'livres numériques', 'llibres electrònics', 'llibres electrònics',
                    'publications électroniques']

conference_proceedings = ['actas de congresos', 'actes de congrès', 'actes de congrès', 'conference "cultures and societies in the
middle euphrates and habur areas in the second millennium bc: scribal education and scribal traditions", 5-6 december 2013,
held in tsukuba, japan', 'conference materials', 'conference papers and proceedings', 'conference proceeding',
'conferences - meetings', 'conferencias', 'congresos', 'congresos, conferencias, etc', 'congress',
'congressen (vorm)', 'congresses', 'congresses (form)', 'congressi', 'congresso', 'congressos',
'congresverslag', 'congrès', 'congrès', 'konferenser', 'konferenz', 'konferenzschrift', 'konferenzschrift 2014',
'kongresi', 'kongress', 'kongress (zaragoza ; 2004)', 'kongress ca', 'kongreß', 'materiały konferencyjne',
'proceedings', 'proceedings of conferences']

```

Figure 7. Examples of terms found in MARC21 field 655a (reserved for genres) and used to count the number of electronic books and conference proceedings in the REF and LT books

The observed variations in genre terms highlight the lack of a single standard for genre assignment in book cataloguing. Additionally, misspellings and other errors suggest that some records were manually added.

The metadata also reflects national genre restrictions and their concomitant inconsistencies. For instance, the REF assessment process allows for submissions of various genres, including fiction and poetry, which appear among the top 20 genres identified in the UK data. However, Lithuanian research assessment policies restrict eligible genres, excluding fiction and poetry. The most highly valued, attracting significantly more funding points than other qualifying genres, are “scholarly monographs” and “collective monographs” (i.e. edited volumes published by “prestigious publishers”). In the Lithuanian data, only 17 books are classified as “monographs”, but edited volumes dominate (Figure 6b).

The ambiguity surrounding the definition of ‘scholarly monograph’ in Lithuanian policies creates uncertainty in how anonymous experts decide on book genres and the associated funding points. Authors and domestic publishers often include genres on title pages, presumably to influence expert assessment. In light of this, it is unsurprising that of the thousands of books available in OCLC databases, only two REF books—but almost 300 LT books—have the word “monograph” in their MARC21 title data.

This analysis highlights the significant challenges associated with using book genres for research evaluation. Discrepancies exist between research assessment policy priorities, book publishing industry standards, and library cataloguing practices, leading to inconsistencies and ambiguities in genre data. Notably, formats such as electronic books often dominate, while genres explicitly required by policies (e.g. “scholarly monographs”) are scarce. This mismatch creates dilemmas for researchers and book evaluators, raising concerns about fairness and effectiveness in research assessment.

Moving forward, further research is necessary to investigate the potential for standardised genre classification systems. Only by addressing these challenges can we ensure the reliable and consistent use of book genres for research assessment purposes.

6.4.5. Publishers: Complexities of collaboration and practice

In research evaluation, publisher prestige reigns supreme, yet the name alone can conceal a labyrinth of collaborations and practices. This subsection exposes the complexities behind the ‘Publisher’ field, revealing undisclosed partnerships and diverse publishing models that impact our understanding of scholarly works (Figure 8). Delving into WorldCat and MARC21 metadata reveals the challenge of identifying publishers and the need for more nuanced and effective research evaluation.

a)

Title	Format	Language	Year	Publisher
 New paradigms in public policy by Peter Taylor-Gooby (Editor), British Academy	eBook	English	2013	Published for the British Academy by Oxford University Press, Oxford
 New paradigms in public policy by Peter Taylor-Gooby (Editor)	eBook	English	2013 First edition	Oxford University Press, Oxford

b)

Title	Format	Language	Year	Publisher
 Early Islamic Iran by Edmund Herzig, Sarah Stewart, London Middle East	Print Book	English	2012	I.B. Tauris in association with the London Middle East Institute at SOAS and the Faculty of Oriental Studies, University of Oxford ; Distributed in the United States and Canada exclusively by Palgrave Macmillan, London, New York
 Early Islamic Iran by Edmund Herzig	Print Book	English	2012	Tauris, London
 Early Islamic Iran by Edmund Herzig, Sarah Stewart (Author)	Print Book	English	2012	I.B. Tauris, New York

Figure 8. An example of a variety of ‘Publisher’ field names associated with the same book title:
(a) ISBN 9780197264935 and (b) ISBN 9781780760612.

Figure 8a exemplifies this complexity, where one record lists only Oxford University Press, while another reveals its partnership with the British Academy. This obscurity intensifies in Figure 8b, showcasing a book with four collaborators: I.B. Tauris, the London Middle East

Institute, the Faculty of Oriental Studies at the University of Oxford, and Palgrave Macmillan. Evaluating such books solely based on publisher name (or prestige) becomes challenging, highlighting the need for deeper analysis. These are not isolated cases: numerous books exhibit similar discrepancies in their “Publisher” field details.

My exploration extended beyond the main “Publisher” field, uncovering valuable information hidden in the “Notes” and “More Information” sections. One book is credited to Crawford House Publishing, with “Notes” specifying a collaboration with the University of Cambridge Museum of Archaeology and Anthropology. Another listed Gandon Editions for Irish Architecture Foundation [and] Irish Architectural Archive, with “Notes” adding publication coinciding with an exhibition. Even ISBNs, often overlooked, unlock hidden connections.

Even more, WorldCat and MARC21 data of ISBNs in UK sample coupled with same ISBNs metadata from Global Register of Publishers revealed a network of partnership linking entities such as Gower, Ashgate, Routledge, the Hakluyt Society, MyiLibrary, and ProQuest. These findings underscore the crucial role of comprehensive data exploration in understanding diverse publishing practices.

A focus on book publication quality, professional digital formats, discoverability, and long-term preservation must take the place of traditional publisher prestige (Kraker et al. 2016). Current assessment often relies heavily on publisher reputation, but metadata reveals that societies, institutions, and museums often curate content before publishers become involved. This suggests focusing on the selection process and publication quality rather than solely on publisher reputation. Such a focus becomes even more critical when “Publisher” fields vary across editions due to mergers and acquisitions, as seen with a book (9781409432661, 9781283367738) whose successive editions were released by Ashgate (2011, 2012), Routledge (2016), and Taylor & Francis (2016).

While multiple ISBNs hint at undisclosed partnerships, the current process to identify them requires manual effort, highlighting the need for standardised and machine-readable metadata to fully comprehend the complexities of publishing practices. Disappearing imprints and changing publishers underscore the critical need for long-term preservation of scholarly books. This includes creating discoverable and freely accessible digital book versions, available in any language and distributed over the internet, even when the original publishers no longer exist. Such digitization ensures accessibility for future generations and aligns with the principles of open science. Additionally, the presence of multiple ISBNs suggests the involvement of various publishers depending on formats and editions, yet current metadata lacks information on the discoverability of digital publications. Further investigation and data enrichment are crucial for a more holistic understanding of the publishing landscape.

Progress towards standardisation has already been made. Rich MARC21 data with WorldCat Entities helps reveal contributors and connections invisible in an ordinary WorldCat search. Utilising metadata from the Global Register of Publishers can further simplify the abundance of publisher titles and imprints (Dagienė 2024). Integrating these resources, along with more comprehensive MARC21 data, presents promising opportunities for book metadata improvement.

By addressing these challenges and leveraging these opportunities, a more precise and comprehensive understanding of book outputs can be obtained, ultimately contributing to fairer and more effective research evaluation practices.

6.4.6. Forms of publication: Print vs. electronic, open access, translations, and editions

The REF sample presents a comprehensive cross-section of books, each with a plethora of editions, translations, and formats disseminated by various publishing houses. This diversity highlights the expansive nature of book metadata, which is readily accessible within both WorldCat and MARC21. To illustrate the vast scope of metadata accumulated over time, I will delve into the specific case of *The Lessons*, analysing the information available in WorldCat for individual book evaluation.

Classified in WorldCat as a fictional work within the “College Students England Oxford” subject area, *The Lessons* transcends publisher boundaries. Viking (an imprint of Penguin Random House), W. F. Howes (UK), and various French and Italian publishers have released various formats and translations. Assessing this work solely through its publisher would prove challenging. Viking released the first English edition in 2010, and subsequent editions followed in 2011 and 2018. Australian publisher Bolinda later issued an audiobook in 2019. Furthermore, the book received an Italian translation in 2020 and French translations across three editions published between 2010 and 2012 (Figure 9). This exploration of *The Lessons* metadata exemplifies the intricate web of publishers, editions, and languages that enrich the bibliographic landscape.

Title	Format	Language	Year	Publisher
Mauvais genre by Naomi Alderman	Print Book	French	2011	Ed. de l'Olivier, Paris
Le lezioni by Naomi Alderman, Silvia Bre (Translator)	Print Book	Italian	2020	Nottetempo, Milano
Mauvais genre : roman by Naomi Alderman (Author), Hélène Papot (Translator)	Print Book	French	DL 2012	Points, [Paris]
Mauvais genre by Naomi Alderman	Print Book	French	2010	Éditions de l'Olivier, [Paris]

Figure 9. Multiple entries for *The Lessons*, which since 2010 has been issued by different publishers and translated into French and Italian (ISBN 9780670916290).

The metadata for each specific book is scattered across various MARC21 fields, making it challenging to explore the completeness of the metadata for all REF and LT books and report it in this article. However, future research could delve into this variety of metadata and its completeness. Technologies can help create a comprehensive picture for each book based on its metadata scattered across many related MARC21 fields. This could aid in the research assessment of books when evaluating the impact of the book and the research behind it.

Research by Ozaygen (2019) and by Neylon et al. (2018) demonstrates that open access books achieve significantly greater visibility and international reach. Unlike their counterparts confined to library shelves or researchers' drawers, open access publications break down barriers, fostering wider engagement and knowledge exchange across borders. Thus, when it comes to open access, the primary format is digital. As anticipated, WorldCat now includes a feature that shows open-access books. Some titles in eBook format display the "Access Free | Open" icon in the top right-hand corner. By clicking on this icon, registered users can access the Internet Archive and read the scanned book for free for one hour.

However, clicking the "DOI" field instead (under "Show more information") shows the eBook price on the publisher's website—hardly the open access that is expected in related recommendations and policies. One might interpret this feature as OCLC's experimental attempt to connect books with digital formats that will be truly open access and properly licensed.

The combination of book formats, editions, and translations, along with publication years, can provide valuable insights into the social impact of books. However, book evaluators can only glean these insights if they assess them over a period of time, not just after the book is published. It would be fascinating to conduct research that combines the extremely rich metadata on all these aspects already available in the MARC21 format.

6.5. The visibility of books in library catalogues

Answering the second research question involves examining the visibility in WorldCat of books submitted to national research assessments. I analyse REF and LT book metadata in specific MARC21 fields, focusing on book formats, library holdings, and metadata completeness. With the UK's REF mandating open access for books and Lithuania's emphasis on the status of publishers as examples of contrasting approaches to research evaluation, I seek to identify potential improvements to book visibility in libraries. Understanding the discoverability and accessibility of scholarly books is crucial for maximising their impact, regardless of policy variations.

6.5.1. Preserving knowledge: Insights from library holdings analysis

This subsection explores the "Library holdings" indicator in WorldCat, highlighting critical challenges in long-term scholarly knowledge preservation. Book discoverability, access to the entire book content, and preservation are intrinsically linked for the future of scholarly communication. Publisher mergers and disappearances, exemplified by the case of LAP LAMBERT Academic Publishing, emphasise the need for accessible digital copies through

libraries or the internet, underscoring the crucial role of digital archiving in ensuring knowledge remains available. While OmniScriptum (the parent publisher of LAP LAMBERT) claims UK and Moldova headquarters, ISBN prefixes point to Mauritius, and copyright pages list German publishers. Their limited print-on-demand model raises concerns about book accessibility. For instance, no library holdings exist for the 2012 book with ISBN 9783848445844 (Figure 10). More such books were found in both the REF and the LT sample.

DEUTSCHE NATIONALBIBLIOTHEK

Kontakt A-Z Träger / Förderer Datenschutz Impressum Hilfe Mein Konto English

↓ Katalog

→ Einfache Suche

→ Erweiterte Suche

→ Browsen (DDC)

→ Suchverlauf

→ Meine Auswahl

→ Hilfe

→ Datashop

→ Mein Konto

→ Ablieferung von Netzpublikationen

→ Informationsvermittlung

KATALOG DER DEUTSCHEN NATIONALBIBLIOTHEK

Gesamter Bestand Musikarchiv Exilsammlungen Buchmuseum

→ Suchformular zurücksetzen

9783848445844 Finden → ☐ Expertensuche ?

Leichte Bedienung, intuitive Suche: Die Betaversion unseres neuen Katalogs ist online! → Zur Betaversion des neuen DNB-Katalogs

Noch nicht die passende Literatur gefunden? → Book a Librarian

Ergebnis der Suche nach: "9783848445844"

Link zu diesem Datensatz	https://d-nb.info/1023986760
Titel	Why Greeks Interrupt each other? : The phenomenon of 'overlaps' in everyday Greek conversations / Athanasia Chalari
Person(en)	Chalari, Athanasia (Verfasser)
Ausgabe	neue Ausg.
Verlag	Saarbrücken : LAP LAMBERT Academic Publishing
Zeitliche Einordnung	Erscheinungsdatum: 2012
Umfang/Format	Online-Ressource
Persistent Identifier	URN: urn:nbn:de:101:1-201207063996
ISBN/Einband/Preis	978-3-8484-4584-4
Anmerkungen	Lizenzpflichtig. - Vom Verlag als Druckwerk on demand angeboten
Sachgruppe(n)	480 Griechisch
Online-Zugriff	Archivobjekt öffnen

DEUTSCHE NATIONALBIBLIOTHEK

English Contact AZ Sponsors data protection imprint Help My account

ACCESS RESTRICTED

Access to the publication is only possible on the reading room computers of the German National Library.

→ to the dataset: <https://d-nb.info/1023986760>

→ to the catalogue

Deutsche Nationalbibliothek Leipzig Frankfurt am Main

Figure 10. Record of a book issued by OmniScriptum is available at the national library (<https://portal.dnb.de>) with restricted access under legal deposit rules (ISBN 9783848445844).

Fortunately, legal deposit rules enable digital access in German national libraries' reading rooms, highlighting the potential of legal deposit for preservation (Figure 10b). By law, publishers or distributors in many countries must submit copies of their publications (usually

books) to a designated library or archive. Libraries designated by the government, often the national library, act as official depositories. The number of copies varies, with some countries requiring one and others multiple, depending on format or edition. Primarily, legal deposit aims to preserve national heritage and ensure comprehensive national bibliographic records. Other benefits include supporting scholarly research, enhancing library collections, and facilitating information access for all. Figure 11 illustrates the evolving distribution of UK and Lithuanian book holdings across libraries.

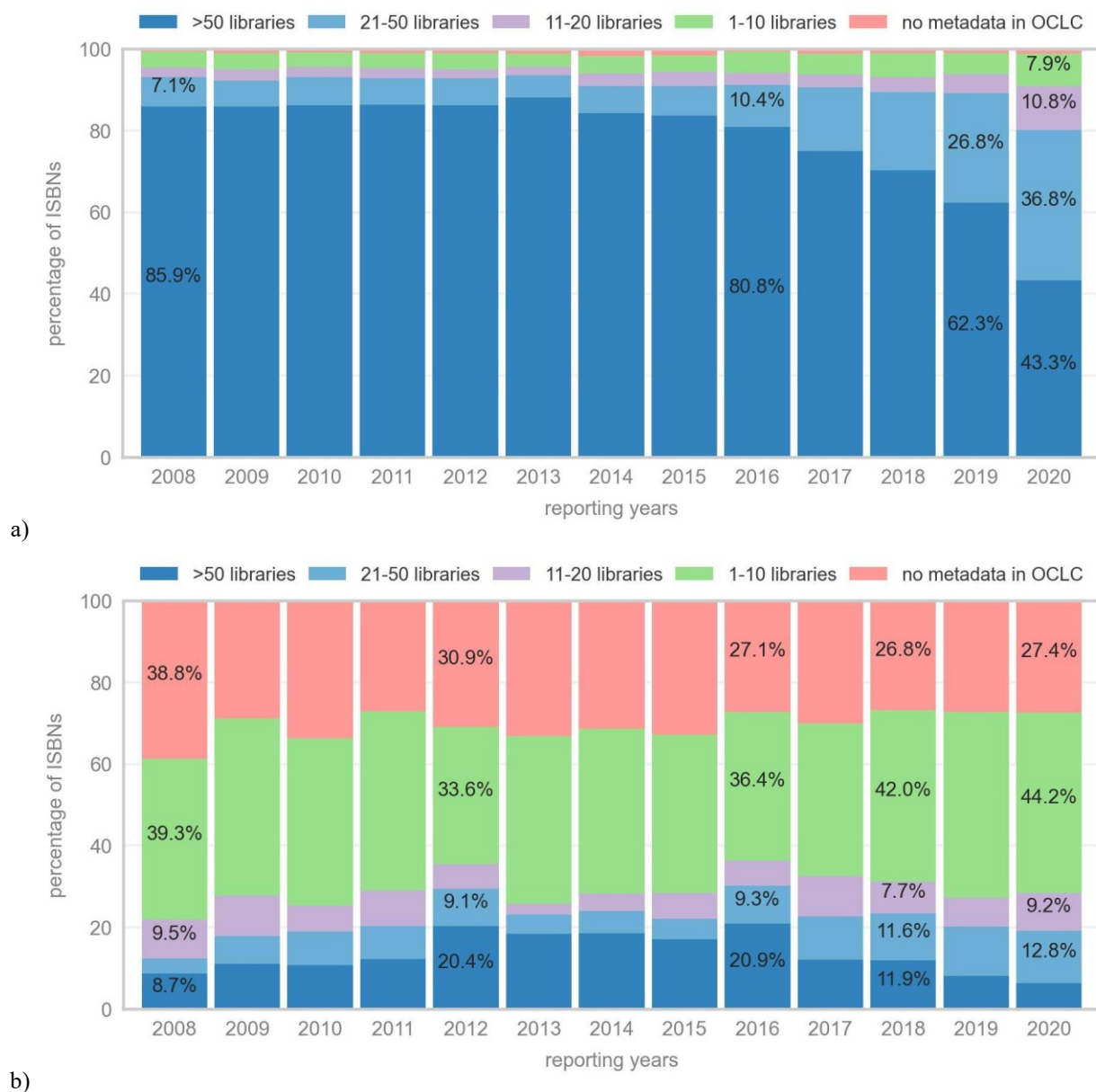


Figure 11. Library holdings and availability of UK (a) and Lithuanian (b) books in the OCLC databases.

While the proportion of UK books in over 50 libraries halved from 2008 to 2020 (Figure 11a), a significant number remain widely available. In contrast, Lithuanian books (Figure 11b) show limited international reach, with most holdings concentrated in fewer libraries. This underscores the need for broader distribution strategies to ensure wider accessibility.

Most concerningly, nearly a third of Lithuanian books are entirely absent from WorldCat, with the majority published in Lithuania itself. This raises concerns about discoverability and international access, highlighting the need for dedicated efforts to increase their visibility.

As this analysis shows, collaborative efforts are urgently required to address the challenges of long-term knowledge preservation. From supporting open access initiatives to strengthening legal deposit frameworks and promoting digital archiving best practices, collective action is crucial to ensuring the enduring availability of scholarly knowledge for future generations.

6.5.2. Providers of book metadata: Top contributors

The absence of many Lithuanian titles in OCLC databases prompted me to investigate the libraries contributing most to book visibility in WorldCat. Further analysis of the metadata helped to explain why half of Lithuanian-origin books are missing from the databases, while suggesting directions for further research into preservation strategies. The results revealed a rich multilingual landscape, with 26 languages used for cataloguing the REF books and 19 for the LT books. To identify the contributing institutions, I examined MARC21 field 040, which contained the source organisation and language used for the original record creation. This revealed 995 distinct MARC codes of organisations contributing to REF books' metadata and 311 for LT books.

For REF books, the top 10 contributors (Figure 12a) accounted for just slightly fewer titles than the remaining 985 institutions (Figure 12b). Notably, since 2018, these top 10 contributors have catalogued approximately as many books as the rest. Further investigation revealed a geographically diverse group, with institutions from Germany, Denmark, the United States, Switzerland, France, and the Netherlands, but only one UK institution—the British Library Group Batchload. Interestingly, none of the top 10 were publishers, but rather resellers, libraries, and bibliographic data professionals.

For LT books, the top 10 contributors (Figure 13a) catalogued more books than the remaining 301 institutions (Figure 13b). Moreover, their output doubled between 2017 and 2020 (Figure 13a). This group comprises national and university libraries, commercial partners, and resellers from the United States, Germany, Poland, and the United Kingdom. Notably, no Lithuanian institution appears among the top 10 contributors. Furthermore, the absence of Lithuanian among the cataloguing languages suggests that the three Lithuanian academic libraries listed in the OCLC membership do not contribute book metadata.

These findings highlight the need for further investigation into the factors limiting Lithuanian book visibility in WorldCat. Future research could explore the reasons behind the lack of Lithuanian libraries among major contributors, potential barriers, and strategies for increasing the representation of Lithuanian titles in global bibliographic databases.

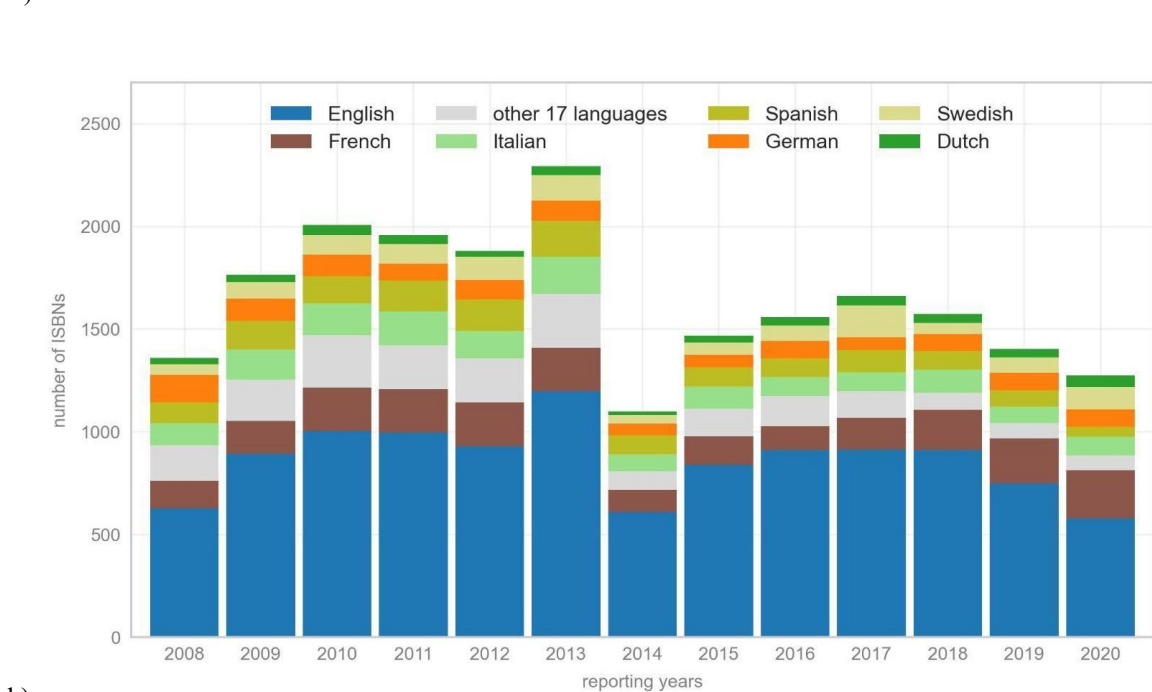
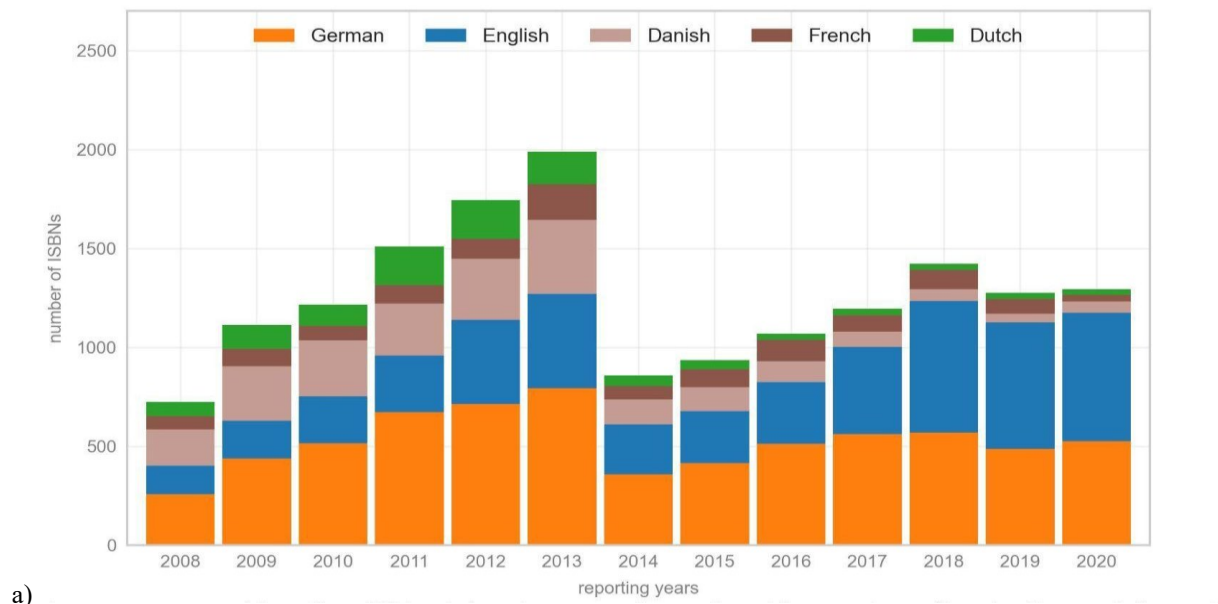


Figure 12a. Top 10 vs. other contributors to REF books' metadata availability in OCLC databases by language: (a) top 10 cataloguing agencies by number of books; (b) the remaining 985 agencies contributing to REF books' metadata availability in OCLC databases by language. Each of the "other 17 languages" was used to catalogue <1000 books over this period.

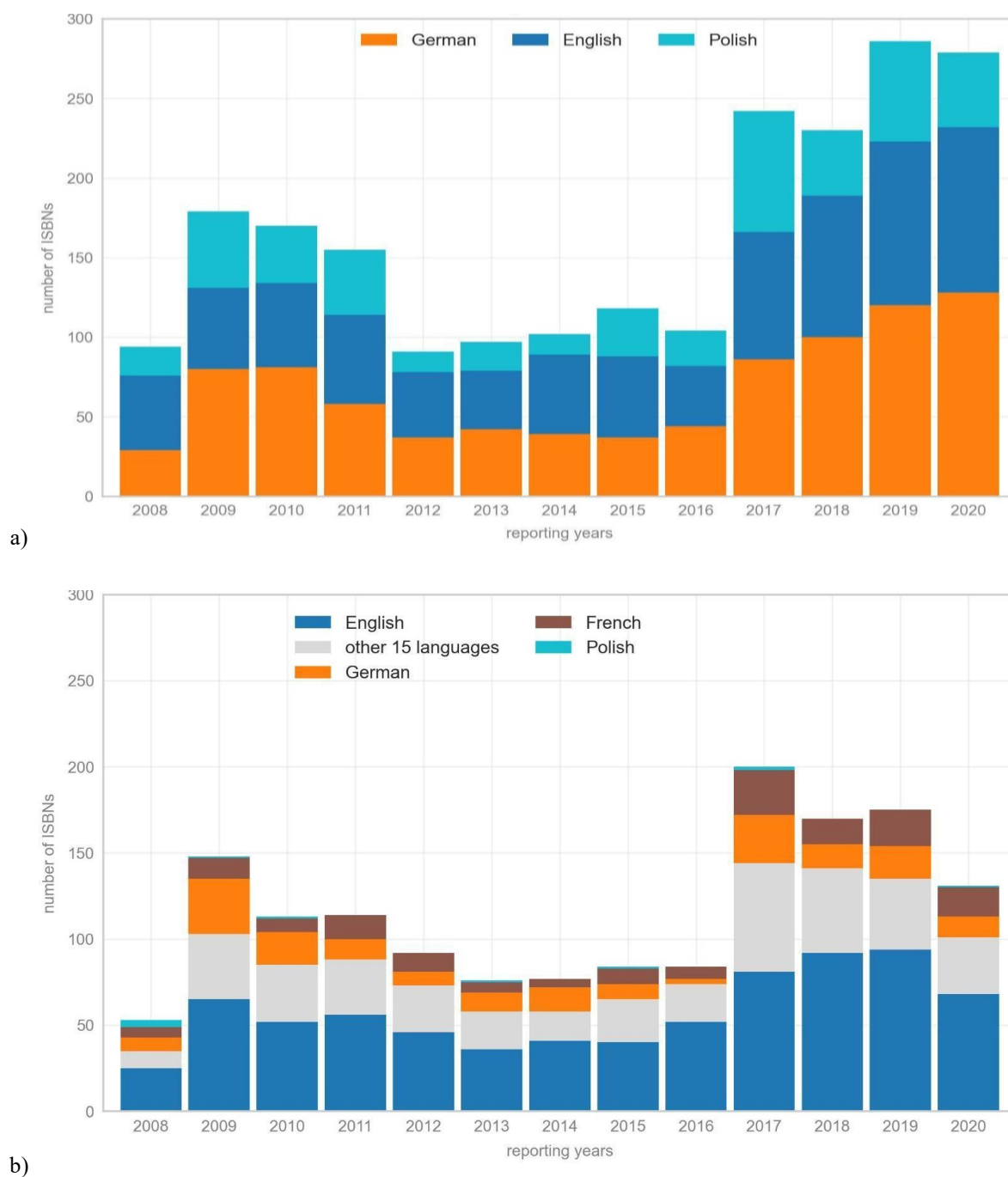


Figure 13. Top 10 vs. other contributors to LT books' metadata availability in OCLC databases by language: a) top 10 cataloguing agencies by number of books, b) the remaining 985 agencies. Each of the "other 15 languages" was used to catalogue <100 books over this period.

6.6. Discussion and conclusions

This research explored the suitability of book metadata available on the WorldCat catalogue and its MARC21 data fields for evaluating individual books, with particular attention to features relevant to scholarly book evaluation.

RQ 1: Which book metadata elements required for research assessment systems are present in WorldCat and are suitable for evaluating individual books?

As the results show, WorldCat and OCLC MARC21 XML records have nearly complete coverage of book metadata elements necessary for research assessment, such as *titles*, *languages*, and *publication years*. These book metadata elements are ideal for automated processing and bibliometric analysis.

While MARC21 XML data fields generally provide comprehensive contributor information (*authors*, *editors*, *translators*, etc.), approximately 10% of records require additional scrutiny and manual cleaning after automated processing.

Additionally, two book metadata elements, *genres* and *peer-review status*, are often absent in MARC21 but may be present in other systems depending on specific evaluation requirements. Some countries integrate and have this book metadata into national repositories or other systems designed for research assessments purposes (Sile et al. 2017). For example, books in the Lithuanian Academic Electronic Library (eLABa) catalogue contain metadata fields on book genres as defined in national research assessment policies. In contrast, the UK's REF datasets identify only authored books and edited volumes because REF policies do not require peer review or any particular genre for 'long-form research' outputs (the term used for monographs and other scholarly books).

Although MARC21 *publisher* related fields contain non-standardized information and are not suited for automated processing, they can be used for qualitative analysis, particularly when combined with fields containing relevant notes. Instead of relying on MARC21 data fields, publisher names can be obtained from the Global Register of Publishers using book ISBN prefixes (Dagienė 2024b). Building on this, future research could assess how publisher names are presented in national lists (Pölönen et al. 2021) and their alignment with both OCLC data and the Global Register of Publishers, potentially informing the development of more standardized and reliable publisher information within research assessment systems.

Finally, the results highlight the richness of library resources in terms of formats, editions, and translations, but reveal that the open access feature in WorldCat is still experimental and does not directly indicate open access.

Overall, while WorldCat and MARC21 are valuable resources for book-level research assessment, their limitations regarding peer review, genre, and open access information should be considered in future research and development efforts.

RQ 2: What is the level of visibility in WorldCat of books submitted to national research assessments, and who are the primary metadata suppliers?

Book visibility in WorldCat depends on book acquisitions and metadata providers. This research found that German libraries significantly contribute to REF book visibility due to their position as top metadata providers. All REF books have MARC21 records, but completeness varies. Interestingly, no publishers were among the top 10 metadata suppliers, with resellers, libraries, and metadata professionals taking the lead.

Only two-thirds of LT books have MARC21 metadata and WorldCat visibility. No Lithuanian institutions currently supply MARC data to OCLC. As a result, German, American, and Polish data providers are primarily responsible for the visibility of Lithuanian books. Notably, German national libraries contribute data even for non-library-held books digitally preserved in their archives. Based on these findings, I recommend encouraging the Lithuanian National Library to become an OCLC member and contribute national book metadata as its German counterparts do.

Understanding contributors to book availability in OCLC can offer valuable insights for addressing visibility concerns around national research outputs.

Limitations, future research, and implications

This study acknowledges limitations in the completeness of its metadata analysis due to the complex nature of MARC21 fields containing forms of publications which were not requested for this project. Additional MARC21 metadata for book ISBNs are crucial if comprehensive analysis of publication forms is intended.

Even though the potential of MARC21 data is evident, as it can support compiling comprehensive metadata and creating rich metrics for individual books, further research is needed to explore the data ownership and technical constraints. Additionally, further research is needed to investigate long-term preservation and data availability in library catalogues.

Ultimately, collaboration among researchers, librarians, publishers, and metadata providers is crucial for developing and implementing effective standardised solutions for completeness of book metadata suitable for research assessment. By fostering such collaboration, we can build a more robust and informative publishing landscape, empowering researchers and evaluators to make informed decisions and transform book evaluation practices.