Catalogue & Index

Periodical of the Metadata & Discovery Group, a Special Interest Group of CILIP, the Library and Information Association

September 2026, Issue 216

ISSN 2399-9667

1-4 **Editorial**
The Editors

5-20 **More time needed**
James T. Clark and Getaneh Alemu

21-27 **Authentic and artificial cataloguing**
Hannes Lowagie

28-39 **Can AI or Python help us catalogue books we can't read?**
Adam Swann and Sally Bell

40-44 **"All our wisdom is stored in the trees"**
Helen K. R. Williams

45-52 **The Metadata Team and institutional research visibility**
Helen K. R. Williams

53-58 **Metadata as cultural stewardship**
Bláithín Hurley

59-68 **Unlocking Southeast Asia's digital heritage**
Chuin Peng Sim and Gandhimathy Durairaj

69-76 **Linked data and OCLC Meridian**
Sarah Glaccum

77-86 **Collaborative cataloguing now!**
Amy Staniforth

87-98 **Metadata in motion**
Fran Frenzel

99-100 **Book Review: Information Science**
Sergio Alonso Mislata

101-115 **Minutes of the 230th MDG Committee Meeting 12 March 2026**

EDITORIAL

Welcome to the September (216) issue of Catalogue & Index which features papers from this year's conference themed around 'metafutures'. Back in March the Metadata & Discovery Group's conference was held at Engineers' House, Bristol. Two packed, informative days, followed by a third day arranged by UKCoR on RDA. Whether you were unable to attend or were there in person joining in the conversations, we are now happy to share several conference papers written up specifically for C&I. We will also be including some further papers in the December (217) issue.

The exploration of the use of AI features in several papers but one thing that appeared to stand out was the idea that "Just because we can..., doesn't mean we should..." and that we should always consider what is the right tool for the job. All the papers included will give you plenty to think about.

The papers are grouped and ordered in the same way they were presented at the conference.

Starting with AI use in cataloguing James T. Clark and Getaneh Alemu's paper looks into the possibilities of improving discoverability of material by using the time period of the subject. They assess the limitations of time periods within LCSH and the benefits of FAST utilising exact dates, before moving to discuss ways of automatically populating the 648 field which had only been used in a small proportion of their records. They experimented with generative AI, using Google's Gemini tool, but also tried some Python scripts for comparison, emphasising in their conclusion that a range of tools should be embraced.

Hannes Lowagie showcases the idea of utilising machine-readable cataloguing rules that provide documentation both for staff and for AI workflows, using JSON as the core structure behind a 'triangle model'. This has immediate benefit on a local level but also has the potential to lead to more interoperability between institutions – with a sharing of both metadata, rules, and the logic behind it all. This strategy has been implemented at the Royal Library of Belgium (KBR).

Adam Swann and Sally Bell provide a thought-provoking look at a project designed to help copy catalogue Russian books at the University of Glasgow Library. They explored the use of AI tools to OCR text and transliterate it in preparation for searches to be performed on World Cat to find records to use. Whilst they had a high success rate, they also began to question the legal, ethical and environmental risks of using AI and thus decided to repeat the project using Python. This was also successful and so they summarise the benefits of Python versus AI, whilst reminding us that human intervention is still always necessary.

Moving into metadata team management, Helen K. R. Williams shares the idea of a 'team purpose tree' that maps institutional strategy to a mission and vision for the metadata team, helping to put into context the work team members do. The presentation at the conference was part workshop and attendees discussed in small groups what they thought their Mission, Vision, and Strategy were, or should be. Williams provided examples from the LSE Library's Metadata team. This is an interesting and practical tool that many of us could benefit from applying in our own workplaces.

Helen K. R. Williams' second contribution rounded up the first day's Collaborative metadata-ing block, looking at the work the LSE Metadata team have done on improving institutional research visibility by improving the discoverability of researchers, their outputs, and their association with LSE departments. This work was achieved through engagement with Wikimedia platforms, something their team has been doing in relation to the library's collections since 2020 but has now been extended to enhancing research visibility.

Day two continued the collaboration theme with Blaithin Hurley providing an interesting look at four historic diocesan libraries in Ireland and explores how approaches to cataloguing their collections have impacted their discoverability and accessibility for both local and scholarly communities. She discusses the idea of metadata as a form of cultural stewardship by recording provenance, preserving context, and making items visible.

Chuin Peng Sim and Gandhimathy Durairaj provide a fascinating insight into the National University of Singapore Libraries' digitisation and digital stewardship programme. As a leading Southeast Asia resource hub there has been an extended programme on digitising a wide range of materials in different languages and formats, such as private papers, maps, photographs etc. They emphasise the importance of metadata for enabling contextual understanding as well as discoverability. Internal and external partnerships have been key in expanding expertise as well as broadening access to private resources. Digital Gems is the public platform through which all the material can be accessed.

On the topic of linked data Sarah Glaccum describes the project undertaken at the University of Southampton Library that experimented with using OCLC's Meridian to add linked data entities to two unique collections. Having been granted free access for one month to Meridian careful preparation was involved and as much data as possible was gathered to populate a spreadsheet before the trial began. Glaccum postulates that engaging with linked data entities made better use of her team's time than training to be NACO contributors, which is an interesting consideration.

Amy Staniforth ran a practical workshop at the MDG conference where a group of attendees catalogued resources for the Drawing Room Library (Bermondsey, London) using a variety of devices and working on a MARC template, without access to the usual Library Management Systems and tools. The workshop was based on work Staniforth had been doing as a freelance cataloguer for the Drawing Room Library on a basic LMS without access to Z39.50. The workshop aimed to test collaborative cataloguing, and discuss thoughts on unlicensed freely shared metadata.

Fran Frenzel describes a pilot project run at the LSE Library to turn metadata into computationally useful datasets. The project took the metadata from their print PhD theses collection and explored the processes needed. The article notes decisions made and lessons learned, and looks to developing skills across the team, demonstrating the relevance that metadata experience has beyond the traditional confines of a library catalogue.

Finally, we are delighted to highlight that from this issue onwards we are also including the minutes and reports from the CILIP Metadata & Discovery Group committee meetings. It was felt that the group should become more transparent to its

members and help keep them informed on current discussions, and reports from associated groups. Many years ago, when a print-based publication, C&I used to included relevant reports to help disseminate information to its members and it was considered that this would also be a benefit to members and readers now. This issue contains the minutes from the March 2026 meeting.

An a friendly reminder that the editors are always happy to receive article proposals, case-studies, feedback, and letters. C&I is a practitioner-focused journal, from the community for the community, and we welcome small, informal pieces as well as longer articles. Please contact us at cataloguеandindex@gmail.com.

Karen F. Pierce & Fran Frenzel, September 2026

Metadata & Discovery Group

More time needed

automation, MARC 648, and discovery by time period of subject

James T. Clark 0000-0001-9936-4661
Library Systems and Discovery Manager, Southampton Solent University

Getaneh Alemu 0000-0003-2424-1725
Cataloguing and Metadata Librarian, Southampton Solent University

Received: 12 May 2026 | Published: 16 September 2026

ABSTRACT

This paper describes an investigation into how existing bibliographic metadata can be used to improve discovery by time period of subject. We describe the benefits of using an exact-year date range for the subject matter of the resource, as recommended by FAST (its Chronological heading) and recorded in the MARC 648 field. We assess the feasibility of automatically generating this date range using either AI or conventional scripting techniques. Finally, we describe ways this metadata can be displayed as a facet in discovery systems.

KEYWORDS metadata automation; retrospective cataloguing; generative AI; faceted discovery; time period; MARC 648

CONTACT James T. Clark james.clark@solent.ac.uk Southampton Solent University

1. Facets and time period of subject

To establish the rationale for our investigation, we begin with a description of the current theory and practice in relation to facets in general and specifically faceting by time period of subject (TPS).

1.1. The faceted approach to information organisation

Broughton (2006) notes the difficulty of defining "facet" concisely, but a facet can be understood as a property or characteristic of an information resource (e.g., topic, place, time, form) that enables users to filter, refine, and navigate search results. The origin of the faceted approach in libraries is often identified as Ranganathan's facet analysis in the 1930s, later formalised in his Colon Classification system, which introduced five fundamental categories: Personality, Matter, Energy, Space, and Time (PMEST) (Ranganathan, 1960, 1.25).

Ranganathan's principles were advanced by the UK Classification Research Group (CRG) in the 1950s, influencing systems such as Bliss Bibliographic Classification (BC2)

(Broughton, 2023, p.414). In 1955, the CRG published a manifesto to make faceted classification the "basis for all methods of information retrieval" (Classification Research Group, 1955). And indeed, the faceted approach to classification has also contributed to the development of subject headings and faceted thesauri (Broughton, 2023, p. 411).

1.2. Facets in library discovery systems

Although the strict faceted approaches to information organisation described above have not been widely adopted, the principle of organising subjects into discrete facets has become foundational for modern information retrieval, common in both library discovery systems and commercial platforms (Broughton, 2006, pp. 61–65). Facets' presence in commercial systems presumably helped make them comprehensible to users when they started to appear in 'next generation library catalogues' in the 2000s. Several studies from this period describe facets' popularity with users (Olson, 2007, p. 555; Sadeh, 2008, p. 22; Denton and Coysh, 2011, p. 316). While facets can be a valuable tool in discovery, their value is sometimes severely limited by practical challenges in their implementation that include "user interface design, metadata quality and complexity, the diversity of library resources, and the theoretical and practical impossibility of completely populating consistent, accurate facets" (McGrath, 2023, p. 440).

1.3. Time Period of Subject (TPS) in the library models

In faceted analysis and classification, including Ranganathan's PMEST, Time is one of the fundamental facets and denotes the temporal extent inherent in subject matter rather than a simple record of bibliographic dates. This demonstrates that time period is an important and analytically significant component of subject representation (Dean, 2004, p. 340). The treatment of temporal information in major bibliographic models and metadata standards reinforces this view, consistently recognising time period as a canonical facet.

The IFLA Library Reference Model (LRM) is a conceptual framework that consolidates three earlier models (FRBR, FRAD, and FRSA) to support user tasks such as finding, identifying, selecting, obtaining, and exploring resources. The LRM defines a set of 11 core entities that structure bibliographic and authority data. One of these entities is Timespan, "a temporal extent having a beginning, an end and a duration" (Riva, Le Boeuf, & Žumer, 2017, p. 36). While this entity can apply to many aspects of bibliographic metadata (e.g., date of publication, lifespan of a person) as well as the TPS, its inclusion as a separate entity in LRM emphasises its distinctness as a concept and thus its suitability for use as a facet.

Mirroring LRM, RDA uses the Timespan element to record the beginning, end and duration of temporal extents, including the chronological scope of subject content, while BIBFRAME represents temporal specificity through its *temporalCoverage*

property, enabling machine actionable encoding of a work's chronological domain, and Dublin Core provides an equivalent mechanism through its dc:date element, which also supports structured temporal specification. This demonstrates that time period is analytically central to bibliographic metadata and must be recognised as a fundamental entity and facet, not only in theoretical models but also in practical library metadata that supports indexing, filtering and browsing. This position is central to the argument developed in this paper.

1.4. Time Period of Subject (TPS) in discovery

That the TPS of a resource might be of interest to users is fundamental both to the theoretical models for subject analysis and resource description discussed above and also to practical approaches: DDC, LCC, and LCSH all allow for the analysis of material by TPS. Librarians seem to agree: a survey by Hall (2016, p. 113) found that only one out of 40 library staff considered a TPS facet to be less than "moderately useful". On this metric, TPS performed better than other potential facets, including "Person (Subject)", "Geographic Location (Subject)", and "LC Call Number". The importance of TPS for discovery seems a reasonable assumption, but we note that it has not been demonstrated empirically.

The idea of TPS as a facet is not entirely novel. Hall (2011) found that 25.6% of the US university libraries she surveyed had such a facet, while Nahotko (2022, p. 91) found that 21.6% of Polish academic libraries had one. We found these figures surprisingly high, and they are not reflected in the current UK context. In brief searches of the library catalogues of all 24 Russell Group universities, we were unable to find one that had a dedicated TPS facet, though several displayed time periods within a general subject facet.

Some libraries have reported a negative experience with the TPS facet, which might account for its relatively low use. Denton and Coysh (2011, p. 317) report that the TPS facet ("Era") was both sparsely populated and confused by users with the date of publication. Snyder (2010, p. 69), whose research focused on the discovery of music resources, reports that users found the "Time Period" facet to be problematic, as "their labels did not clearly indicate that they contained subject information and … their content did not seem logical or intuitive, nor was it consistently applicable to music materials." These case studies illustrate some of the challenges with faceting, but they are not fatal to the *principle* of faceting by TPS, as its success often depends on the details of implementation (see 1.2. above). As discovery systems have grown more sophisticated, it seemed a good time to revisit this potentially under-used facet.

1.5. Time Period of Subject (TPS) in LCSH

LCSH is a sophisticated subject vocabulary that has facilitated access to resources across the anglophone world for over 100 years. However, its structure was developed for use in card catalogues, and it is not optimised for modern discovery systems,

especially as regards faceting. McGrath (2007) notes several problems that LCSH would cause for a TPS facet; we summarise these below with her examples:

1. Time period is omitted from some subjects where it is considered redundant, e.g., "Underground Railroad$xHistory."
2. Time period is sometimes mentioned or implied in words, e.g., "Iron Age", "Post-Communism", "Literature, Ancient".
3. Time period can often only be entered in pre-selected ranges, not reflecting the actual range covered by the resource, e.g., a book on 1868 might have the subdivision "19th Century."

The difficulties this causes for faceting by TPS are obvious. Resources on a historical topic might have no recorded chronological information to extract into a facet (problem 1); the information might be there but vague and difficult to extract (problem 2); information that is easy to extract might be grossly imprecise (problem 3).

1.6. Time Period of Subject (TPS) in FAST

OCLC developed FAST (Faceted Application of Subject Terminology) by simplifying and reworking LCSH into at first eight, later nine, heading groups, one of which is the Chronological heading (Danskin et al., 2023, pp. 507–508). There are no prescribed time periods in FAST; rather, the Chronological heading is populated with the exact date range that the resource covers. Chan and O'Neill, in an authoritative FAST textbook, state:

> "Chronological headings represent the time period treated or covered in the document or resource at hand. Chronological headings are assigned to reflect the actual time coverage in the resources and are not necessarily limited to specific periods associated with particular events." (Chan and O'Neill, 2010, p. 99)

Thus, in theory at least, FAST avoids the problems that LCSH has with regard to faceting by TPS. The Chronological heading should be present whenever it is relevant, and it should give an accurate reflection of the period that the resource treats.

1.7. LCSH and FAST comparison

The increased precision and consistency offered by using the exact dates makes the FAST approach superior for modern discovery methods, including faceting.[1] The use of exact dates also has the further benefit of simplicity: there are no overheads for the maintenance of a vocabulary, few requirements for training in its application, and it is

[1] See section 3, below, for a discussion of faceting using exact dates. Although future AI-based search may apply filters dynamically, potentially making traditional visual facets and faceted navigation redundant (see Campbell, Jeffery, and Nowicki, 2025), the precision provided by exact date ranges will likely remain relevant.

internationally applicable.[2] Exact dates also seem more suited for applications beyond the purely bibliographic. Increasingly, discovery is happening outside the library, and books are discovered alongside a range of other resource types. In this environment, an approach that can be easily understood out of context and easily applied by non-specialists has distinct advantages.

Similar arguments to those made above have been adduced for the adoption of FAST generally: its simplicity makes it easier to learn and apply, and it is better suited to the modern discovery environment of faceted navigation and linked data (Danskin et al., 2023, p. 507). Our intention here is not simply to restate these arguments, but rather to argue that in the recording of *chronological* data specifically the advantages of FAST over LCSH are greatest. Though simplified from LCSH, most FAST headings are still part of a controlled vocabulary requiring training to apply and continual maintenance; not so the Chronological heading. And it is only in Chronological headings that FAST offers greater precision than LCSH. Furthermore, it is impossible to extract accurate FAST Chronological headings automatically from LCSH; this is not the case for other FAST headings.

Thus, the approach we have adopted in our own catalogue, and that we would advocate for, is to continue to use LCSH alongside a FAST-style exact date range in the 648 when the work has a historical focus.

1.8. FAST Chronological heading: current practice

While in theory the FAST Chronological heading is ideal metadata for a TPS facet, there are problems with the state of current practice. The first is that the headings are not sufficiently widely applied. A search of the UK's National Bibliographic Knowledgebase (NBK) for the 648 MARC field, in which the heading is recorded, found it in just 0.7% of records. By contrast, the relevant subfields for LCSH chronological subdivisions (650$y and 651$y) are present in 8.3% of records. Furthermore, it is likely that many 648 fields do not contain FAST data. From a sample of 11,879 MARC 648 fields only 4,375 (36.8%) were coded as FAST. Thus, the proportion of FAST Chronological headings in the NBK might be as low as 0.25%.

The second problem is that very often the value in the FAST Chronological heading does not in fact attempt to record exact dates for the TPS but instead has been automatically generated from the dates in the LCSH. Thus, a book exclusively covering the 1970s is given the Chronological heading "Since 1960" in FAST, this being a combination of the earliest and latest dates found in its LCSH ("$y1960-1980" and "$y1969-"). Of the 4,375 FAST Chronological headings in our NBK sample, 470 (10.7%) end "-2099" and 227 (5.2%) are open-ended ("Since ..."). These headings, at least,

[2] In this context, it is worth noting that the *Regeln für die Schlagwortkatalogisierung* (RSWK), the German subject headings system, also tends to assign exact dates in the 648 field (Frommeyer, 2004, p. 202); the National Library of Poland Descriptors (NLD) uses both exact dates (in the 045 field) and preassigned ranges that apply across the whole vocabulary, in the 648 field (Cieloch-Niewiadomska, 2019, p. 52).

appear to have been automatically converted from LCSH, and it is likely that many others have too. This is in line with guidance: the FAST Converter tool provided by OCLC generates such headings[3], and Chan and O'Neill (2010, p. 176) recommend the use of these "when more specific dates are not available." This approach, which allows automatic conversion from LCSH to FAST, is presumably founded on the belief that this data is better than nothing. We agree, but it is unfortunate that the provenance of this data, that it was automatically generated from LCSH, was not recorded.

2. Automatically populating the 648 field

We have seen that an exact date TPS currently exists in only a tiny proportion of records, and manual retrospective addition of the data at that scale is clearly out of the question. Therefore, we decided to investigate the feasibility of fully automating the generation of this metadata, i.e., finding an automation method that is sufficiently accurate (for a subset of records, at least) to be applied without human oversight at the record-by-record level.

2.1. Trying generative AI: methodology

Given the current excitement (or perhaps hype) about the potential of generative AI (GenAI) to 'understand' natural language, we decided to investigate whether GenAI tools could accurately create an exact TPS based on existing metadata. We chose Google's Gemini as our tool, as its API allowed generous usage limits without charge.

2.1.1. Generate a set of sample records

After early experiments with GenAI, we decided to provide three MARC fields to the tool: the title (245, minus $c), contents (505), and summary (520). We chose not to provide subject headings as the date ranges there can be imprecise. In selecting a sample of records to work with, we merely wanted records that were likely to have historical subjects and that had some indication of a date in these fields. We located these by a crude methodology, first searching our catalogue for the keyword "history", then filtering for 'date terms', i.e., four consecutive numbers (i.e., a year) or the string "centur" (i.e., the word "century" or "centuries"). This approach returns some false positives, but this was not a concern. We also needed records to have a single publication date (extracted from the MARC 008) to use as the upper bound.

As a final criterion, we excluded two types of material that early experiments suggested the tool would perform badly on: multi-volume works and exhibition catalogues. The problem with multi-volume works is that the dates mentioned in our selected metadata fields often refer to the period covered by the whole work, while we might only own a single volume. Often the only date that records for exhibition catalogues contain is the date of the exhibition, and arguably this is not relevant to the TPS.

[3] Available online at https://fast.oclc.org/lcsh2fast/.

Our initial search for the keyword "history" retrieved 10,955 records. The further filters reduced the sample to 4,491 records. We treated these as three separate sets: those with a 'date term' in the 245 only (516 records), those with a 'date term' in the 245 and at least one of the 505 and 520 (974 records), and those with no 'date term' in the 245 but in at least one of the 505 and 520 (3,001 records).

2.1.2. Push through Gemini with prompt

These records were pushed through the Gemini API with the following prompt, refined from early experiments:

> "Below is the title, table of contents, and summary of a student essay. Give me a date range in years for the subject-matter of the essay and a percentage value for how confident you are that this date range is correct. For example, if there was nothing in the data to indicate the date range, you could return a confidence level of 0%. Give me the date range only first, then confidence percentage only, then any further information. Separate these three sections with the tab character."

The first thing to address is our referring to the resource as a "student essay". We were concerned that if we described the resource as a book, the tool would search for information about it, find bibliographic data, and return the dates from its LCSH. We did not run a controlled experiment to find if our approach was necessary or had any impact, but anecdotally we noticed no evidence of LCSH dates being returned by the tool.

The prompt displays some features of a good prompt: it gives some context and, while the text could be sharper and more concise, it has no ambiguity and specifies the output format. However, we did not use advanced techniques of prompt engineering such as scenario framing ("imagine you are a library cataloguer") or temperature setting. The great number of different patterns to be found in our input data made us sceptical of the value of providing a worked example ("shot") or chain-of-thought reasoning, and our early experiments did not change our mind.[4] But no doubt the prompt could be improved in some respects.

2.1.3. Process and filter the results

We anticipated that the tool would not be able to produce accurate data in every case, and we were keen to find ways to automatically identify data that was more likely to be accurate. One way was to ask the tool to rate its confidence in the result. We also decided to run the queries through the tool twice and see whether the same result was obtained. We left approximately a month between the two query runs.

[4] See Khan, 2024, pp. 35–40 for the characteristics of a good basic prompt and (p. 50) on temperature setting. See Debnath et al., 2025, pp. 6–8 for the concept of "shots" and chain-of-thought prompting.

Once we had results from the two runs, several short Python scripts were used to process the results. First, the few occasions where the tool had failed to return a valid result (either a single year or a year range) were excluded. The second process checked the upper bound (if one existed) from the returned date range against the publication date, choosing the earlier of these two dates as the new upper bound. Finally, we checked the results from the two different runs against each other; only the results where the two runs matched were passed to the validation stage.

2.1.4. Validate for acceptability

Random samples were taken from the result sets for validation. Our validation method was to check the date ranges for acceptability, that is, to check that they did not misrepresent the resource. This differs from the gold-standard approach to validating GenAI tools, which would have meant assigning dates to the resources independently of the tool and then checking for matches. The gold-standard approach is a stricter test, but we do not think it is a more appropriate one here.[5] What we are interested in, after all, is how often the tool can apply useful metadata, not how often it can imitate a human. Furthermore, checking for acceptability acknowledges that there is a subjective element in assigning a date range to the subject, and it is the approach we would take if we were assessing the work of a new human cataloguer.[6]

The gold-standard approach to validation would have been superior in one respect: it removes the potential for validators to skew the results through some (possibly unconscious) desire for the tool to appear either effective or ineffective. We tried hard to be objective in our judgements of acceptability, but it is impossible for us to say whether we were successful; the best test of this is whether our results are replicable.

Unless there was a strong reason not to, we considered self-defined date ranges – those stated explicitly in the title or summary – as acceptable. And when the metadata we had for the resource was very full, we did not always consult the full text. The logic here was that even if we had the book in hand, we would form our judgements initially on the title, the contents, and the summary – the very metadata that we had – and would not necessarily scan the main text if we found those elements informative enough. In one respect we took a consciously lenient approach: if the upper bound was the publication date of the resource, we deemed this acceptable even if the upper bound should have been a year earlier. This is because it would probably have been more sensible to apply the publication year minus one as an upper bound, given the usual gap between a manuscript being closed and the book being published.

[5] Others investigating the use of AI for subject analysis also deviate from the gold-standard approach: Chow, Kao, & Li (2024, p. 582) use an acceptability validation, while Dobreski and Hastings (2024, p. 5) check against a gold standard but also pass acceptable headings.

[6] Deciding on an acceptable date range is not a simple task. Consider a book of 300 pages that mainly covers the years 1830-1850, but whose ten-page introduction reviews the years 1780-1830. Should we accept the date range 1780-1850 for this book? Probably not, but what if the introduction is 60 pages and thus 20% of the resource? And then what if the years 1780-1820 are dealt with in the first 40 pages of those 60?

2.2. Trying generative AI: results

Table 1 gives the results of our investigation. The three rows represent the three input sets obtained from the process described above (2.1.1.). The first column gives the number of records in the set. The second column gives the number of records that progressed to validation, that is, for which we got a valid and matching result from both runs through the AI tool. So, in the case of those records with a date term in the 245 only, a valid matching result was produced for 449 of the 516 records (87%).

The next column ("Sample 1 acceptability") shows the acceptability rate for the first sample we took, which for this group was 97%. We noticed that when the combined confidence level the tool produced for the result was 198 (i.e., 99% on both runs, or 98% on one and 100% on the other) or above, the results were more accurate, so we took a second sample that included only these results. In validating this set we noticed that when the result was a single date, rather than a date range, it was often faulty, so we excluded these too. The next column ("Sample 2 acceptability") shows the rate of acceptability here. The final column shows the number of results that meet the criteria for sample 2, i.e., that were available for sampling.

Group	Number of records	Two AI answer match	Sample 1 acceptability	Sample 2 acceptability	Meet criteria for Sample 2
Date in 245 only	516	449 (87%)	97/100 (97%)	197/199 (99%)	350 (68%)
Date in 245 and 505/520	974	821 (84%)	95/100 (95%)	187/195 (96%)	628 (64%)
Date in 505/520	3,001	1,654 (55%)	68/100 (68%)	86/100 (86%)	394 (13%)

Table 1: Acceptability rates of AI-generated date ranges

The groups with a date in the 245 (the top two rows in the table) performed very similarly. The tool's responses from both runs matched in around 85% of cases, and the acceptability rate for Sample 1 was 95% and above. We can increase the acceptability rate a little by applying the filters that produced our Sample 2, but there are diminishing returns here: the acceptability creeps up a few percentage points, but the proportion of records that now meet the criteria drops to around 65% (the final column in the table).

Records without a date in the 245 (the bottom row) perform less well. We can reach an acceptability rate of 86% for Sample 2, but then only 13% of records meet the criteria.

2.3. Trying generative AI: errors

Table 2 gives examples of errors that persisted into our Sample 2. Each one illustrates a different kind of error. For *An economic history of Europe*, the tool has been unable to recognise from "and the subsequent crisis" that the book treats the period beyond 2008. For *Pocketbook politics* it incorrectly applies the date 1970 as the end of the "Great Inflation of the 1970s." Even if it had done better here, we would still have considered the date range unacceptable, as the book mainly deals with 1900-1960, the period after 1960 being confined to a short epilogue. *Style 1930* treats only 1930 and the few years either side, not the whole 1930s. The real error here is one of over-confidence; with such limited information we would expect the confidence rating to be lower than the threshold required for our sample.

Title	Extract of 505/520	Gemini response
An economic history of Europe since 1700	... The book ends with an assessment of the impact of the global crash of 2008 and the subsequent crisis of the Eurozone ...	1700-2008
Pocketbook politics : economic citizenship in twentieth-century America	... ending with the Great Inflation of the 1970s.	1900-1970
Style 1930		1930-1939

Table 2: Representative errors in AI-generated date ranges

It is notable that we have seen instances where the tool dealt well with situations just like these, so there is a certain amount of apparent randomness to the responses. The tool performs much more inconsistently than a human would. The quality of input data is important, but even apparently good quality input data can be misleading, as in the Pocketbook politics example. In combination, these two factors mean that whether the tool gives an acceptable response involves a fair amount of chance. This might be disconcerting, but it does not necessarily obviate its use: the key is to find the right balance of filters that means the tool's failures occur at an acceptably low rate. A further important point is that in most cases when the response was deemed unacceptable, it was wrong by only a few years and, we suspect, closer to the truth than LCSH time periods would have been.

2.4. Trying regular expressions

The AI tool's results for those records that have a date in the 245 may seem impressive, but of course many of these records include an explicit declaration of the date range, such as *Women's suffrage in Britain 1867-1928*. The tool is not doing anything especially sophisticated in extracting this range, and it is possible to do the same using regular expressions. We used a Python script to identify a date range from 14 different regular expression patterns, such as "in the ... century" (*The British working*

class in the twentieth century), "of the [decade]" (*Living up to the ads : gender fictions of the 1920s*), and "[date] to now" (*Photography at MoMA : 1960 to now*).

This approach generated date ranges for 1,077 of the 1,490 records (72%), and manual checking revealed that in only two instances did this date range fail to reflect the title of the resource, due to unanticipated patterns in the data. There is scope for this approach to be improved still further by accounting for more complex patterns and perhaps incorporating natural language processing (e.g., to identify parts of speech).

2.5. Review

When records include a date in the title, automatically generating a TPS is relatively simple. Regular expressions seem to give the best results, with an acceptability of almost 100% for 72% of records, though GenAI is not far behind. Once the low hanging fruit that can be handled by regular expressions is removed, GenAI is less successful, but it can achieve an acceptability rate of 91% (97/107) for 26% of the records with a date in the title that our regular expression script passed over, and 86% (86/100) for 13% of records with no date in the title.

3. Discovery

3.1. Faceting with raw data: problem and solution

There is an obvious difficulty in presenting exact date ranges in a facet: the user will see many overlapping date ranges and need to select multiple options. The predefined time periods in LCSH alleviate this a little, but with these too there will be considerable overlap; works about the 1500s might, depending on their main subjects, have time periods defined as "16th century", "Early modern and Elizabethan, 1500-1600", "Early modern, 1500-1700", "Renaissance, 1450-1600", "Wars of the Huguenots, 1562-1598", etc.

The solution is similarly obvious: to allow the user to type in a date range (or use a slider to select it). Such tools are common on commercial websites, normally for selecting a price range, and already exist on some library discovery systems for filtering by publication date. These tools will need adapting to work with a date range (as opposed to a specific date or dates), but this should be well within the capabilities of library software providers. Such a tool could also be used for other chronological facets, such as the creation dates now recorded in the 046 MARC field (Mullin, 2024, pp. 138–139).

3.2. Faceting with processed data: an interim solution

But we do not have to wait for the perfect faceting tool before using this data. At Solent we developed an interim solution that takes the year range from the 648 field

and converts it to two groups of data, stored in local fields in the MARC record: one group contains the centuries covered by the year range, the other contains the decades covered by it. Thus a 648 of 1756-1779 would be given "18th century" in the centuries field and "1750s", "1760s", and "1770s" in the decades field. This conversion work is fully automated via normalisation rules and the Alma API. As in most instances LCSH time periods identify centuries correctly, the script also generates centuries (but not decades) for records that have LCSH time periods but no 648.

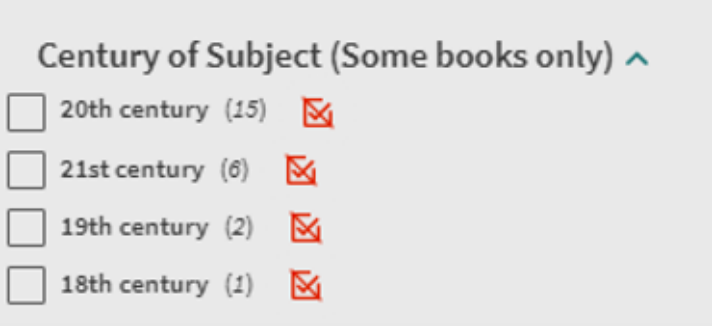


Figure 1: Century of subject facet in Solent's catalogue

Once created, the century and decade fields are then used as facets in the catalogue (see Figure 1). The level of precision used here is a matter of choice; with a small tweak the script could output periods of, say, 25 or 50 years, instead of decades and centuries. As discussed, this is not our ideal solution, and there is always some maintenance overhead to a customisation, even when fully automated like this one. But this shows what is possible with existing technology. These facets went live in Solent's catalogue in February 2026, and we will monitor their use. A prototype of the centuries facet was available for the previous nine months. In absolute terms it was not heavily used, but once allowance is made for the fact that it appears as a facet on only about 20% of searches (due to the small proportion of records where TPS is relevant), it was used about half as frequently as the general subject term facet. This is not a bad return, considering time period is only one aspect of a subject, and it justifies its continued inclusion in our catalogue.

4. Conclusions

4.1. Reflections on Time Period of Subject (TPS) data

Describing the TPS of a resource is a fundamental part of subject analysis. We are convinced that in the contemporary discovery context, the best way to do this is through assigning the exact date range of the subject matter.

The main obstacle to adopting this practice is the enormous number of records that are missing this data. To combat this, FAST allows the use of LCSH time periods as an alternative, allowing automated conversion from LCSH to FAST. Despite the problems with LCSH time periods, we approve of this approach on the basis that this data is better than nothing. But the publication date of the resource should be used as the

closing date of the time period, thus avoiding date ranges such as "2000-2099". And more importantly, the provenance of the data should be identified, e.g.,

> 648 #7 $a1960-2000$2fast$7(dpsfa)Derived from LCSH subdivisions and publication date.

Identifying the provenance of AI-generated data is widely acknowledged to be important, and it is similarly important for other automated methods that are prone to producing errors. Recording data provenance will make it easier to revisit the LCSH-derived dates later and improve upon them, either manually or by some automated process.

In this project we investigated the potential for automatically generating exact time periods using three key metadata fields: the title, contents, and summary. We were able to generate this data with a high degree of confidence for about a quarter of records. We note, however, that most successes came from the 'low-hanging fruit', where the date was mentioned in the title; when this information is absent, the success rate drops dramatically.

The project also investigated methods for the use of this data in modern discovery tools. While a 'date slider' facet appears best suited for use with this data and is presumably not beyond the capabilities of library systems providers, we were able to demonstrate that a workable alternative is available now.

4.2 General reflections on automated metadata generation

It is a commonplace to insist that GenAI in cataloguing must be used with human supervision.[7] We agree, but we contend that the kind of supervision used here, where quality is maintained by sampling, can be sufficient.[8] This project looked at the retrospective addition of the 648 field to older MARC records. In the vast majority of records, this field has not previously been added by human cataloguers, and there is no realistic prospect of the cataloguing community achieving this retrospective addition manually. In cases like this, then, the question is whether we would rather have no data at all or data that is potentially erroneous. That will depend on factors such as the potential harm done by erroneous metadata, the value added by the extra metadata, and the proportion of errors. In this particular case, where we can generate the 648 for a subset of records with an acceptability rate above 85%, we think the automated approach deserves further investigation. It might put this question in some sort of perspective to remember that the guidance for the FAST Chronological heading is that LCSH time periods should be used where an exact date range is not available (Chan and O'Neill, 2010, p. 176). The date ranges generated by this approach are almost always considerably less satisfactory than the 'unacceptable' ranges generated by our AI tool. If the imprecise ranges derived from LCSH are deemed better than

[7] For bibliography, see the review article by Engel et al., 2025, who conclude (p. 8), "there is general agreement that human oversight is a necessary component for the use of AI technologies today."

[8] The German National Library have used this approach for some time; see Mödden, 2022, p. 262.

nothing, then a smaller proportion of more minor errors from GenAI should not cause more concern.

Many investigations into cataloguing applications of GenAI look for ways to make what we already do more efficient. This is important work, of course, but we should also look for ways that GenAI can allow us to do more, to add new or rarely used metadata elements. That is what this project does, in seeking to generate retrospectively a field that exists in only a tiny proportion of records currently. There are historical instances of librarians responding to technological change by expanding their scope (Abbott, 1988, pp. 138–184 and pp. 138–120), and GenAI provides such an opportunity to cataloguers.

We have elsewhere stressed the value of the 505 and 520 fields in MARC records for helping users find and select resources (Alemu, 2022, pp. 183–14; Clark, 2024, p. 196). This project has shown that they are also useful in the automatic generation of further metadata. In the absence of full text (and possibly even when it is available), ensuring that good quality content and summary fields are available for as many records as possible should be foundational to any attempt to automate subject analysis.

Finally, while we are generally positive about the potential of GenAI both in this particular application and in general, we would temper that with an acknowledgement of the need for caution.[9] In particular, we point to the way that simple regular expressions returned better results for records that had a date in the title. If we intend to embrace technology, then we should embrace a suite of tools, and not rush headlong for GenAI every time.

References

Abbott, A. (1988) *The system of professions: An essay on the division of expert labor*. University of Chicago Press.

Alemu, G. (2022) *The future of enriched, linked, open and filtered metadata: Making sense of IFLA LRM, RDA, Linked Data and BIBFRAME*. Facet Publishing.

Broughton, V. (2006) 'The need for a faceted classification as the basis of all methods of information retrieval', *Aslib Proceedings*, 58(1–2), pp. 49–72. Available at: https://doi.org/10.1108/00012530610648671 [Accessed: 10 August 2026]

Broughton, V. (2023). 'Facet analysis: The evolution of an idea', *Cataloging and Classification Quarterly*, 61(5–6), pp. 411–438. Available at: https://doi.org/10.1080/01639374.2023.2196291 [Accessed: 10 August 2026]

Campbell, L., Jeffery, K., and Nowicki, R. (2025) 'AI, acronyms, and the future of facets: Mnemonics as heuristic instructional tools for structuring GenAI prompts', *Qualitative and Quantitative Methods in Libraries*, 14(1), pp. 121–134. Available at: https://www.qqml-journal.net/index.php/qqml/article/view/906 [Accessed: 10 August 2026]

[9] For a balanced assessment of the use of GenAI in cataloguing, see Moulaison-Sandy and Coble, 2024.

Chan, L. M., and O'Neill, E. T. (2010) *FAST: Faceted application of subject terminology: Principles and applications*. Libraries Unlimited.

Chow, E. H. C., Kao, T. J., and Li, X. (2024) 'An experiment with the use of ChatGPT for LCSH subject assignment on electronic theses and dissertations', *Cataloging and Classification Quarterly*, 62(5), pp. 574–588. Available at: https://doi.org/10.1080/01639374.2024.2394516 [Accessed: 10 August 2026]

Cieloch-Niewiadomska, J. (2019) 'Introducing the National Library of Poland descriptors to the Polish national bibliography', *Cataloging and Classification Quarterly*, 57(1), pp. 37–58. Available at: https://doi.org/10.1080/01639374.2019.1573774 [Accessed: 10 August 2026]

Clark, J. T. (2024) 'The format and display of the MARC 505 contents note: Some recent practical developments', *Cataloging and Classification Quarterly*, 62(2), pp. 188–208. Available at: https://doi.org/10.1080/01639374.2024.2350467 [Accessed: 10 August 2026]

Classification Research Group (1955) 'The need for a faceted classification as the basis for all methods of information retrieval', *Library Association Record*, 57(7), pp. 262–268.

Danskin, A., et al. (2023) 'FAST the inside track: Where we are, where do we want to be, and how do we get there?', *Cataloging and Classification Quarterly*, 61(5–6), pp. 506–524. Available at: https://doi.org/10.1080/01639374.2023.2215762 [Accessed: 10 August 2026]

Dean, R. J. (2004) 'FAST: Development of simplified headings for metadata', *Cataloging and Classification Quarterly*, 39(1–2), pp. 331–352. Available at: https://doi.org/10.1300/J104v39n01_03 [Accessed: 10 August 2026]

Debnath, T., et al. (2025) 'A comprehensive survey of prompt engineering techniques in large language models', *TechRxiv* [Preprint]. Available at: https://doi.org/10.36227/techrxiv.174140719.96375390/v2 [Accessed: 10 August 2026]

Denton, W., and Coysh, S. J. (2011) 'Usability testing of VuFind at an academic library', *Library Hi Tech*, 29(2), pp. 301–319. Available at: https://doi.org/10.1108/07378831111138189 [Accessed: 10 August 2026]

Dobreski, B., and Hastings, C. (2025) 'AI chatbots and subject cataloging: A performance test', *Library Resources & Technical Services*, 69(2). Available at: https://doi.org/10.5860/lrts.69n2.8440 [Accessed: 10 August 2026]

Engel, J. Y., et al. (2025) 'Artificial intelligence in library cataloging: A review of literature', *Journal of Library Metadata*, 25(4), pp. 261–276. Available at: https://doi.org/10.1080/19386389.2025.2526913 [Accessed: 10 August 2026]

Frommeyer, J. (2004) 'Chronological terms and period subdivisions in LCSH, RAMEAU, and RSWK', *Library Resources & Technical Services*, 48(3), pp. 199–212. Available at: https://doi.org/10.5860/lrts.48n3.199-212 [Accessed: 10 August 2026]

Hall, C. E. (2011) 'Facet-based library catalogs: A survey of the landscape', *Proceedings of the American Society for Information Science and Technology*, 48(1). Available at: https://doi.org/10.1002/meet.2011.14504801195 [Accessed: 10 August 2026]

Hall, C. E. (2016) *Facets in library catalogs: The beliefs, behaviors, policies and practices that guide implementation*. PhD thesis. Drexel University. Available at: https://doi.org/10.17918/etd-7078 [Accessed: 10 August 2026]

Khan, I. (2024) *The quick guide to prompt engineering*. Wiley.

McGrath, K. (2007) 'Facet-based search and navigation with LCSH: Problems and opportunities', *Code4Lib Journal*, 1. Available at: https://journal.code4lib.org/articles/23 [Accessed: 10 August 2026]

McGrath, K. (2023) 'Musings on faceted search, metadata, and library discovery interfaces', *Cataloging and Classification Quarterly*, 61(5–6), pp. 439–490. Available at: https://doi.org/10.1080/01639374.2023.2222120 [Accessed: 10 August 2026]

Mödden, E. (2022) 'Artificial intelligence, machine learning and bibliographic control: DDC short numbers – Towards machine-based classifying', *JLIS.it*, 13(1), pp. 256–264. Available at: https://doi.org/10.4403/jlis.it-12775 [Accessed: 10 August 2026]

Moulaison-Sandy, H., and Coble, Z. (2024) 'Leveraging AI in cataloging: What works, and why?', *Technical Services Quarterly*, 41(4), pp. 375–383. Available at: https://doi.org/10.1080/07317131.2024.2394912 [Accessed: 10 August 2026]

Mullin, C. A. (2024) *Many pathways for discovery: Describing music resources using faceted vocabularies*. Music Library Association & A-R Editions.

Nahotko, M. (2022) 'Knowledge organization affordances in a faceted online public access catalog (OPAC)', *Cataloging and Classification Quarterly*, 60(1), pp. 86–111. Available at: https://doi.org/10.1080/01639374.2021.2015734 [Accessed: 10 August 2026]

Olson, T. A. (2007) 'Utility of a faceted catalog for scholarly research', *Library Hi Tech*, 25(4), pp. 550–561. Available at: https://doi.org/10.1108/07378830710840509 [Accessed: 10 August 2026]

O'Neill, E. T., and Chan, L. M. (2003) 'FAST (Faceted application of subject terminology): A simplified LCSH-based vocabulary', *World Library and Information Congress: 69th IFLA General Conference and Council*, Berlin. 219–2 August. International Federation of Library Associations and Institutions. Available at: https://webdoc.sub.gwdg.de/ebook/aw/2003/ifla/vortraege/iv/ifla69/papers/010e-ONeill_Mai-Chan.pdf [Accessed: 10 August 2026]

Ranganathan, S. R. (1960) *Colon classification*. 6th edn. Asia Publishing House.

Riva, P., Le Boeuf, P., and Žumer, M. (2017) *IFLA Library Reference Model: A conceptual model for bibliographic information*. IFLA. Available at: https://www.ifla.org/files/assets/cataloguing/frbr-lrm/ifla-lrm-august-2017.pdf [Accessed: 10 August 2026]

Sadeh, T. (2008) 'User experience in the library: A case study', *New Library World*, 109(1–2), pp. 7–24. Available at: https://doi.org/10.1108/03074800810845976 [Accessed: 10 August 2026]

Snyder, T. (2010) 'Music materials in a faceted catalog: Interviews with faculty and graduate students', *Music Reference Services Quarterly*, 13(3), pp. 66–95. Available at: https://doi.org/10.1080/10588167.2010.528746 [Accessed: 10 August 2026]

Metadata & Discovery Group

Authentic and artificial cataloguing

Machine-readable cataloguing rules as a bridge to unified metadata practice

Hannes Lowagie 0000-0002-0671-3568

Head of Bibliographic Information Agency, Koninklijke Bibliotheek van België - Bibliothèque royale de Belgique (KBR)

Received: 17 April 2026 | Published: 16 September 2026

ABSTRACT

Cataloguers and metadata professionals are entering a period in which local cataloguing rules must be not only transparent and consistent, but also usable by machines. This article examines a metadata strategy developed at KBR, Royal Library of Belgium that transforms traditional cataloguing guidance into a modular, machine-readable ruleset structured in JSON. Rather than maintaining separate documentation for staff and systems, the JSON file functions as a single source of truth that supports two outputs: a dynamic HTML interface for human users and a machine-readable interface that can be integrated into generative AI workflows and other metadata processing environments.

The article explores how this "triangle model" – JSON as the core structure, HTML as the human interface, and prompt-ready data as the machine interface – offers a scalable approach to maintaining consistency, authority control, and transparency in cataloguing practice. It also considers the wider implications for interoperability across libraries and cultural heritage institutions. If local rules are encoded in comparable machine-readable forms, alignment between institutions may move beyond shared terminology towards shared cataloguing logic. The article concludes by reflecting on how metadata professionals can shape AI-enhanced cataloguing environments while preserving professional judgement, local context, and accountability.

KEYWORDS metadata standards; artificial intelligence; interoperability; JSON

CONTACT Hannes Lowagie Hannes.Lowagie@kbr.be Koninklijke Bibliotheek van België - Bibliothèque royale de Belgique

1. Introduction

Metadata work is changing rapidly as libraries and cultural heritage institutions begin to explore the possibilities of artificial intelligence and automation. Tasks that were once carried out entirely by human cataloguers are now increasingly supported by digital tools capable of extracting, analysing, and suggesting metadata at scale. This does not diminish the importance of professional expertise; rather, it changes the context in which that expertise is applied. Cataloguers and metadata specialists are now expected not only to understand descriptive standards and local practices, but

also to work confidently within AI-assisted environments where speed, consistency, and data quality are essential.

At the same time, many institutions still rely on local cataloguing documentation presented in traditional formats such as printed manuals, PDF files, Word documents, or static web pages. While these formats remain useful for human reference, they are no longer sufficient in environments where systems need to interpret and apply rules automatically. Static documents can be difficult to update, challenging to search efficiently, and often disconnected from the tools used in day-to-day metadata creation. They may also lead to inconsistent interpretation when different staff members apply rules in slightly different ways.

This creates a clear need for guidance that is not only understandable to people but also structured in a way that machines can process. Modern metadata workflows require local rules that are consistent, transparent, reusable, and machine actionable. If institutions want to benefit from automation and generative AI while maintaining control over quality and local policy, their cataloguing guidance must evolve accordingly.

The method described in this article was developed using the Microsoft Power Platform. Power Platform is a low-code suite of tools that enables organisations to build applications, automate workflows, analyse data, and create intelligent processes without the need for extensive traditional software development. It includes tools such as Power Apps, Power Automate, Power BI, and AI Builder, which together make it possible to design flexible internal solutions quickly and efficiently[1].

The starting point for this project was retrospective cataloguing. Large volumes of legacy material still required catalogue records, and manual processing alone would have taken considerable time and resources. To address this, a Document Processing Model was implemented to extract metadata automatically from digitised sources such as title pages, scanned cards, and other structured documents. This model could identify and capture useful information including titles, names, dates, publishers, and other descriptive elements.

However, extracting metadata was only the first step. Raw data produced by automated systems does not automatically conform to local cataloguing rules, authority practices, or institutional standards. A separate tool was therefore needed to transform, validate, and enrich the extracted metadata so that it aligned with local cataloguing practices and could be imported reliably into the catalogue. Without this additional layer, automatically generated metadata would risk inconsistency, reduced quality, and limited usefulness.

The project therefore developed beyond simple extraction into a broader metadata governance solution: one that bridges the gap between automated data capture and

[1] For more guidance see Lowagie H. (2025) *AI Powered Cataloguing*. Facet Publishing.

professional cataloguing standards. This became the foundation for a system in which local rules could guide both human cataloguers and machine processes, ensuring that efficiency gains did not come at the expense of accuracy or coherence.

2. The problem with traditional cataloguing documentation

For many years, local cataloguing guidance has been created primarily for human readers. Institutions have traditionally relied on printed manuals, internal policy documents, PDF handbooks, spreadsheets, Word files, and static intranet pages to explain how metadata should be created and maintained. These resources are often detailed and valuable, reflecting years of professional expertise and carefully developed local practice. They help cataloguers understand institutional preferences, interpret national and international standards, and apply rules consistently in day-to-day work. However, although these formats are readable and useful for staff, they are not designed for machines. A PDF or written manual may clearly explain a rule to an experienced cataloguer, but a computer system cannot easily interpret the same information in a reliable and structured way. Rules written in paragraphs of prose may contain exceptions, conditions, and nuanced decisions. Another significant challenge is the maintenance of multiple versions of local documentation. At KBR, we had for example a French version and a Dutch version that could differ. But we also accumulated guidance in several places: a master manual, separate departmental notes, training slides, older policy files, email instructions, and informal updates shared among staff. When rules change, every version ideally needs to be reviewed and amended. In practice, this can be slow and difficult to manage. Some documents may be updated immediately, while others remain outdated, creating uncertainty about which version is current. Staff may unknowingly rely on superseded instructions simply because they are easier to find or more familiar.

This fragmentation can lead to inconsistent interpretation by people. Even highly experienced cataloguers may apply a rule differently if wording is unclear or if several versions of guidance exist. New staff members may find local practice difficult to navigate, particularly when institutional decisions are scattered across numerous files. To address this, we moved away from separate PDF documents and created a single central HTML page that brings together all cataloguing guidance in one place, available in both French and Dutch, making access simpler, clearer, and easier to maintain.

Although moving from PDF documents to a central HTML page was an important improvement, HTML still has limitations. At its core, an HTML page is primarily a text-based format organised into headings, paragraphs, lists, and links. It is excellent for presenting information clearly to human readers, but it is not ideal as a structured rules engine for machines. The content may be easier to read and maintain than scattered PDF files, yet the logic behind cataloguing decisions often remains embedded in narrative text rather than expressed in a form that systems can readily interpret and apply. This became especially clear when I began experimenting with

prompts for generative AI tools. In order to guide the AI correctly, I repeatedly found myself copying and rewriting large sections of the same cataloguing guidance that already existed on the HTML page. Rules, exceptions, preferred wording, local decisions, and examples all had to be restated in prompts so that the system could understand what was required. Very quickly, it became obvious that this approach was inefficient and unsustainable. It created unnecessary duplication and increased the risk that the guidance used by staff and the instructions given to AI tools would gradually diverge. I wanted to unify both environments and avoid maintaining two parallel versions of the same cataloguing rules: one version written for human cataloguers and another rewritten for artificial intelligence. If a local rule changed, it should only need to be updated once. Both staff guidance and machine workflows should then reflect the same decision immediately and consistently. This need for a shared and maintainable foundation led to the development of what we call the triangle model.

3. The triangle model

The triangle model is built around the principle of a single authoritative source that can serve multiple audiences and systems at the same time. Rather than creating separate documentation for people and separate instructions for machines, the model uses structured data at the centre, from which different outputs can be generated according to need. At the heart of the model is a JSON file that acts as the single source of truth[2]. JSON is a lightweight structured data format that is both human-readable and machine-processable. It allows cataloguing rules to be stored in a clear and modular way, including rule names, descriptions, conditions, examples, language versions, exceptions, and workflow notes. Because the information is structured, it can be reused consistently across multiple platforms.

The first output of this central JSON source is the HTML interface used by staff. Instead of manually writing and updating static web pages, the HTML view is generated dynamically from the JSON content[3]. This creates a clear staff-facing interface where cataloguers can browse rules, search guidance, and consult examples in a familiar web environment. It also ensures that the information visible to staff always reflects the latest approved version of the rules.

The second output is prompt-ready structured data for AI systems and other automated tools. The same JSON content can be passed directly into prompt workflows, retrieval systems, or metadata processing applications. This means that AI tools no longer rely on loosely copied instructions or manually assembled prompts.

[2] The KBR's JSON file is available on github: https://github.com/kbrbe/cataloging-guidelines/blob/main/guidelines.json

[3] The human-readable interface is available here: https://kbrbe.github.io/cataloging-guidelines/cataloging-guidelines.html. The HTML and the scripts we use to fetch the information from the JSON and create the human-readable interface can be accessed via the site's source code or at https://github.com/kbrbe/cataloging-guidelines/blob/main/cataloging-guidelines.html.

Instead, they receive accurate, current, and institutionally approved guidance drawn from the same source used by staff.

The strength of the triangle model lies in its simplicity: one structured core, multiple reliable outputs. JSON provides the logic and governance layer, HTML provides the human interface, and AI-ready data provides the machine interface. This reduces duplication, improves consistency, and makes future development far easier.

4. Wider implications for libraries

The approach described here has significance far beyond a single institution. While it was developed to solve local cataloguing and metadata management challenges, the same principles could be applied across libraries, archives, museums, and other cultural heritage organisations. If adopted more widely, machine-readable cataloguing rules could help move the sector towards a new level of collaboration, consistency, and interoperability.

One of the most promising outcomes is the potential for cross-institutional harmonisation. Libraries have long worked together through shared standards, cooperative cataloguing, authority control, and record exchange. However, collaboration has often focused primarily on sharing the metadata itself rather than sharing the logic behind how that metadata was created. Institutions may use similar standards while interpreting them differently through local policy statements, workflows, and cataloguing decisions. As a result, records can appear compatible on the surface while still reflecting significant differences in practice. A machine-readable rules framework offers the possibility of sharing not only data, but also the rules, decisions, and structures used to create that data. This is an important shift. If institutions can exchange metadata together with structured descriptions of the cataloguing logic behind it, receiving systems gain a clearer understanding of how records were produced, what local choices were made, and how those choices relate to broader standards. In this sense, transparency becomes as valuable as the metadata itself. This contributes directly to the long-standing vision of Universal Bibliographic Control (renewed in 2025) (International Federation of Library Associations and Institutions, 2025). The principle behind Universal Bibliographic Control is that bibliographic data should be created once, according to recognised standards, and then shared for the benefit of all. In practice, this goal has always been challenged by local variation, language differences, policy interpretations, and technological limitations. Machine-readable cataloguing rules offer a new pathway forward by making local variation explicit, structured, and easier to align across institutions. It could also support deeper integration with RDA Toolkit and institutional policy statements. Many organisations already maintain local application profiles of Resource Description and Access. If these policies were expressed in machine-readable formats, they could be connected more directly within the RDA Toolkit, local systems, training environments, and AI-assisted tools, reducing duplication and making updates easier to manage.

This model also makes a shared used of Artificial Intelligence tools possible. Rather than every library needing to build entirely separate systems, it becomes possible to imagine a common AI-enabled cataloguing platform that can be refined locally through institution-specific rulesets. In other words, one shared system could operate with local customisation. A library in Belgium, for example, could apply bilingual authority practices, while another institution elsewhere could apply its own language, descriptive traditions, or subject conventions using the same technical framework. This creates economies of scale while preserving local identity and professional standards. To make this possible, institutions could collaborate on a unified, machine-readable structure for expressing cataloguing rules. Formats such as JSON, persistent identifiers such as Uniform Resource Identifier, and linked data models could be used to represent rules in a consistent and reusable way. If the underlying structure is shared, individual institutions can adapt the content without losing interoperability.

5. Conclusion

The motivation for redesigning local cataloguing rules arises from both practical and strategic needs. On a practical level, libraries and other cultural heritage institutions need workflows that reduce repetitive manual tasks, improve consistency in metadata creation, and make policy updates easier to manage. Metadata infrastructures must now be capable of supporting future developments in automation and AI-assisted cataloguing. Library systems are no longer isolated local environments; they operate within increasingly connected networks of discovery platforms, data exchange services, repositories, and external knowledge systems. In this context, local rules that exist only in static documents are difficult to integrate into modern workflows. This article proposes a solution by making the cataloguing rules structured and ready to use in human and automated environments. The objective is to move towards a more sustainable and future-ready approach to cataloguing documentation. They move local policy from static reference files into active operational frameworks that can support both human expertise and technological innovation.

There is a strong case for collaborative development of interoperable, machine-readable rules across the sector. No single institution should need to solve these challenges in isolation. Shared models, common structures, and reusable approaches would benefit libraries of all sizes and create stronger foundations for collective progress. Now is the right time to act. Artificial intelligence is developing quickly, system expectations are changing, and institutions are already reconsidering how metadata work will be organised in the coming years. If libraries take the initiative now, they can shape these developments according to professional values of quality, transparency, access, and cooperation rather than adapting later to systems designed without them in mind.

References

International Federation of Library Associations and Institutions (2025) *IFLA Professional Statement on Universal Bibliographic Control*. Available at: https://repository.ifla.org/handle/20.500.14598/4259 [Accessed: 10 August 2026]

Lowagie, H. (2025) *AI Powered Cataloguing*. Facet Publishing.

Lowagie, H. and Van Woensel, J. (2025) *Royal Library of Belgium (KBR) - Bilingual Cataloging Guidelines* (version 0.1.0). Available at: https://github.com/kbrbe/cataloging-guidelines and https://doi.org/10.5281/zenodo.15018951 [Accessed: 18 August 2026]

Metadata & Discovery Group

Can AI or Python help us catalogue books we can't read?

Adam Swann 0000-0003-1041-1308

Metadata Migration Assistant Librarian, Digital Library Team, University of Glasgow Library

Sally Bell 0000-0001-8647-305X

Head of Collection Development, Collection Development Team, University of Glasgow Library

Received: 14 August 2026 | Published: 16 September 2026

ABSTRACT

The potential for using tools like AI and Python to increase the efficiency of library cataloguing is being increasingly recognised. The University of Glasgow Library has a substantial backlog of uncatalogued non-English material, but few staff with the specialist language skills required to tackle it. Recent advances in computerised image analysis and transliteration have made it possible to begin cataloguing this backlog without requiring specialist language skills.

This article discusses a project which started by investigating the viability of using AI tools to help us copy catalogue Russian books. The project found that ChatGPT and Copilot can carry out OCR and transliteration of Russian text quickly and accurately. AI tools, however, come with serious environmental, legal, and ethical drawbacks.

We therefore started developing our own tool written in Python which has similar OCR and transliteration functionality to AI tools but without the associated environmental or ethical costs. The initial results from our Python tool are promising but more development is required.

Our project concluded that both AI and Python can be successfully used to help cataloguers without specialist language skills transliterate non-English texts. To maintain metadata quality, however, cataloguers with language skills must remain an integral part of our workflows.

KEYWORDS transliteration; non-English cataloguing; AI; Python, OCR

CONTACT Adam Swann adam.swann@glasgow.ac.uk University of Glasgow Library

The Centre for Russian, Central and East European Studies is a long-established and internationally renowned research partnership based at the University of Glasgow. To support the work of this unit the University of Glasgow Library (UoGL) has a substantial volume of stock in Russian, but no dedicated cataloguer for this material. Consequently, a backlog of uncatalogued Russian material has built up which is divided into two equal parts.

The first is approximately 1000 Russian books which require retrospective cataloguing, as they are on the shelves and recorded in a card catalogue but not yet added to our electronic catalogue. The second part is a donation of approximately 1000 Russian books which have not been appraised or catalogued. Some records from the card catalogue have been added to our electronic catalogue, but the Russian-language cataloguer working on this has subsequently moved to another team in the library.

This article discusses a pilot project which explored the viability of using AI and Python to help us address our Russian cataloguing backlog. It begins by situating the project in the broader context of AI use in the sector, then outlines the workflow and discusses the results of our project and the viability of using AI to help us catalogue non-English books. The project's initial scope was limited to using AI and so was originally intended to conclude there. However, the project was conducted against the backdrop of a broader debate about the tensions between what AI promises and what it delivers, and at what cost.

By the end of the AI phase of the project we, like many librarians, had become increasingly uncomfortable about AI use. We therefore started a second phase of the project which recreated the same workflow but using the Python programming language instead of AI. This article documents not just our project, but our evolving attitudes to AI use which are a microcosm of the shifting debates in our sector.

Recent discussions of AI use in academic libraries describe it as not only "rapidly changing the library landscape" (Togia et al., 2026, p. 1), but even "the next frontier in library cataloguing" (Ogungbenro et al., 2025, p. 105). It has particular potential for addressing cataloguing backlogs, since it offers "a level of productivity and efficiency in metadata creation that is unthinkable for humans" (Dover and Grzegorski, 2025, p. 605).

However, this must be balanced against the reality that for all its apparent sophistication, AI is currently a somewhat blunt instrument for metadata work. It has phenomenal data-processing capability but lacks the nuanced understanding of context necessary for accurate cataloguing (Oyighan et al., 2024, p. 24); simply put, AI alone "cannot consistently create quality metadata" (Sussmeier and Henry, 2025, p. 245).

We find a similar pattern in recent experiments which used AI to catalogue non-English books. When asked to OCR images of book title pages to create RDA MARC records, ChatGPT did so effectively for major languages like Arabic, French, and Spanish, but struggled with minority languages like Tamil and Sinhala (Gamage and Wanigasooriya, 2024, p. 241, 227). Similarly, it successfully converted Chinese book titles to Latin characters but was less accurate with authors "due to the intricacies of Chinese names" (Jiang et al., 2024, p. 48).

Our pilot project's initial aim was to investigate whether AI tools can help staff without specialist language skills to copy catalogue non-English books. The scope of the pilot was limited to Russian Cyrillic texts, but there is potential to include both Latin non-English languages and other non-Latin scripts in the future. We aimed to be able to use AI tools to identify the correct catalogue record to use for a non-English book with at least 80% accuracy.

When cataloguing books written in non-Latin scripts, it is standard practice to transliterate them into Latin script, e.g. 'Государственные финансы царской России в эпоху империализма' becomes 'Gosudarstvennye finansy tsarskoi Rossii v epokhu imperializma'. This allows readers of the secondary script "to comprehend the orthography and the approximate pronunciation of the original text" (Tour et al., 2025, p. 1) and enhances discoverability by providing both original script and Latinised catalogue entry points. When transliterating non-Latin scripts like Russian Cyrillic into Latin characters, we use the Library of Congress Romanisation standard[1].

We began by selecting 50 random Russian books from the card catalogue and photographing their title pages, ready for the AI tool to transliterate this Russian Cyrillic text into Library of Congress-compliant Latin characters. We would then use the transliterated author, title, and publication details to identify the correct record on Worldcat to download to copy catalogue the book.

We decided to use ChatGPT for the project because it supports both OCR from images and LC-compliant transliteration. Our initial workflow started by uploading a title page image to ChatGPT and asking for an LC-compliant transliteration of its bibliographic details. We would then format and collate these into a spreadsheet and search for the appropriate record on Worldcat.

However, the repeated formatting and collation required soon became laborious. Two important and intertwined principles for "obtaining precise and useful responses from AI models" are specifying the desired output format and reviewing the output to refine the input prompt (Geroimenko, 2025, p. 24, 31). We therefore provided ChatGPT with a bibliographic details template to use in its responses, and after some iteration and refinement we were satisfied with the format of the output.

We then streamlined the start of the process by asking ChatGPT to respond to an input image of a Cyrillic title page by immediately supplying the transliterated version in our preferred format, with an English translation for reference. This simplified workflow greatly increased the speed of the transliteration process.

After we had collated all 50 transliterations, a colleague fluent in Russian reviewed them alongside the original images and confirmed their accuracy. We then felt confident proceeding with the project and using the transliterations to search for records on Worldcat.

[1] See the Library of Congress's Russian Romanisation Table (Library of Congress, 2012a) and its source document (Library of Congress, 2012b).

We searched for all 50 books on Worldcat using the transliterated bibliographic details and the results were excellent:

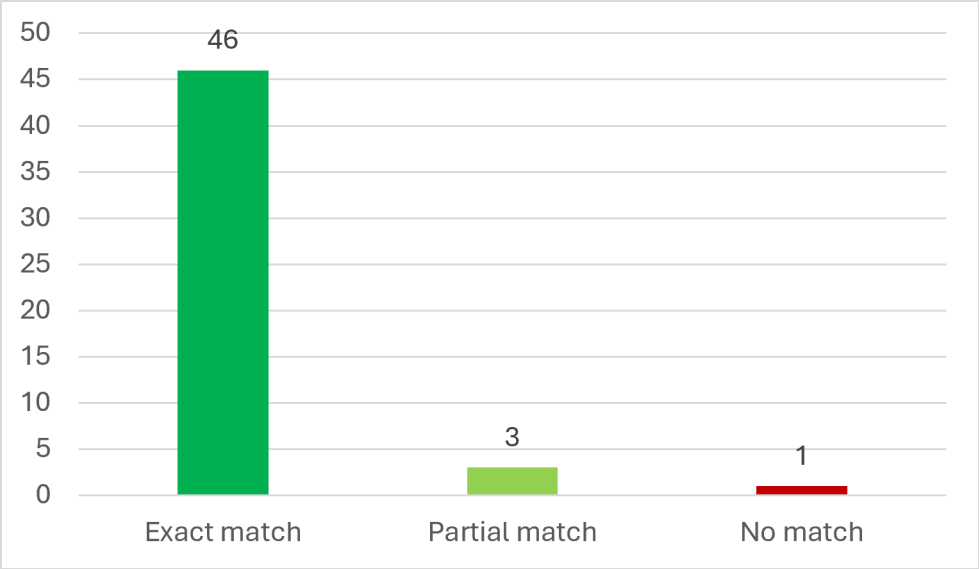


Figure 1: Retrospective cataloguing results

We were able to identify 46 exact matches on Worldcat (a 92% success rate), and only one book was not identified because it had no Worldcat record. The remaining three records were partial matches. They weren't exact matches as ChatGPT had transliterated them due to minor transliteration differences or errors:

Our ChatGPT Transliteration	Transliteration on Worldcat Record
ekonomike	ekonomiko
razvitogo	razvitoi
– [en-dash]	- [hyphen]

While these are minor differences a human cataloguer can still identify the correct catalogue record, so with some tweaking of search queries the final success rate was 98%.

We then applied the same workflow to 50 random books from the uncatalogued donation which had been identified by an academic as rare or unusual. As before, we uploaded photographs of their title pages into ChatGPT then used the transliterated details to search for records on Worldcat.

The results were not as successful as with the retrospective cataloguing:

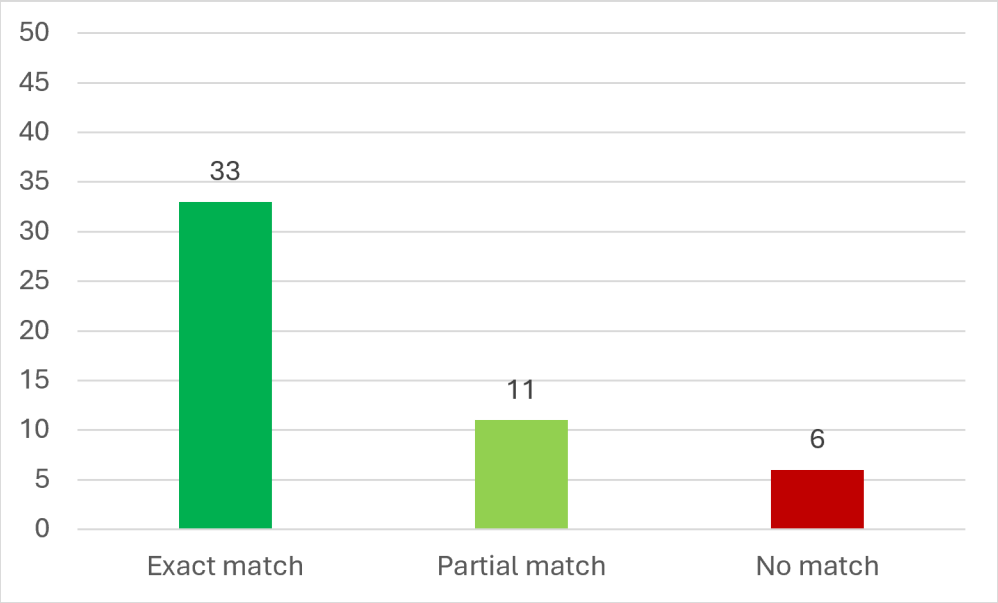


Figure 2: Donation cataloguing results

We had 33 exact matches (a 66% success rate), and 6 books had no Worldcat records. The remaining 11 books were partial matches; 5 titles had been catalogued on Worldcat using contractions (e.g. 'univers.' instead of 'universitov'), 3 titles had minor differences or errors in transliteration, and 3 titles were different editions to our own copies. The lower match rates are possibly attributable to these being rarer books, meaning there will be fewer holdings elsewhere for us to match against.

As before, with human input we could identify the correct catalogue record to use, so the final success rate for the books from the uncatalogued donation was 88%. Between retrospective cataloguing and donation cataloguing, therefore, we had an average final success rate of 93%.

A key part of the donation appraisal process is scarcity checking items against local and national holdings to assess rarity and research value. Using our ChatGPT transliterations, we scarcity checked this donation against both our own holdings and other Scottish and UK libraries on Library Hub just as easily as if the books were in English:

Author	Title	Publication details	GUL copies	Other copies	Scottish copies	UK copies
Alekseev, B. K. & M. N. Perfil'ev	Printsipy i tendentsii razvitiia predstavitel'nogo sostava mestnykl	Leningrad: Lenizdat, 1976	0	0	0	3
Vinogradov, A. N.	Verkhovnyi Sovet RSFSR	Moskva: Izdatel'stvo Iuridicheskaia literatura, 1967	0	0	0	4
Kurashvili, B. P.	Ocherk teorii gosudarstvennogo upravleniia	Moskva: Nauka, 1987	1	0	0	4
Kryshtanovskaia, O. V.	Inzhenery : stanovlenie i razvitie professional'noi gruppy	Moskva: Nauka, 1989	1	0	0	4
	Pervichnaia partiinaia organizatsiia : opyt, formy i metody raboty	Moskva: Izdatel'stvo politicheskoi literatury, 1979	1	0	0	4
Iudin, I. N.	Sotsial'naia baza rosta KPSS	Moskva: Izdatel'stvo politicheskoi literatury, 1973	1	0	0	5
Tkachenko, N. G.	Izbiratel'nyi uchastok	Moskva: Iuridicheskaia literatura, 1975	0	0	0	2
Grigor'ev, V.	Poriadok provedeniia vyborov v Verkhovnye Sovety soiuznykh, a	Moskva: Izdatel'stvo Izvestiia, 1975	0	0	0	1
Smirnov, V. P.	Referendum v pechati	Moskva: Mysl, 1978	1	0	0	0

Figure 3: Transliterated scarcity checks

While ChatGPT OCR transliteration works well, its batch image processing requires a paid subscription. As part of our library's existing Microsoft subscription we already have full access to Microsoft Copilot, an AI tool based on ChatGPT. We had tested an earlier GPT-4-based version of Copilot early in the project and the results were not very

accurate. In August 2025 a new version of Copilot based on GPT-5 was released, so we decided to rerun the entire project using Copilot. The results closely matched those from ChatGPT:

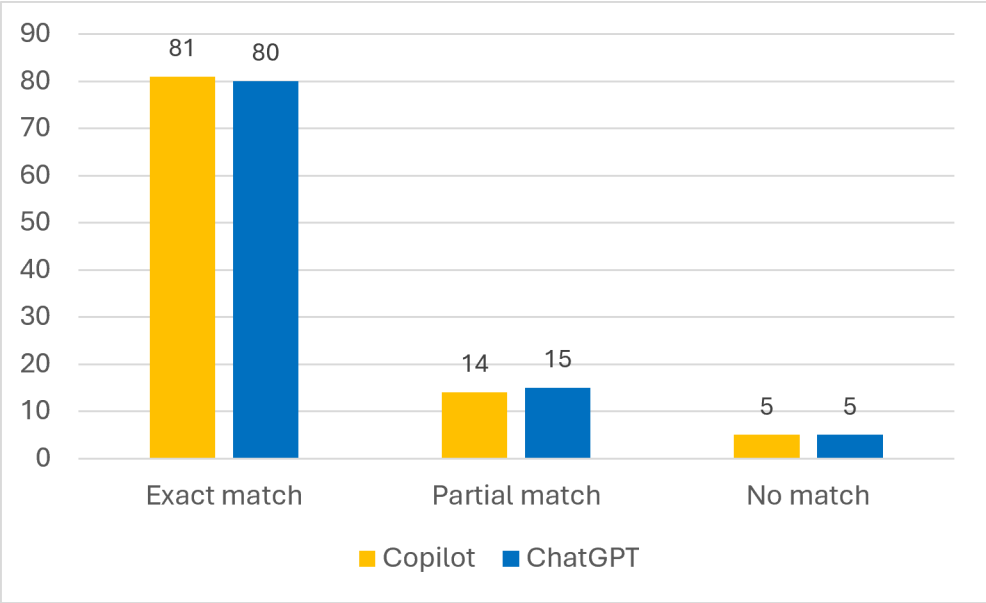


Figure 4: Copilot & ChatGPT overall results

From our study, we can therefore conclude that using AI transliterations to help catalogue non-English material works very well. AI can OCR non-English text with high accuracy, and AI transliteration is both effective and LC-compliant with around a 90% success rate in this pilot. With the help of AI tools, cataloguers can now catalogue books written in languages they can't read.

We can do this, but should we? AI is a powerful tool but it comes with serious environmental, ethical, and legal risks which we cannot ignore. For AI, analysing training data and responding to queries is an extremely intensive process, leading to "expanding demand for computing power, larger carbon footprints, shifts in patterns of electricity demand, and an accelerated depletion of natural resources" (Bashir et al., 2024, p. 2). According to a recent United Nations report on AI, these environmental impacts "are growing significantly" (United Nations Independent International Scientific Panel on Artificial Intelligence, 2026, p. 34). With England having just recorded its hottest June on record (Carrington and Kassam, 2026), we urgently need to start reducing, not increasing, our energy usage and carbon footprints.

There are also ethical and potential legal issues surrounding AI use. The training data used for AI models is based on a "vast amount of publicly available data, some of which may include copyrighted content or unpublished proprietary information" (Dover and Grzegorski, 2025, p. 607). ChatGPT's developer OpenAI claim their ingestion of copyrighted material falls under fair use ('The New York Times Company v. Microsoft Corporation', 2023, p. 4), but Hyung Doo Nam has argued that while fair use "was originally designed as an exceptional system for little users, ... AI developers and Big Tech companies have abused this doctrine in the name of 'public interest'"

(Nam, 2026, p. 606). As librarians who observe both the letter and spirit of copyright law, it is problematic if we use tools which arguably do not.

There are also practical implications to integrating AI into our workflows. AI models are opaque since "vendors do not always disclose the sources of their training data and how their models work, making their output difficult to understand, explain, and replicate" (Dover and Grzegorski, 2025, p. 607). The difficulty of replicating results is compounded by the observed phenomenon of AI tools "mimic[ing] human-like fatigue when performing too many similar or tedious tasks in succession" (Sun, 2025, p. 276), which makes them unreliable for metadata tasks which demand consistency. Integrating programs provided by external companies into our core workflows also exposes us to both pragmatic and financial risks, since the service offered by these companies could at any time become prohibitively expensive or even withdrawn entirely (Michalak and Ellixson, 2024, p. 317).

So while it has been recently asserted that "there is an urgent need for the adoption of AI by libraries" (Harisanty et al., 2023, p. 520), for this workflow we disagree. Instead, we propose using Python. Python is a programming language designed for data analysis and used extensively in data science. At the University of Glasgow Library, we have developed and started using our own Python tools to carry out automated scarcity checking and large-scale metadata analysis. We therefore decided to investigate whether Python could be used for OCR and transliteration of non-English texts.

Well-established open-source Python libraries are already available for both OCR and LC-compliant transliteration of Cyrillic text; our own project uses Pytesseract (Lee, 2023) for OCR and Domovyk (Goodale, 2023) for transliteration. One of the key benefits of using open-source software is that in contrast to AI's black box, we can inspect the code and see exactly what is going on:

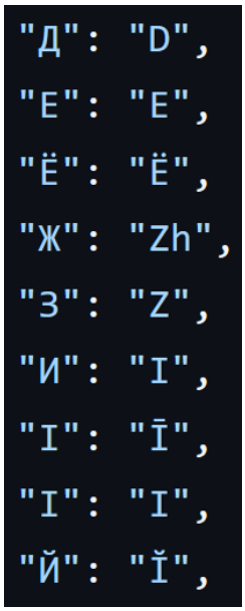


Figure 5: Excerpt of Domovyk's transliteration table

If the LC transliteration tables were to be updated, we could amend our code immediately to reflect this rather than having to wait months or potentially years for an AI tool to be updated.

Python has already been successfully used to transliterate languages as diverse as Arabic, Manipuri, Middle Persian, and even Hittite cuneiform (Farsi et al., 2025; Tour et al., 2025; Moirangthem and Nongmeikapam, 2026). Monika Kilić has used Python to transliterate Russian and Serbian Cyrillic texts. Kilić used text rather than images as input, but she nevertheless found that "the computer's ability to parse and transliterate hundreds of titles in a few seconds offers a rate of efficiency that is unmatched" (Kilić, 2025, p. 90). She used Copilot in her project to generate and revise the Python code and found that several stages of iteration were necessary to achieve accurate results ((Kilić, 2025, p. 82). Several other projects have also used Python for both OCR and transliteration (Shrividya et al., 2025; Roy, 2025).

We have now written an experimental Python program which follows the same workflow we previously used with ChatGPT. It takes an image of a Russian title page, uses OCR to convert it to Cyrillic text, then transliterates this into LC-compliant Latin text which can then be used for copy cataloguing or scarcity checking.

This has several benefits compared to using AI:

- The program requires no subscription, account, or even internet access.
- It comes at a much lower environmental cost than AI, using negligible computing power while still processing an image in a few seconds.
- Being based on open-source code which was made freely available for use and reuse, it sidesteps the ethical and legal concerns of using AI tools.
- All the code is inspectable and modifiable, so we can adapt the Python code to our evolving requirements.
- Perhaps most valuably, developing our own in-house Python solution is an opportunity to upskill our own staff rather than outsourcing expertise to external companies.

While the development of our Python program for OCR and transliteration of Russian books is still in its early stages, the initial results have been promising. The program can transliterate Russian text very effectively, likely because this is a straightforward process of converting one character to another according to a crossmapping table.

We have, however, found the OCR part of the process more challenging. In our testing, Python OCR works very well with full pages of text but is less effective on pages with areas of blank space (such as title pages). However, if we crop out smaller sections of text to use as the input images the results are much more accurate:

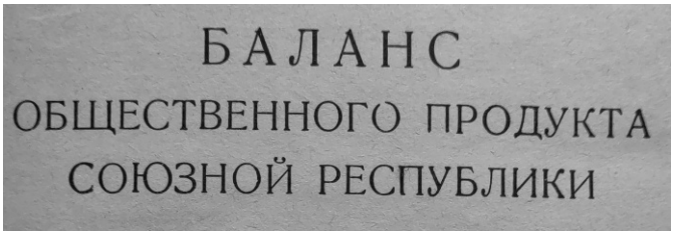


Figure 6: UoGL Python programme input

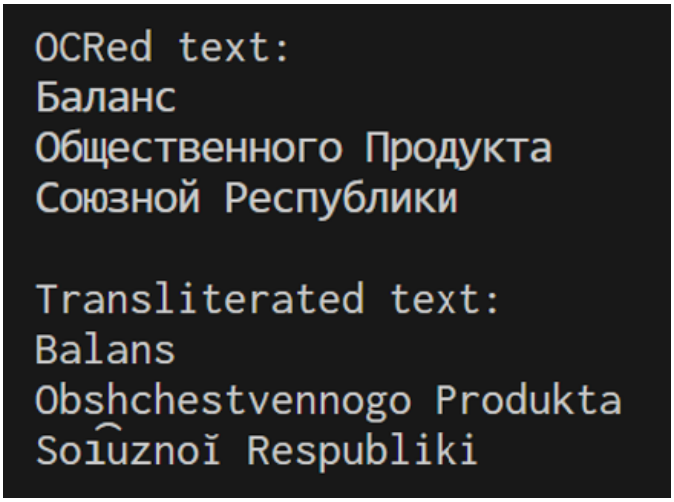


Figure 7: UoGL Python programme output

Python OCR is also much more sensitive to input image quality. While AI tools are generally quite forgiving of poorer image quality, effective Python OCR requires images which are in sharp focus and black and white or grayscale.

Our future plans for this Python OCR project are initially centred on further development and refinement of the process. We want to carry out more testing to determine the optimal image preprocessing pipeline for the input images, because other projects using the same Tesseract OCR engine as us have improved output quality by applying image filtering (Shrividya et al., 2025, p. 2). Our program could then automatically apply the optimal preprocessing pipeline to each input image.

We will carry out more extended testing to compare the accuracy of our Python OCR and transliteration tool against both AI tools and human cataloguers. Once we are satisfied with the accuracy of the output, we will investigate expanding the tool to support other languages and scripts. When our Python tool is feature-complete and working reliably, we will make it open-source and freely available online for anyone to use and modify.

In conclusion, AI tools such as ChatGPT and Copilot are capable of accurate OCR and transliteration of Russian texts. They can provide high-quality output even with low-quality input images and are very easy to use. However, they come with serious environmental, legal, and ethical drawbacks.

Python can transliterate accurately but its OCR quality can be more variable without image preprocessing. Our Python tool's great strength is that it is fully customisable, transparent, efficient, and does not have any associated environmental or ethical costs. It is, however, still in its early experimental stages and requires more development before it is ready to be integrated into daily work.

AI tools are currently the best technological option to help cataloguers without specialist language skills transliterate non-English texts. But with further development, we are confident that a Python tool will be a viable alternative which for us is preferable to using AI.

But whether we use Python or AI tools, we must remember that for all their sophistication they are just that: tools. The Program for Cooperative Cataloguing's recent argument that "AI cannot replace human expertise" (PCC Task Group on AI and Machine Learning for Cataloging and Metadata, 2025, p.5) applies to Python just as much as AI. If we rely solely on these tools for transliteration and don't have cataloguers with language skills to check the output, errors will accumulate over time with a consequent gradual degradation of metadata quality.

These tools can be useful to help us create a basic catalogue record to make an uncatalogued item discoverable, but we cannot and should not assume their outputs are correct. Cataloguers with language skills must remain an integral part of our workflows, because "while the fantasy of being able to delegate our responsibilities to the touch of a button is potentially alluring, we must not sacrifice the quality and depth of our work in leaping to chase it" (Engel et al., 2025, p. 273).

References

Bashir, N. et al. (2024) 'The Climate and Sustainability Implications of Generative AI', in *An MIT Exploration of Generative AI*. Available at: https://doi.org/10.21428/e4baedd9.9070dfe7 [Accessed: 18 August 2026]

Carrington, D. and Kassam, A. (2026) 'UK and Switzerland record hottest ever June day as health emergencies surge in Europe', *The Guardian*, 25 June. Available at: https://web.archive.org/web/20260629171559/https://www.theguardian.com/environment/2026/jun/25/highest-june-minimum-temperature-record-broken-cardiff-savage-heatwave-continues [Accessed: 18 August 2026]

Dover, A. and Grzegorski, J. (2025) 'Artificial Intelligence Through the Lens of the Cataloguing Code of Ethics', *Cataloging & Classification Quarterly*, 63(6–7), pp. 600–620. Available at: https://doi.org/10.1080/01639374.2025.2544137 [Accessed: 18 August 2026]

Engel, J. Y. et al. (2025) 'Artificial Intelligence in Library Cataloging: A Review of Literature', *Journal of Library Metadata*, 25(4), pp. 261–276, Available at: https://doi.org/10.1080/19386389.2025.2526913 [Accessed: 18 August 2026]

Farsi, F. et al. (2025) 'ParsiPy: NLP Toolkit for Historical Persian Texts in Python', *Second Workshop on Ancient Language Processing*, Albuquerque, 3–4 May. Association for Computational Linguistics, pp. 238–250. Available at: https://doi.org/10.18653/v1/2025.alp-1.17 [Accessed: 18 August 2026]

Gamage, R. and Wanigasooriya, P. (2024) 'Using Generative AI for Bibliographic Description: A Study with ChatGPT 4', *Journal of the University Librarians Association of Sri Lanka*, 27(2), pp. 257–284. Available at: https://doi.org/10.4038/jula.v27i2.8083 [Accessed: 18 August 2026]

Geroimenko, V. (2025) *The Essential Guide to Prompt Engineering: Key Principles, Techniques, Challenges, and Security Risks*. Springer.

Goodale, I. (2023) *Domovyk* (Version 1.0.1) [Computer program]. Available at: https://github.com/ian-nai/domovyk [Accessed: 18 August 2026]

Harisanty, D. et al. (2023) 'Is Adopting Artificial Intelligence in Libraries Urgency or a Buzzword? A Systematic Literature Review', *Journal of Information Science*, 51(2), pp. 511–522. Available at: https://doi.org/10.1177/01655515221141034 [Accessed: 18 August 2026]

Jiang, Y. et al. (2024) 'Exploring the Potential Applications of Generative AI in East Asian Librarianship: Use Cases, Reflections, and Inspirations', *Journal of East Asian Libraries*, 179, pp. 43–58. Available at: https://scholarsarchive.byu.edu/jeal/vol2024/iss179/5/ [Accessed: 18 August 2026]

Kilić, M. (2025) 'Using AI Tools for Slavic Transliteration to Support Cataloging Workflows: Potential Use Cases and Limitations', *Slavic & East European Information Resources*, 26(1–2), pp. 80–91. Available at: https://doi.org/10.1080/15228886.2025.2518369 [Accessed: 18 August 2026]

Lee, M. A. (2023) *Pytesseract* (Version 0.3.13) [Computer program]. Available at: https://github.com/madmaze/pytesseract [Accessed: 18 August 2026]

Library of Congress (2012a) *Russian Romanization Table*. Available at: https:/www.loc.gov/catdir/cpso/romanization/russian.pdf [Accessed: 3 September 2026]

Library of Congress (2012b) *Russian Romanization Table* [Source document]. Available at: https://www.loc.gov/catdir/cpso/romanization/russian.doc [Accessed: 3 September 2026]

Michalak, R. and Ellixson, D. (2024) 'Buy Versus Build: Navigating Artificial Intelligence (AI) Tool Adoption in Academic Libraries', *Information Services and Use*, 44(4), pp. 316–326. Available at: https://doi.org/10.1177/18758789241296755 [Accessed: 18 August 2026]

Moirangthem, G. and Nongmeikapam, K. (2026) 'Optimizing Low-Resource Machine Transliteration: A Case of Script Transition from Manipuri in Bengali Script to Meetei-Mayek', *ACM Transactions on Asian and Low-Resource Language Information Processing*, 25(5). Available at: https://doi.org/10.1145/3806198 [Accessed: 18 August 2026]

Nam, H. D. (2026) 'The Paradox of Fair Use – A Giant on a Seesaw: Restoring the Balance in Copyright and Media-Specific Issues', *GRUR International*, 75(7), pp. 605–606. Available at: https://doi.org/10.1093/grurint/ikag001 [Accessed: 18 August 2026]

'The New York Times Company v. Microsoft Corporation' (2023) United States District Court for the Southern District of New York, 1:23-cv-11195. Court Listener. Available at: https://www.courtlistener.com/docket/68117049/the-new-york-times-company-v-microsoft-corporation/#entry-1 [Accessed: 18 August 2026]

Ogungbenro, O. D. et al. (2025) 'Revolutionizing Library Services: The Impact of Artificial Intelligence on Cataloguing and Access to Information in Nigeria Academic Libraries',

Journal of Library Metadata, 25(2), pp. 99–118. Available at: https://doi.org/10.1080/19386389.2025.2475418 [Accessed: 18 August 2026]

Oyighan, D. et al. (2024) 'The Role of AI in Transforming Metadata Management: Insights on Challenges, Opportunities, and Emerging Trends', *Asian Journal of Information Science and Technology*, 14(2), pp. 20–26. Available at: https://doi.org/10.70112/ajist-2024.14.2.4277 [Accessed: 18 August 2026]

PCC Task Group on AI and Machine Learning for Cataloging and Metadata (2025) *PCC Task Group on AI and Machine Learning in Cataloging and Metadata: Final Report*. Available at: https://web.archive.org/web/20260708094646/https://www.loc.gov/aba/pcc/taskgroup/AI-and-Machine-Learning-TG-final-report.pdf [Accessed: 18 August 2026]

Roy, S. (2025) 'Global Scriptualization of Sanskrit: AI-Assisted OCR, Transliteration, and Translation Across 50+ Languages and Scripts' [Preprint]. Available at: https://doi.org/10.13140/RG.2.2.35119.60326 [Accessed: 18 August 2026]

Shrividya, M. L. et al. (2025) 'Multilingual Digitization of Ancient Manuscripts: OCR-Based Preservation and Translation', 5th International Conference on Emerging Research in Electronics, Computer Science and Technology (ICERECT), Mandya, 12–13 September. IEEE. Available at: https://doi.org/10.1109/ICERECT65215.2025.11377349 [Accessed: 18 August 2026]

Sun, L. (2025) 'Enhancing Cataloging with Generative AI: Converting Wade-Giles to Pinyin', *Cataloging & Classification Quarterly*, 63(4), pp. 267–283. Available at: https://doi.org/10.1080/01639374.2025.2508956 [Accessed: 18 August 2026]

Sussmeier, S. and Henry, J. A. (2025) 'Mind the Gap!: How Do We Ethically Bridge the Divide Between the Cataloging/Metadata Community and the World of AI?', *Journal of Library Metadata*, 25(4), pp. 241–259. Available at: https://doi.org/10.1080/19386389.2025.2525720 [Accessed: 18 August 2026]

Togia, A. et al. (2026) 'Artificial Intelligence in Greek Libraries: Perceptions, Challenges, and Readiness for Adoption', *Journal of Librarianship and Information Science. Available at:* https://doi.org/10.1177/09610006251410240 [Accessed: 18 August 2026]

Tour, E. et al. (2025) 'Potnia: A Python Library for the Conversion of Transliterated Ancient Texts to Unicode', *Journal of Open Source Software*, 10(108). Available at: https://doi.org/10.21105/joss.07725 [Accessed: 18 August 2026]

United Nations Independent International Scientific Panel on Artificial Intelligence (2026) *Preliminary Report of the Independent International Scientific Panel on AI: Evidence-Based Assessment of Opportunities, Risks and Impacts of Artificial Intelligence*. United Nations. Available at: https://web.archive.org/web/20260710091141/https://www.un.org/independent-international-scientific-panel-ai/sites/default/files/2026-07/en_Preliminary%20Report_.pdf [Accessed: 18 August 2026]

Metadata & Discovery Group

"All our wisdom is stored in the trees"[1]

using a tree to model metadata mission and vision

Helen K. R. Williams 0000-0003-1259-7097
Metadata Manager, The London School of Economics and Political Science (LSE) Library

Received: 29 June 2026 | Published: 16 September 2026

ABSTRACT

How well are the mission and vision of your metadata team articulated? Is there a clear strategy in place to achieve that vision? Are there clearly defined goals to support it? Do wider colleagues understand what happens in the metadata team, and the centrality of metadata to the institution?

A team purpose tree can help to visualise the role of a metadata team as a central component in supporting discovery, which enables teaching, learning and research in the wider context of a library and its parent institution. Mapping the library and institutional strategy to a mission and vision for the metadata team brings clarity and focus to the value of day-to-day work. It also demonstrates how metadata contributes to strategic priorities, a particularly important skill when communicating with senior managers and decision makers.

KEYWORDS metadata strategy; metadata advocacy; metadata leadership

CONTACT Helen K. R. Williams h.k.williams@lse.ac.uk The London School of Economics and Political Science

The Metadata team purpose tree is a model that has been in use at LSE since 2018. The concept was first encountered as a diagram designed to support thinking about employee engagement, in the context of an LSE Leadership programme. (Credit for the original idea goes to Tim Fuller (Fuller, no date) of the Reality Business and is used with permission).

I adapted this to a more visual framework and we used it in response to structural changes that were having an impact on our team. Specialist colleagues, such as those in the Metadata team, had ceased to be part of day-to-day customer facing activities, and were entirely focused on office-based metadata work. Conversations suggested that this shift made it harder to maintain a clear sense of the centrality of the team's work to the wider institutional mission. This carried a risk of reduced engagement and, potentially, lower job satisfaction. In response the model helped to demonstrate the importance and purpose of our work in the context of the Library as a whole and the wider institution.

[1] Kalwar, no date

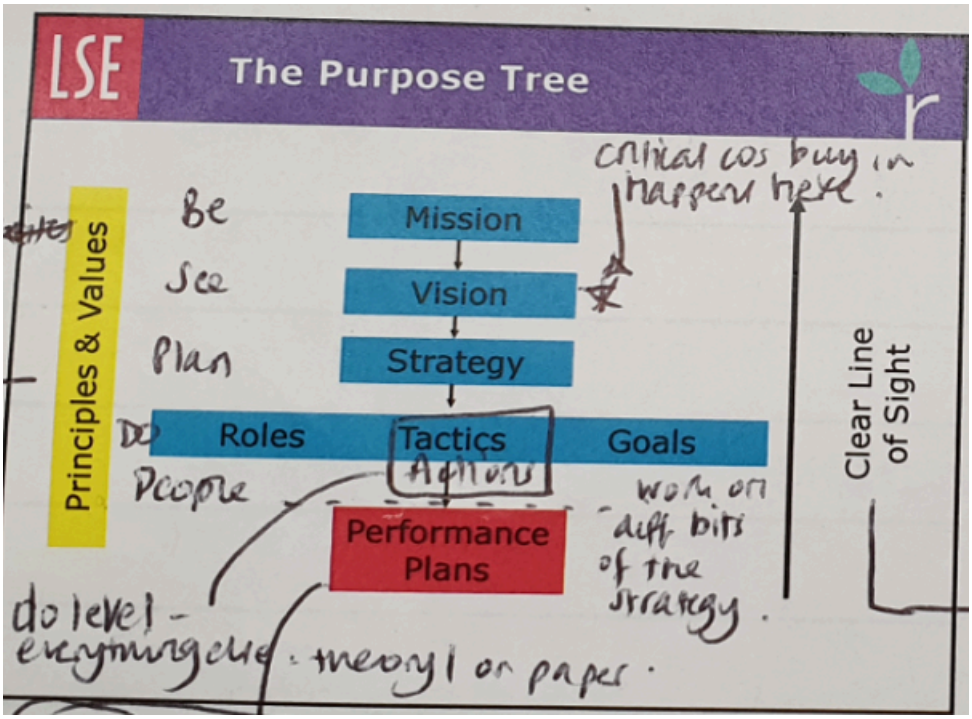


Figure 1: Original purpose tree. © Tim Fuller, The Reality Business

Over time we have discovered that in addition to supporting the team with a greater understanding of their shared purpose, the clarity on shared goals and objectives has also had wider value in supporting:

- confidence in contributing to the strategic direction of the university,
- increased ability, across the team, to advocate for metadata and its central role in institutional priorities,
- a more accessible route for wider Library colleagues to gain a better understanding of the Metadata team and its value proposition,
- increased visibility of the team and its transferrable skills, leading to more collaborative opportunities and an extended portfolio of work,
- development of a wider remit and skillset which helps with future proofing as technology, services and institutional needs evolve, and
- making the case for staffing and tying funding requests into the strategic direction of the university.

The purpose tree also aligns with John Stepper's "working out loud' model (Stepper, no date), an approach I was keen to foster in the team. The details and information we include in our purpose tree help us to have transparency within the team, and to be purposeful about eliminating the invisibility that can all too easily occur when individuals within the team are involved in different areas or projects. It supports attitudes of openness, visibility and connection as key values across different strands of work and helps to build an environment for shared problem solving and emerging thinking. At LSE I created our purpose tree and then shared it with the team, but in hindsight it would have been more beneficial if we had developed it together as a

team. This would have collated richer and broader input, as well as encouraging greater ownership within the team.

Karen Smith Yoshimura has observed,

> "An indispensable skill in these times is the ability to explain in a concise, compelling, and relatable way to administrators and other decision makers why resources should be allocated to creating and maintaining structured metadata. The audience for these appeals frequently will not have a background or past work experience in this type of work. Many professionals in need of ideas for making arguments for metadata as a strategic priority could benefit from exchanging ideas and techniques for effective advocacy." (Smith Yoshimura, 2017)

Reflecting on this, and on the increasing importance of metadata advocacy, it seemed appropriate not only to share the purpose tree model with the wider metadata community via the MDG conference, but to facilitate an interactive session whereby participants could discuss together in small groups how to articulate the mission, vision and strategy of a metadata team, and then share feedback with all conference participants to foster knowledge exchange of valuable professional expertise.

With a blank purpose tree[2] in hand our interactive session took participants through some questions to help them consider the following:

Mission – think about why you exist as a team:

- What does your team do? What is your main reason for being a team?
- Why do you do what you do? What is the purpose behind it?
- Who are you doing it for? Who is your audience?
- What is the benefit of what you do?

Based on our experience I recommend making the mission statement organisation-specific, timeless, and around 15–30 words so that it is easily memorable.

Participants were provided with LSE Library's Metadata team mission statement as an example to stimulate discussion if needed.

> "Facilitate discovery of LSE Library collections and LSE research for the LSE community and the wider world."

Vision – think about what it looks like if you fulfil your mission:

- What do you want to achieve and see in your team?
- What is it that your everyday work contributes towards the mission?

[2] A blank purpose tree template is available as an additional resource at https://journals.cilip.org.uk/catalogue-and-index/article/view/971

Due to fast paced technology changes, I would recommend thinking in a 1–5 year time frame, depending on what is most applicable in the local context. Like the mission statement ideally the vision will be something that teams can share and communicate clearly.

We summarise the LSE Metadata team vision as:

> "Create and manage comprehensive and authoritative metadata which adds value to LSE's outstanding collections, contributes to the global web of data, and facilitates wide use of the collections."

Strategy – think about your plan to achieve the vision:

The strategy begins to get a bit more practical, but is still quite high level and general, rather than focusing on day-to-day tasks. I suggest aiming for 5–6 bullet points, which might cover processes, training, current awareness, collaboration and promotion.

At LSE our plan to make our vision a reality is to:

- "Promote metadata as an institutional asset
- Embed Metadata representation across Library activity
- Encourage team development
- Ensure efficient processes at all stages of the metadata lifecycle
- Maintain current awareness of international metadata trends
- Engage with internal and external colleagues on collaborative opportunities."

Mission, vision and strategy feed into the **goals**, **actions and roles** section of the purpose tree. This is much more organisation specific, so the MDG interactive session did not break participants into groups for further discussion, but at LSE we use this section to represent:

- library and institutional strategic aims as our goals,
- business as usual and operational library project work as our actions, and
- people working on each area as our roles.

This helps to depict the connection between the work undertaken by the team, and Library and institutional priorities. The purpose tree is updated, on average, 2 or 3 times a year to maintain currency. Specific work tasks or areas identified here feed directly into annual career development reviews and to the objectives contained within those, which are monitored in one-to-one meetings with line managers throughout the year.

The purpose tree has helped embed our work within the wider institutional context and has enabled the team to see how they fit into the bigger picture. It has also been a constructive visual depiction of the unique value the team brings to institutional

strategies and priorities. Its use as an advocacy tool can help colleagues outside the metadata community of practice to understand our roles and our contributions to the wider organisational context. We have found this to be a very valuable model for us as a team, and it is shared in the hope that it may be adaptable to other metadata contexts.

AI statement

I used Microsoft CoPilot to proof-read my final draft of this article, incorporating some of its suggestions to improve clarity.

References

#

Metadata & Discovery Group

The Metadata Team and institutional research visibility

Helen K. R. Williams 0000-0003-1259-7097
Metadata Manager, The London School of Economics and Political Science (LSE) Library

Received: 29 June 2026 | Published: 16 September 2026

ABSTRACT

In recent years metadata teams in many institutions have evolved to support the creation and management of metadata for institutional research outputs as well as for library collections. As organisational priorities around research visibility, impact and open scholarship grow there are new opportunities for metadata teams to contribute to related strategic goals.

This article explores how the Metadata team at LSE Library has extended its traditional metadata role to support LSE's Research for the World strategy. It outlines the creation of new metadata sources to strengthen open, structured metadata representing institutional infrastructure, and examines engagement with Wikimedia platforms to improve the discoverability and accessibility of researchers and their outputs in an increasingly complex digital ecosystem. The article also discusses the development of a combined Wikimedian in Residence and Research Visibility Champion role, highlighting the potential for metadata teams to support open science, and research dissemination and impact through collaboration, innovation and the strategic application of metadata expertise.

KEYWORDS research visibility; Wikimedia platforms; scholarly communication

CONTACT Helen K. R. Williams h.k.williams@lse.ac.uk The London School of Economics and Political Science

The Metadata team at LSE is committed to ensuring that its activities support the strategic priorities of the institution. Metadata is positioned as a foundational service which underpins the effective management of institutional knowledge. This alignment is made explicit in a "purpose tree" model which illustrates how metadata contributes to wider organisational goals. By applying metadata expertise beyond traditional library boundaries, and by engaging in collaborative partnerships, the team demonstrates its value in supporting LSE's *Research for the World strategy* (London School of Economics and Political Science, 2025). The strategy defines the "timeless purpose of a university – to advance the frontiers of knowledge, deepen understanding of the world and share that knowledge for the public good." (London School of Economics and Political Science, 2025, p. 4) It also sets out an intention to improve "the visibility and accessibility of our research to a wide range of audiences"

with the aim of achieving a "global impact" in tackling "the world's most urgent challenges" (London School of Economics and Political Science, 2025, pp. 8, 12 and 14).

Since 2020 the Metadata team has worked with Wikidata to increase the discoverability of the Library's unique and distinctive collections. In 2025 this work was extended to explore how such an approach could be transferrable to supporting the strategic priorities of the university's research strategy, particularly in enhancing research visibility.

Collaborative work with colleagues at Montana State University (MSU) on the use of Wikidata and knowledge graphs to enhance the discoverability of library people, places and services in machine-readable contexts (Clark, Williams and Rossman, 2022), demonstrates that research impact in a complex, disparate and crowded landscape depends on strong metadata foundations. Institution-level metadata describing organisations, researchers and research outputs needs to be optimised for the Semantic Web and AI tools, supporting discovery across the online environments used by both human audiences and machine agents.

Metadata practice often focuses on the creation of metadata for research outputs themselves, but it is also important for search engines, web-based tools, LLMs, and the AI tools they power, to have a correct understanding of institutional structure, enabling them to interpret the metadata relating to researchers and outputs. This is captured in an LSE Impact blog post in which the authors write:

> "Infrastructure ... reflects how we organise trust, authority and participation. In scholarly communication, infrastructure shapes what gets seen, who gets recognised and how research is understood and used. These decisions, often invisible, carry real consequences for equity, access and accountability." (Chodacki et al., 2025)

In support of LSE's *Research for the World strategy*, this work aimed to explore how metadata could:

- increase the discoverability of researchers and their association with LSE,
- ensure research outputs are accurately linked to appropriate organisational units,
- surface previously unrecognised associations between people, institutions and concepts, and
- provide a structured foundation to support the potential for greater reach, impact and re-use of LSE scholarship by both human and machine audiences.

An initial review of Google search results for LSE departments and research centres indicated a strong web-based representation of institutional infrastructure. However, existing metadata sources did not appear to support a comparable understanding in AI tools, such as CoPilot, which were unable to provide a complete or accurate picture, despite experimenting with various prompts. This suggested that these tools lacked

access to the foundational underlying metadata that would appropriately contextualise information about people and research outputs.

In response, structured organisational metadata was collated and added to Wikidata with the aim of improving this situation. Creating and editing Wikidata entities for institutional units, and their alignment with external identifiers, proved more complex than initially anticipated. Existing institutional identifiers were inconsistent and incomplete, underscoring the importance of actively managing organisational data in open, shared infrastructures such as Wikidata, from which other services may subsequently draw.

An audit revealed that 62% of departments and research institutes were either incorrect or missing in the Library of Congress Name Authority File (LCNAF). The Policy, Training, and Cooperative Programs Division of The Library of Congress responded to requests for correction and the creation of new authority records. Similarly, many International Standard Name Identifiers (ISNIs) contained errors, lacked key information, or did not exist. Amendments could be submitted via the ISNI online improvements tool, and new identifiers requested via the ISNI team at the British Library. These identifiers were then reconciled and linked across services to ensure consistency.

A review conducted six months after this work was completed showed that Copilot was subsequently able to provide a correct list of departments and research institutes, including URLs – functionality that had not been observed previously. While a direct causal relationship cannot be demonstrated, this outcome suggests that improvements to shared, structured organisational metadata may positively influence the quality of information returned by LLM-based tools. Subject to team capacity, this approach could be extended to all research centres affiliated with top-level organisational units. In addition, following the example of work at MSU (Clark, Williams and Rossman, 2022), Google Business Profiles could be claimed and curated for each unit in discussion with the university's communications division.

This foundational metadata work formed the basis for a research optimisation and visibility pilot exploring the role of Wikimedia platforms within these workflows. This was limited to four researchers due to the manual effort required to test the viability of sustained activity in this area. As with all Wikimedia-related work in the Library, a risk-managed approach was adopted. Creating metadata in external systems, beyond institutional control, is a departure from the tighter governance over metadata in LSE's internal systems. This was addressed using a Data Protection Impact Assessment and transparent communication about engagement with Wikimedia. This included articulating the benefits of contributing LSE-related metadata so that it can be enhanced by other editors, linked to external data, and re-used by web-based tools, researchers and machine agents.

Initially the pilot focused on the role of metadata in providing further value for our researchers and their outputs using Wikidata. In practice this involved a structured workflow:

1. Create or edit a Wikidata entity for the individual.
2. Use SPARQL queries to obtain publications on Wikidata already linked to the individual.
3. Use SPARQL queries to obtain a list of publications using the "author name string" property (indicating that links between the author and outputs are not yet established).
4. Use the Author Disambiguator tool to create missing links between the author and their outputs.
5. Export research outputs for the individual from the institutional CRIS and crosscheck with Wikidata.
6. Use the BHL to Wikidata tool to bulk add missing publications with DOIs, linking these to the individual via the Author Disambiguator where necessary.
7. Create a spreadsheet of metadata for outputs without DOIs and upload these to Wikidata in bulk using OpenRefine.

The bibliographic and person data contributed to Wikidata through this workflow connects to related entities already present in the platform. Researchers can then use tools such as Scholia to access visualisations of their outputs and associated data.

Wikidata has also been identified as a tool with "considerable potential...in social sciences in particular" (Arroyo-Machado, W., 2024, p. 4)[1], reinforcing the value of adding metadata for LSE's research outputs to Wikidata to enable connections and support wider discovery. Researchers participating in the pilot were provided with a short slide deck outlining the role of Wikidata, Wikipedia and Wikimedia Commons in research visibility, their relevance to the research process, and the potential benefits for individual researchers, including:

- enables Wikipedia editors to reference research outputs more easily in relevant articles,
- increases the visibility of research metadata to the LLMs underpinning generative AI tools,
- connects information about researchers and their work within a continually expanding network of related entities,
- supports visualisation through tools such as Scholia,
- enhances re-use by external systems, including library catalogues, that draw on Wikidata, and
- improves discoverability, with the potential to increase citations and research impact.

[1] Author's translation.

While these generalised benefits are persuasive and can be demonstrated to researchers through connected metadata and visualisations, measuring their impact remains challenging. Concrete evidence of value is difficult to obtain, particularly when seeking to justify the investment of staff time.

One of LSE's founding principles is that widely shared knowledge can contribute to a better society. In practice, however, research can remain hidden beyond academic audiences, presenting a barrier to the institutional aim of producing research which has an impact on society's greatest challenges. Metadata teams are well placed to address this, playing a pivotal role in improving the discoverability and accessibility of institutional research for non-specialist audiences. Such activity can encourage wider dissemination and understanding, support open science, and potentially increase opportunities for public engagement.

Wikipedia has extensive global reach and is used by both human readers and machines as a data source for other tools, including search engines, voice assistants and AI tools. Consideration was therefore given to whether engagement with Wikipedia could complement the emerging Wikidata activity in the team. This revealed that similar work was already underway elsewhere, indicating growing interest in this area. The UK Reproducibility Network (UKRN) primer on Wikimedia and Open Research advocates for the role of Wikimedia in maximising the reach of research (Poulter, Sheppard and Mietchen, 2024), and the National Institute of Health Research (NIHR) has reported success in using Wikipedia to disseminate health research (Harangozó, 2024).

Drawing on this work the research visibility pilot was expanded to include editing Wikipedia subject pages based on insights from the academic outputs of the four pilot researchers. These edits were verified with metadata citations to the relevant research publications. As with the Wikidata work, it was necessary to demonstrate the impact of this to justify the staff time required to translate specialist academic content into material suitable for a lay audience. A range of metrics was explored, including:

- Altmetric scores,
- article metrics, and
- Wikipedia page views.

Each of these showed some modest increases, but the measures were necessarily limited by the small scale of the pilot and the restricted set of research outputs. A larger dataset would be required to demonstrate robust evidence of impact and support firm conclusions.

Nevertheless, the initiative demonstrated clear potential, and there was an opportunity to apply for funding to support its expansion, with the appointment of a Wikimedian in Residence (WiR) identified as the most effective means of scaling activity. Developing a job description for this role highlighted the potential for

confusion in universities when engaging with Wikipedia, particularly where there is an expectation of creating and editing biographical Wikipedia pages for institutional faculty.

Readers familiar with Wikipedia will recognise that this would be incompatible with its principles and at odds with a number of policies, including *Conflict of Interest* (Wikipedia, 2026a), *Paid editing* (Wikipedia, 2025), and *Neutral point of view* (Wikipedia, 2026b), as well as raising potential issues in relation to *Verifiability* (Wikipedia, 2026d) and *Notability* (Wikipedia, 2026c). Conflict of Interest (COI) edits are likely to be removed, and editors who persist in such activity may be blocked from contributing. Despite this the COI policy frames such activity as "strongly discouraged" (Wikipedia, 2026a), which can lead to ambiguity in interpretation.

In this context the intention was not self-promotion, but to support the sharing of knowledge in ways that enhance the global discoverability and accessibility of research and thereby contribute to positive societal impact. While this may appear distinct from the marketing approach of commercial enterprises, the same risks arise, and consequently the same policies and procedures apply. It was important to establish that the focus of the WiR role would be on supporting open science and open knowledge, and enabling institutional subject experts to contribute to this through Wikimedia platforms. Page view data indicates that subject pages generally receive significantly higher levels of engagement than biographical pages. This helped to demonstrate the value of researchers contributing specialist subject expertise to support the wider accessibility of knowledge.

With the provision of funding the Library established a combined post of Wikimedian in Residence and Research Visibility Champion. The role is based in the Metadata team for three days a week focusing on Wikimedia activity and spends two days a week with the Open Research Services team contributing to bibliometrics and wider research visibility work. The post champions the use of tools to improve research visibility and develops, promotes and integrates Wikimedia in supporting open scholarship at LSE and in the wider social science community. Its remit includes advocacy, training, advisory support and workshop delivery, alongside serving as the institutional expert in Wikimedia and wider research visibility to foster knowledge exchange, impact and engagement. The role was developed in close collaboration with Wikimedia UK and continues to be delivered in partnership with them. At the time of writing the project is approximately six months into a two-year residency, during which its impact will be evaluated and key performance indicators reported.

Historically, metadata teams have sometimes been positioned as 'back room' functions, with their contributions to strategic priorities not always fully recognised or communicated. The work outlined here demonstrates how experience and expertise in creating and managing research metadata, combined with alignment to institutional priorities, can extend the function of metadata teams into new areas that enhance institutional research visibility and make a measurable contribution to

organisational priorities. Many years ago, LSE's metadata team was described, by a kind blogger, as "some kind of indispensable repository metadata ninja unit" (Playforth, 2012), a depiction that continues to inform its work. As the metadata landscape continues to evolve, the strategic contribution of metadata teams within institutions is an area in which metadata managers increasingly seek to demonstrate value. Engagement with Wikimedia represents one way in which metadata teams can remain central and indispensable in new ways within a changing environment.

AI statement

I used Microsoft CoPilot to proof-read my final draft of this article, incorporating some of its suggestions to improve clarity.

References

Arroyo-Machado, W. (2024) 'Explorando Wikidata para la investigación en Ciencias Sociales', *Infonomy*, 2(2). Available at: https://doi.org/10.3145/infonomy.24.034 [Accessed: 14 May 2026]

Chodacki, J., Buttrick, A., Alperin, J. P., Dean, C. and Mentis, D. (2025) 'Infrastructure is a collective responsibility – lessons from scholarly publishing and metadata', *LSE Impact of Social Sciences blog*, 3 June. Available at: https://blogs.lse.ac.uk/impactofsocialsciences/2025/06/03/infrastructure-is-a-collective-responsibility-lessons-from-scholarly-publishing-and-metadata/ [Accessed: 14 May 2026]

Clark, J. A., Williams, H. K. R. and Rossman, D. (2022) 'Wikidata and knowledge graphs in practice: using semantic SEO to create discoverable, accessible, machine-readable definitions of the people, places, and services in libraries and archives', *Information Services & Use*, 42(3–4), pp. 377–390. Available at https://doi.org/10.3233/ISU-220171 [Accessed: 11 August 2026]

Harangozó, A. (2024) How Wikipedia can help to disseminate research: an innovative NIHR project, *NIHR Blogs*, 5 February. Available at: https://www.nihr.ac.uk/blog/how-wikipedia-can-help-disseminate-research-innovative-nihr-project [Accessed 14 May 2026]

London School of Economics and Political Science (2025) *Research for the World strategy*. The London School of Economics and Political Science. Available at https://www.lse.ac.uk/Research/Assets/RftW-Strategy-FINAL.pdf [Accessed: 14 May 2026]

Playforth, S. (2012) 'The value of cataloguing: CILIP Cataloguing and Indexing Group conference 2012', *The toast in the machine*, 12 September. Available at: https://playforth.wordpress.com/2012/09/12/the-value-of-cataloguing-cilip-cataloguing-and-indexing-group-conference-2012/ [Accessed 14 May 2026]

Poulter, M. L., Sheppard, N. and Mietchen, D. (2024) *Wikimedia and open research: a primer from UKRN*. Available at: https://doi.org/10.31219/osf.io/au48t [Accessed 14 May 2026]

Wikipedia (2025) *Paid editing (guideline)*. Available at: https://en.wikipedia.org/wiki/Wikipedia:Paid_editing_(guideline) [Accessed: 11 August 2026]

Wikipedia (2026a) *Conflict of interest*. Available at: https://en.wikipedia.org/wiki/Wikipedia:Conflict_of_interest [Accessed: 11 August 2026]

Wikipedia (2026b) *Neutral point of view*. Available at: https://en.wikipedia.org/wiki/Wikipedia:Neutral_point_of_view [Accessed: 11 August 2026]

Wikipedia (2026c) *Notability*. Available at: https://en.wikipedia.org/wiki/Wikipedia:Notability [Accessed: 11 August 2026]

Wikipedia (2026d) *Verifiability*. Available at: https://en.wikipedia.org/wiki/Wikipedia:Verifiability [Accessed: 11 August 2026]

Catalogue & Index | Issue 216 (2026)

Metadata & Discovery Group

Metadata as cultural stewardship

rediscovering Irish diocesan libraries through cataloguing

Bláithín Hurley 0000-0001-9872-0281
Associate Lecturer, Open University

Received: 11 July 2026 | Published: 16 September 2026

ABSTRACT

Irish diocesan libraries represent an important yet overlooked element of the nation's ecclesiastical and cultural heritage. Although many of these collections survive physically, inadequate cataloguing has limited their visibility within contemporary research environments. This article argues that metadata should be understood not simply as a technical mechanism for organising information but as an essential form of cultural stewardship. Through a comparative study of four historic diocesan libraries formerly associated with the Diocese of Cashel, Ferns and Ossory: Christ Church Cathedral Library, Waterford; St Canice's Cathedral Library, Kilkenny; the Bolton Library, Cashel, Co. Tipperary; and the Cotton Library, Lismore, Co. Waterford, it explores how differing approaches to cataloguing have shaped discoverability, preservation, institutional identity and public engagement. The article suggests that comprehensive metadata can reconcile scholarly accessibility with local stewardship, allowing historic collections to remain embedded within their original communities while participating fully in wider networks of research and cultural interpretation.

KEYWORDS metadata; cultural stewardship; diocesan libraries; heritage collections; discoverability

CONTACT Bláithín Hurley blaithin.hurley@open.ac.uk Open University

Introduction

In the preface to *Fasti Ecclesiae Hibernicae*, Rev. Henry Cotton wrote that his purpose was "to preserve ... the memory of those who have held high office in the Church, and to supply materials for the future historian" (Cotton, 1848–1878). His words capture the original purpose of many Irish diocesan libraries. Established during the eighteenth and nineteenth centuries, these collections supported theological scholarship, preserved institutional memory, and sustained the intellectual life of the clergy. They also served as repositories of local knowledge, reflecting the religious, educational, and cultural life of the dioceses in which they were created. Today, however, many remain physically intact but are only partially visible within contemporary discovery systems, limiting their contribution to historical scholarship and public understanding.

While historic libraries are often discussed in terms of conservation, preservation alone does not ensure scholarly accessibility. A collection may survive physically yet remain effectively invisible if it cannot be identified, searched, or interpreted. Metadata enables the identification, discovery, and interpretation of collections, recording provenance and preserving the institutional context in which they were created (Baca, 2016; Gilliland, 2016). It establishes intellectual connections between collections, institutions, and researchers, allowing historic libraries to participate in wider systems of knowledge. Without effective cataloguing, even significant collections risk becoming inaccessible to all but a small number of specialists.

This article argues that metadata is more than a technical mechanism for organising information; rather, it is a form of cultural stewardship. Drawing on four historic diocesan libraries associated with the Diocese of Cashel, Ferns and Ossory, it demonstrates how cataloguing has influenced discoverability, preservation, and institutional identity. In doing so, it suggests that comprehensive metadata can enable heritage collections to remain both locally rooted and intellectually accessible, ensuring that they continue to serve not only scholars but also the communities from which they emerged.

Metadata Beyond Description

Metadata is commonly defined as the structured information that enables information resources to be identified, managed, discovered, and used (Baca, 2016; Gilliland, 2016). Within heritage collections, however, its significance extends beyond technical description. Metadata records provenance, preserves institutional context, and creates the intellectual connections through which collections become visible to researchers. It is therefore not simply a tool of organisation but an essential component of cultural stewardship.

Historically, Irish diocesan libraries operated as working collections rather than heritage repositories. Their holdings of biblical commentaries, patristic writings, sermons, and ecclesiastical law supported the daily pastoral and scholarly work of the clergy (Brown, 2001; Burns, 1999). Access depended less upon formal catalogues than upon institutional memory: successive generations of clergy knew the collections intimately and transmitted this knowledge through continued use. As patterns of theological education changed during the nineteenth and twentieth centuries, these informal systems gradually disappeared. Metadata succeeded institutional memory at precisely the point when institutional memory itself began to fade.

The subsequent histories of four diocesan libraries formerly associated with the Diocese of Cashel, Ferns and Ossory illustrate how differing approaches to cataloguing have shaped their survival and accessibility. Christ Church Cathedral Library, Waterford, was transferred to South East Technological University, where professional cataloguing restored its scholarly visibility while simultaneously removing it from the cathedral community that had created it (Walton, 1989; South East Technological

University Library, 2026). At St Canice's Cathedral, Kilkenny, the transfer of the Ossory Diocesan Library to Maynooth University likewise transformed a largely local clerical collection into one accessible through modern discovery systems and national research networks (Campbell and McCormack, 2017).

The Bolton Library at St John the Baptist's Cathedral, Cashel, followed a similar trajectory. Professional cataloguing and conservation secured the collection's future only after its relocation to the University of Limerick, demonstrating how metadata can restore scholarly accessibility while fundamentally altering a collection's institutional context (Coey, Turner and McGuinn, 2007; University of Limerick Library, 2025). By contrast, the Cotton Library at St Carthage's Cathedral, Lismore, remains within its original cathedral setting but continues to rely largely upon Henry Cotton's nineteenth-century catalogue (Cotton, 1889; St Carthage's Cathedral, no date). Although physically preserved, it remains only partially represented within contemporary discovery systems and consequently underutilised by researchers.

Taken together, these four libraries reveal that metadata does far more than describe collections. It influences where they are housed, who can discover them, and how they continue to participate in scholarly and public life. Cataloguing is therefore not merely an administrative activity undertaken after preservation; it is one of the principal means through which heritage collections remain intellectually present.

Metadata as Cultural Stewardship

The four libraries demonstrate that metadata should be understood as more than descriptive information. It is an active form of cultural stewardship, shaping not only whether collections are discoverable but also how they are valued, interpreted, and sustained. Professional cataloguing creates intellectual access, allowing collections to participate in catalogues, authority files, digital discovery services, and wider scholarly discourse (Gilliland, 2016; Baca, 2016). In doing so, metadata extends the life of collections beyond their physical preservation.

The case studies also reveal a more complex relationship between discoverability and place. In Waterford, Kilkenny, and Cashel, professional cataloguing followed relocation to university libraries, where specialist expertise secured both preservation and scholarly accessibility. These interventions have undoubtedly benefited the collections, yet they have also altered their relationship with the cathedral communities in which they were created. Metadata has enabled these libraries to enter national and international research networks, but only after they had ceased to function within their original ecclesiastical settings.

The Cotton Library presents a different possibility. As the only surviving diocesan library within the Diocese of Cashel, Ferns and Ossory still housed in its original cathedral, it offers an opportunity to establish comprehensive metadata before relocation becomes necessary. A programme of professional cataloguing, authority

control, and integration with shared discovery systems would make the collection visible to researchers while allowing it to remain within the architectural, liturgical, and historical environment that gives it much of its significance. Rather than preserving access through relocation, metadata offers the possibility of preserving both access and place.

This suggests that discoverability and locality need not be regarded as competing objectives. Advances in cataloguing standards, digital metadata, and collaborative discovery systems increasingly allow intellectual access to be separated from physical location. Collections can remain embedded within the communities that created them while participating fully in wider networks of scholarship.

Towards Collaborative Stewardship

Such an approach requires collaboration that extends beyond the cathedral itself. The experience of Irish public libraries demonstrates that stewardship is no longer confined to preservation alone but encompasses interpretation, partnership, education, and community engagement (Department of Rural and Community Development, 2023). Discoverability is the essential first stage in that process. Once collections become visible, opportunities emerge for exhibitions, educational programmes, collaborative research, genealogy, digital interpretation, and wider public engagement. Metadata therefore provides the foundation upon which these activities can develop, transforming collections from largely inaccessible repositories into active cultural resources.

For diocesan libraries, this points towards a model of shared stewardship involving cathedral custodians, university libraries, public library services, heritage organisations, metadata specialists, and local historians. Each contributes complementary expertise, while comprehensive cataloguing provides the intellectual infrastructure that connects their work. My own experience within the Irish public library service has demonstrated that successful stewardship depends upon collaboration as much as conservation. When collections become discoverable, they attract new audiences, encourage partnerships, and stimulate fresh research. Metadata therefore becomes more than an exercise in information management; it is the means through which historic collections remain rooted within their local communities while participating confidently in national and international scholarship. Through collaboration, discoverability becomes not an end in itself but the beginning of a collection's continuing intellectual, cultural, and educational life.

Conclusion

The four diocesan libraries considered here demonstrate that metadata shapes far more than bibliographic description. It influences discoverability, institutional identity, scholarly engagement, and ultimately the long-term stewardship of heritage collections. While the histories of Waterford, Kilkenny, Cashel, and Lismore differ

significantly, each illustrates the central role that cataloguing plays in determining whether a collection remains intellectually accessible or gradually recedes from scholarly awareness. Together, they show that the future of heritage collections depends not only upon their physical preservation but also upon their ability to participate in contemporary systems of knowledge.

The Cotton Library offers a particularly significant opportunity. Unlike the other collections, it remains within its original cathedral setting and therefore provides the possibility of establishing comprehensive metadata before relocation becomes necessary. Its future suggests that discoverability and locality need not be regarded as competing priorities. Instead, professional cataloguing can enable historic collections to remain embedded within their original communities while participating fully in contemporary research, education, and public engagement. In doing so, it offers a model that may be applicable to many other locally significant heritage collections.

Metadata should therefore be understood not simply as a technical process but as an act of cultural stewardship. By recording provenance, preserving institutional context, and enabling discovery, it reconnects collections with the researchers, communities, and institutions they continue to serve. Henry Cotton hoped to "supply materials for the future historian" (Cotton, 1848–1878). More than a century and a half later, that aspiration remains remarkably relevant. In this sense, cataloguing is not the final stage of preservation but the beginning of a collection's continuing intellectual, cultural, and community life.

References

Baca, M. (ed) (2016) *Introduction to Metadata*. 3rd edn. Getty Research Institute. Available at: https://www.getty.edu/publications/intrometadata/ [Accessed: 1 July 2026]

Brown, S. J. (2001) *The National Churches of England, Ireland and Scotland 1801–46*. Oxford University Press.

Burns, J. H. (1999) *The Church in History*. Oxford University Press.

Campbell, Y. and McCormack, B. (2017) 'Cataloguing the St Canice's Cathedral Library Collection at Maynooth University', *An Leabharlann*, 26(1), pp. 16–20. Available at: https://www.libraryassociation.ie/wp-content/uploads/2018/11/An_Leabharlann_26_1.pdf [Accessed: 19 June 2026]

Coey, A., Turner, B. and McGuinn, N. (2007) *Bolton Library Conservation Plan*. Heritage Council.

Cotton, H. (1848–1878) *Fasti Ecclesiae Hibernicae: The Succession of the Prelates and Members of the Cathedral Bodies in Ireland* (5 vols). Hodges and Smith.

Cotton, H. (1889) *A Catalogue of the Books in the Cotton Library, Lismore*. Hodges, Figgis & Co.

Department of Rural and Community Development, City and County Management Association and the Local Government Management Agency (2023) *The Library is the Place: Information, Recreation, Inspiration. National Public Library Strategy 2023–2027*. Government of Ireland.

Available at: https://assets.gov.ie/static/documents/the-library-is-the-place-information-recreation-inspiration-national-public-library-st.pdf [Accessed: 5 July 2026]

Gilliland, A. J. (2016) 'Setting the Stage', in Baca, M. (ed.) *Introduction to Metadata*. 3rd edn. Getty Research Institute. Available at: https://www.getty.edu/publications/intrometadata/setting-the-stage/ [Accessed: 10 July 2026]

South East Technological University Library (2026) *Special Collections*. Available at: https://library.setu.ie/culture-and-heritage/collections/ [Accessed: 5 July 2026]

St Carthage's Cathedral (no date) *The Cotton Library*. Available at: https://www.stcarthagescathedral.ie/visit/the-cotton-library/ [Accessed: 10 July 2026]

University of Limerick Library (2025) *Special Collections & Archives: The Bolton Library*. Available at: https://libguides.ul.ie/special-collections-and-archives/boltonlibrary [Accessed: 19 June 2026]

Walton, J. C. (1989) 'The Library of Christ Church Cathedral, Waterford', *Decies: Journal of the Waterford Archaeological and Historical Society*, 41, pp. 2–21.

Metadata & Discovery Group

Unlocking Southeast Asia's digital heritage

discovery, partnerships, and engagement

Chuin Peng Sim 0000-0002-0740-6201
Deputy University Librarian, National University of Singapore Libraries

Gandhimathy Durairaj
Associate University Librarian, National University of Singapore Libraries

Received: 8 April 2026 | Published: 16 September 2026

ABSTRACT

National University of Singapore (NUS) Libraries' digitisation and digital stewardship programme expands open access to rare and unique Southeast Asian materials while ensuring long-term usability, trustworthiness, and preservation. This paper traces the shift from ad-hoc digitisation to a coordinated, standards-based programme anchored by Digital Gems, NUS Libraries' public gateway to digitised special collections. It argues that metadata is the operational backbone linking workflows, multilingual description and authority control, rights and ethics review, interoperability (stable identifiers, APIs, OAI-PMH harvesting), user experience, aggregation, and preservation actions. The paper also examines partnership models – loan-and-return digitisation with rights holders and co-creation with NUS units and regional collaborators – that responsibly expand collections and enrich context. Finally, it shows how engagement through teaching, research, exhibitions, and outreach moves collections from "available" to "used," generating educational and scholarly impact and encouraging further donations. It concludes with practical lessons: align discovery with preservation, adopt community standards, design for interoperability, foreground rights and ethics, and invest in partnerships and outreach to sustain living digital heritage.

KEYWORDS digitisation; metadata; preservation

CONTACT Chuin Peng Sim chuinpeng@nus.edu.sg National University of Singapore Libraries

1. Introduction: why a Southeast Asia resource hub

NUS Libraries aims to be a leading Southeast Asia resource hub by expanding discovery and access to primary sources that support teaching, research, and public understanding of the region. Southeast Asia's documentary heritage is widely dispersed and often fragile, under-described, and difficult to consult – challenges compounded by multilingual materials, diverse scripts, and shifting historical geographies. Digitisation therefore functions not only as conversion, but as an institutional strategy for access, stewardship, and scholarly enablement.

Two anchor collections shape this work: the Singapore/Malaysia Collection (SMC), a long-developed research collection on Singapore, Malaysia, and ASEAN (Lee, 2025), and the Chinese Southeast Asia Collection, which includes extensive runs of Chinese-language newspapers and periodicals, commemorative publications, school records, genealogies, and community histories (Sim, 2025). Together with NUS Libraries' broader holdings – manuscripts, rare books, private papers, maps, photographs, ephemera, and audiovisual materials – these resources support research ranging from colonial administration and socio-political change to migration networks and cultural life.

Digitisation reduces physical handling and widens access, but it also introduces long-term responsibilities: rights and ethics review, durable identifiers, reliable metadata, interoperable delivery, and preservation. This paper presents NUS Libraries' programme through three reinforcing pillars – discovery, partnerships, and engagement – anchored by Digital Gems, the Libraries' public gateway to digitised special collections. It argues that standards-based metadata is the operational backbone that connects workflows, multilingual description, integrations, user experience, aggregation, and preservation, and that partnerships and outreach are essential to keep digital heritage not only available, but actively used.

2. From ad-hoc to programmatic digitisation

Digitisation programmes often begin with opportunistic projects driven by available funding, equipment, or immediate demand. NUS Libraries' trajectory has similarly included early projects that were valuable but not yet integrated into a coordinated pipeline. Over time, the Libraries' approach has shifted from isolated digitisation efforts to a programme with clearer governance, selection criteria, metadata practices, and preservation alignment.

2.1. Early groundwork and the shift to coordination

Earlier digitisation initiatives demonstrated the research value of providing remote access to rare sources. Examples include *Lat Pau* (founded in 1881), recognised as the earliest Chinese-language newspaper in Singapore, and *Journal of Three Voyages* (Lee, 2025), which captures travel and observation during a period when China was largely closed to foreigners. These projects established proof of concept: digitised collections could be used by scholars beyond campus, could support teaching, and could preserve access to sources otherwise constrained by fragility and reading-room logistics.

The key transition occurred in 2016, when NUS Libraries established the Special Collections Unit to move digitisation from ad-hoc selection to a coordinated programme. The intent was not only to increase the volume of digitisation, but also to improve consistency in imaging practices, metadata creation, rights review, and publication workflows, and to situate digitisation within longer-term preservation planning.

2.2. Capacity building and workflow maturation

Digitisation at scale depends on throughput, quality control, and repeatable processes. In 2018, upgraded scanning infrastructure supported higher output and better image quality, raising throughput to an average of about 12,000 pages per month. Scaling required the library to clarify selection criteria (significance, rarity, condition, demand, rights feasibility, language/script representation) and to strengthen internal coordination between curatorial decision-making, imaging, metadata creation, systems, and user-facing publication.

This shift from ad-hoc to programmatic digitisation also reframed how success is measured. Rather than viewing digitisation as a completed deliverable ("we scanned X pages"), the programme increasingly evaluated outcomes such as discoverability in multiple channels; quality of descriptive metadata; inclusion of rights statements; usability for reading, citation, and teaching; and durability through preservation workflows that support fixity, replication, and future migration.

2.3. Digital Gems as a programme anchor

In 2019, NUS Libraries launched Digital Gems[1], an open gateway to rare and unique materials spanning the humanities, social sciences, and natural sciences (NUS Libraries, 2020). Digital Gems represents the public interface of the programme and a unifying publication platform for multi-format digitised content. By providing a central access point, Digital Gems also operationalises consistent workflows: items are digitised, described, rights-reviewed, quality-checked, published, and then integrated for discovery via harvesting and other pathways.

3. Metadata as the backbone: discovery, integration, and preservation

Digitisation without metadata produces visibility without findability. For rare and multilingual collections, metadata is not peripheral: it is the backbone that enables discovery, contextual understanding, interoperability, and preservation.

3.1. Metadata for discovery: consistency, authority, multilingual description

Digital collections become useful when users can locate what they need and understand what they find. Discovery depends on consistent application of core descriptive elements such as titles, creators, dates, subjects, and places. Consistency supports meaningful faceting and helps users refine large result sets. It also enables computational use: digital humanities workflows, linked data approaches, and large-scale text analysis require structured and predictable fields.

[1] https://digitalgems.nus.edu.sg/

Authority control makes search more accurate by standardising names and places, even when spellings, scripts, or boundaries change. It links different name forms to the same person or place and avoids mixing up different entities with similar names.

Multilingual description matters for Southeast Asian collections. It keeps original scripts while also supporting discovery across languages through translated or transliterated fields and shared vocabularies.

It is also important to record how items relate to each other – such as which collection or series they belong to, and whether they are versions or translations – so users can understand provenance and context.

3.2. Metadata for integration: crosswalks, URIs, APIs, and harvesting

NUS Libraries designs metadata with interoperability in mind. Interoperability requires more than choosing schemas; it requires mapping fields across systems, aligning definitions, and preserving relationships during transformation. The programme uses crosswalks to map elements between schemas while maintaining semantic integrity.

To connect metadata across systems, the Libraries emphasises the use of URIs and stable identifiers so that records can be linked reliably in discovery layers, external portals, and scholarly citations. Stable identifiers also support course links, references in publications, and integrations with external research databases – for example, databases documenting postwar Chinese popular culture in Singapore and Malaya, as reported in *Lianhe Zaobao* (An, 2025).

Standardised protocols enable machine-to-machine exchange. OAI-PMH supports metadata harvesting so that portals and aggregators can ingest records at scale and keep them updated. This reduces the need for bespoke integrations. A key outcome is that viewers and external discovery services can incorporate NUS Libraries content without custom development, lowering barriers to collaboration and expanding reach.

3.3. Preservation alignment: structure and evidence

Preservation is not only storage; it is the ability to demonstrate authenticity and sustain usability over time. NUS Libraries frames preservation as structure and evidence. Structural metadata keeps complex digital objects intelligible by recording how components relate. Preservation metadata records fixity checks and actions, providing an audit trail that supports trust and accountability. By aligning with established long-term archiving practices, the Libraries aims to ensure that packages can move across systems with minimal effort, reducing vendor lock-in and future migration risk.

4. Digital Gems: the public gateway for open access

Digital Gems is the primary public platform through which NUS Libraries shares digitised special collections. It provides access to rare and unique materials across multiple disciplines and formats, supporting both close reading and broader exploration.

4.1. Scope and highlights

Digital Gems is open to all and includes rare books, manuscripts, private papers, journals, newspapers, drawings, pamphlets, photographs, and maps. It hosts over 30 collections and more than 210,000 items.

Key highlights include

- the *Biodiversity Library of Southeast Asia* (BLSEA)[2], an open access library of rare and unique biodiversity literature on Southeast Asia drawn from the Lee Kong Chian Natural History Museum (LKCNHM) and NUS Libraries (NUS Libraries, 2017); and
- the *Bugis–Makassar Repository*[3], developed with Universitas Muslim Indonesia and NUS Southeast Asian Studies, featuring rare manuscripts, journal-logs, correspondence, and oral histories from South Sulawesi, including the 19th-century Daeng Paduppa journal-log written primarily in Lontara (NUS Libraries, 2024).

Digital Gems also features major private papers collections, such as

- the Bai Yan collection[4] – Bai Yan (1920–2019) was a Singapore performing-arts figure – documenting local stage culture through scripts, clippings, and correspondence (Cai, 2023),
- the Wang Sha private papers[5] – Wang Sha (1925–1998) was a well-known Singapore comedy and entertainment personality – comprising scripts, costumes/props, recordings, and memorabilia related to popular entertainment (Sim, Toh & Lin, 2025), and
- the Yeap Chor Ee private papers[6] – Yeap Chor Ee (1867–1952) was a Penang businessman and philanthropist – containing family and business records that shed light on diasporic enterprise and philanthropy (Sim et al., 2024).

Digital Gems further includes Guy Charles Madoc's 1943 wartime manuscript *An Introduction to Malayan Birds*[7]; Madoc (1911–1999) is associated with ornithological

[2] Biodiversity Library of Southeast Asia: https://digitalgems.nus.edu.sg/browse/collection/50224
[3] Bugis–Makassar Repository: https://digitalgems.nus.edu.sg/browse/collection/484179
[4] Bai Yan collection: https://digitalgems.nus.edu.sg/browse/collection/67695
[5] Wang Sha private papers: https://digitalgems.nus.edu.sg/browse/collection/508505
[6] Yeap Chor Ee private papers: https://digitalgems.nus.edu.sg/browse/collection/82544
[7] *An Introduction to Malayan Birds*: https://digitalgems.nus.edu.sg/view/466393

work in Malaya, and the manuscript contributes to the documentation of Malayan birdlife during the wartime period (NUS Lee Kong Chian Natural History Museum, 2025). More than 100 historical newspapers in Chinese, Malay, and English further strengthen the platform as a source base for research on society, culture, and everyday life in Southeast Asia.

4.2. User experience and discovery features

Digital Gems supports discovery through faceted search by date, creator, subject, place, and language; thumbnails and structural navigation for complex items; deep-zoom viewing; full text search for OCR-enabled newspapers; collection pages that add context; and stable identifiers that support citation and course linking.

4.3. Aggregation and external discovery pathways

Digital Gems is designed to be harvested and integrated. A key example is metadata integration with the National Library Board (Singapore) OneSearch platform (National Library Board Singapore, n.d.), enabling national-level discovery. Digital Gems metadata is also exposed for harvesting by other aggregators, reducing duplicated effort and widening reach by placing materials where users already search.

5. Governance and workflow: from selection to publication

Sustaining a digitisation programme requires governance that protects materials, ensures ethical and legal compliance, and maintains quality at scale.

Selection balances scholarly value, institutional priorities, and feasibility: significance and rarity; condition and handling; rights status; language and script representation; anticipated use; and partnership potential (e.g., translation and contextualisation). Digitisation follows careful handling protocols, calibrated imaging, and quality control. Metadata creation includes descriptive, administrative, technical, and preservation components. Rights and sensitivity reviews occur prior to publication. After publication, usage and feedback inform iterative improvement.

6. Partnerships: rights holders, departments, and regional collaboration

Partnerships expand access, bring subject expertise into description, and strengthen legitimacy through co-creation and trust.

6.1. Rights holders and loan-and-return

Important heritage sources often remain in private custody. NUS Libraries therefore uses a loan-and-return model where appropriate: materials are loaned for digitisation and returned, while digital surrogates (with agreed rights statements) are made available.

Examples include the Riau Lingga Sultanate Family Papers[8] – historical documents, manuscripts, and correspondence of the royal family of the Riau–Lingga Sultanate (1824–1911), based on Penyengat Island, which shed light on the administration, culture, literary production, and Islamic intellectual traditions of a Malay polity that later came under Dutch control (Diyana, 2025a) – and Guy Charles Madoc's wartime manuscript *An Introduction to Malayan Birds* (NUS Lee Kong Chian Natural History Museum, 2025). Both were digitised and returned to their owners. These collaborations depend on trust, clear permissions, careful handling, and transparent documentation.

6.2. Internal partnerships within NUS

With the Lee Kong Chian Natural History Museum, BLSEA selects, digitises, and shares biodiversity resources, supporting research, conservation, and teaching. With NUS Geography, historical maps have been scanned and georeferenced under relevant permissions to support spatial analysis[9] (NUS Libraries, 2018). With Architecture partners, collections such as Singapore Pioneer Architects[10] preserve early studio materials and student projects that illuminate Singapore's architectural heritage.

6.3. Flagship external partnership: Bugis–Makassar and Daeng Paduppa

The Bugis–Makassar Repository is a flagship partnership with Universitas Muslim Indonesia (UMI) and NUS Southeast Asian Studies. In October 2023, UMI donated the original Daeng Paduppa manuscript[11] to NUS Libraries (NUS Department of Southeast Asian Studies, 2023). The programme is preserving, digitising, and translating the journal into Indonesian and English for open access via Digital Gems.

Fieldwork in Sulawesi has broadened the initiative to include oral histories with nakhoda and master shipbuilders, documentation of keris craftsmanship, identification of additional manuscripts (including a possible 1600s keris text), and on-site digitisation of a ship captain's diary. These sources illuminate maritime networks and their ties to Singapore, grounding regional history in local voices (Diyana, 2025b).

6.4. Subject expertise and co-description

Librarians rely on researchers and cultural custodians to guide selection and contextual description for complex sets such as private papers and photographs. Expert inventories and identification accelerate metadata creation and improve quality, shortening time to publication while deepening interpretive context.

[8] Riau Lingga Sultanate Family Papers: https://digitalgems.nus.edu.sg/browse/collection/486746
[9] Historical maps of Singapore: https://libmaps.nus.edu.sg/
[10] Singapore Pioneer Architects: https://digitalgems.nus.edu.sg/browse/collection/12377
[11] Daeng Paduppa manuscript: https://digitalgems.nus.edu.sg/view/520515

7. Digital preservation infrastructure

Long-term access requires continuous preservation as technologies change. In 2023, NUS Libraries adopted Libnova as a Digital Preservation System to manage the long-term preservation of born-digital and digitised assets, including NUS Archives and photo databases. Preservation work includes scheduled fixity checks, storage replication across locations, format identification, and detailed logging of preservation actions. Governance includes multiple checksums, multiple copies across varied storage media and locations, adequate documentation and descriptive metadata, and security controls such as encryption and access restrictions.

A core objective is to preserve integrity and portability so that preservation packages can move across systems with minimal effort, supporting future migration and reducing dependency on any single vendor.

8. Engagement, teaching, research, and public value

Digitisation creates value when materials are used. NUS Libraries supports engagement through social media highlights, donor presentations, seminars, talks, exhibitions, and online showcases.

Collections such as Bai Yan's papers, the Bugis–Makassar programme, Yeap Chor Ee, Wang Sha, and Madoc's wartime manuscript have been activated through public programmes and outreach (Cai, 2023, NUS Libraries, 2024; Sim et al., 2024; Zhang & Xin, 2025; NUS Lee Kong Chian Natural History Museum, 2025). Engagement can also generate donations. After the Wang Sha programme, Chin Whai, an 87-year-old veteran of Singapore's Chinese performing arts scene, announced a donation of private papers and artefacts (Zhong, 2025), and the family of the late Tang Hu – Wang Sha's stage partner – indicated plans to organise and donate materials. Together, the private collections of Wang Sha, Chin Whai, and Tang Hu will help reconstruct post-war popular entertainment and expand primary sources for study.

9. Evidence of impact and usage

Digital Gems shows sustained growth in usage, with page views and downloads increasing over time. The platform's importance for teaching and research is also visible in spikes in user queries and helpdesk tickets during maintenance downtime, suggesting reliance by learners and researchers.

BLSEA appears in coursework and theses across ecology, conservation, geography, and environmental history. Bugis–Makassar materials and private papers feature in assignments, seminars, and exhibitions. Partnerships further broaden reach, while preservation infrastructure provides technical assurance of integrity, supporting continued use.

10. Lessons learned and reflections

Discovery and preservation must advance together. Consistency and authority control, combined with multilingual description, improve search quality and reach. Rights and ethics underpin trust through clear item-level rights, sensitivity review, transparent documentation, loan-and-return models, and preservation records. Collaboration multiplies value by bringing in content and expertise; fieldwork anchors collections in local voices and practices. Workflow discipline – selection, handling, imaging, description, rights review, quality control, publication, and preservation – raises standards and reduces rework.

11. Conclusion

Digital heritage is a shared endeavour. Technology enables it; standards stabilise it; rights and ethics guide it; partnerships and engagement bring it to life. NUS Libraries' approach – programmatic digitisation, metadata-driven discovery, interoperable integrations, partnership-based expansion, active engagement, and preservation-aligned governance – aims to connect Southeast Asian heritage to learners, researchers, and the public, accessible today and safeguarded for the long term.

References

An, S. Y. (2025) '"战后新马华人的大众文化"数据库启用 [Database on Chinese Popular Cultures in Postwar Singapore and Malaya Launched]', *Lianhe Zaobao*, 14 February. Available at: https://www.zaobao.com.sg/news/singapore/story20250214-5799326 [Accessed: 28 March 2026]

Cai, X. (2023) '家属将白言生前收藏捐赠给国大图书馆 [Bai Yan Collection Gifted to NUS Libraries by Family]', *Lianhe Zaobao*, 14 August. Available at: https://www.zaobao.com.sg/entertainment/story20230814-1423626 [Accessed: 27 March 2026]

Diyana, N. (2025a) 'A Bugis Journal-Log', *LINUS @ NUS Libraries*, 30 April. Available at: https://blog.nus.edu.sg/linus/2025/04/30/a-bugis-journal-log/ [Accessed: 29 March 2026]

Diyana, N. (2025b) 'Manuscripts and Masterpieces: Librarians Exploring Sulawesi's Treasures', *LINUS @ NUS Libraries*, 4 February. Available at: https://blog.nus.edu.sg/linus/2025/02/04/manuscripts-and-masterpieces-librarians-exploring-sulawesis-treasures/ [Accessed: 29 March 2026]

Lee, C. E. (2025) 'From SMC to SMC – Special Malaysian Collection to Singapore/Malaysia Collection', *LINUS @ NUS Libraries*, 30 October. Available at: https://blog.nus.edu.sg/linus/2025/10/30/from-smc-to-smc-special-malaysian-collection-to-singapore-malaysia-collection/ [Accessed: 29 March 2026]

National Library Board Singapore (n.d.) '*OneSearch*'. Available at: https://search.nlb.gov.sg/onesearch/ [Accessed: 29 March 2026]

NUS Department of Southeast Asian Studies (2023) *'Collaboration on the preservation, digitisation, and translation of the Daeng Paduppa manuscript by NUS Libraries and Universitas Muslim Indonesia (UMI)'*. Available at: https://fass.nus.edu.sg/sea/2023/10/10/collaboration-

on-the-preservation-digitisation-and-translation-of-the-daeng-paduppa-manuscript-by-nus-libraries-and-universitas-muslim-indonesia-umi/ [Accessed: 26 March 2026]

NUS Lee Kong Chian Natural History Museum (2025) *'A Family Legacy: Ms Fenella Madoc-Davis and the Preservation of Ornithological History'*. Available at: https://lkcnhm.nus.edu.sg/a-family-legacy-ms-fenella-madoc-davis-and-the-preservation-of-ornithological-history/ [Accessed: 26 March 2026]

NUS Libraries (2017) 'BLSEA: The Beauties and Beasts of Southeast Asia', *LINUS @ NUS Libraries*, 18 May. Available at: https://blog.nus.edu.sg/linus/2017/05/18/blsea-the-beauties-and-beasts-of-southeast-asia/ [Accessed: 26 March 2026]

NUS Libraries (2018) 'Introducing libmaps.nus.edu.sg', *NUS Libraries News*, 8 May. Available at: https://nus.edu.sg/nuslibraries/whats-on-listing/news/2018/05/08/introducing-libmaps-nus-edu-sg [Accessed: 22 March 2026]

NUS Libraries (2020) 'Digital Gems', *LINUS @ NUS Libraries*, 21 September. Available at: https://blog.nus.edu.sg/linus/2020/09/21/digital-gems/ [Accessed: 24 March 2026]

NUS Libraries (2024) '200-Year-Old Bugis-Makassar Manuscript', *LINUS @ NUS Libraries*, 11 January. Available at: https://blog.nus.edu.sg/linus/2024/01/11/200-year-old-bugis-makassar-manuscript/ [Accessed: 29 March 2026]

Sim, C. P. (2025) 'Unlocking Regional Insights: The Chinese Southeast Asia Collection', *LINUS @ NUS Libraries*, 25 April. Available at: https://blog.nus.edu.sg/linus/2025/04/25/unlocking-regional-insights-the-chinese-southeast-asia-collection/ [Accessed: 29 March 2026]

Sim, C. P., Toh, M. J. G. and Lin, A. (2025) 'Preserving a Cultural Icon: The Wang Sha Private Papers Collection', *LINUS @ NUS Libraries*, 1 September. Available at: https://blog.nus.edu.sg/linus/2025/09/01/preserving-a-cultural-icon-the-wang-sha-private-papers-collection/ [Accessed: 24 March 2026]

Sim, C. P., Wong, M., Wu S., Diyana, N. and Chow, C. K. (2024) 'Late Philanthropist Yeap Chor Ee's Private Papers Newly Added to Our Special Collection', *LINUS @ NUS Libraries*, 19 August. Available at: https://blog.nus.edu.sg/linus/2024/08/19/late-philanthropist-yeap-chor-ees-private-papers-newly-added-to-our-special-collections/ [Accessed: 23 March 2026]

Zhang, X. P. and Xin, M. (2025) '王沙逾300件珍贵物件捐赠国大公开展出 [Over 300 rare artefacts from Wang Sha donated to NUS, now on public display]', *Lianhe Zaobao*, 1 September. Available at: https://www.zaobao.com.sg/entertainment/story20250901-7442243 [Accessed: 9 November 2025]

Zhong, Y. L. (2025) '87岁秦淮宣布把个人物件赠予国大图书馆 [Chin Whai, 87, donates personal items to NUS Libraries]', *Lianhe Zaobao*, 4 September. Available at: https://www.zaobao.com.sg/entertainment/story20250904-7461054 [Accessed: 1 April 2026]

Metadata & Discovery Group

Linked data and OCLC Meridian

a collaboration between OCLC and the Metadata and Discovery Team at the University of Southampton Library

Sarah Glaccum

Discovery Manager, University of Southampton Library

Received: 30 July 2026 | Published: 16 September 2026

ABSTRACT

In early 2025 the Metadata and Discovery Team at the University of Southampton Library embarked on a project in partnership with OCLC, engaging with their Meridian linked data entity management system. This article describes the project from conception to completion and assesses the benefits of linked data both for the library community and for users.

KEYWORDS linked data; entity management; OCLC Meridian

CONTACT Sarah Glaccum seg1@soton.ac.uk University of Southampton Library

Introduction

As metadata specialists, it will not have escaped any of us that the metadata landscape is changing. More recently, the narrative of change has become more assertive. Those of us with long careers behind us are still absorbing the idea that "MARC must die" (Tennant, 2002). Despite the riposte that BIBFRAME must also die (Edmunds, 2024), it appears likely that the future of metadata lies in BIBFRAME and linked data, with an opportunity, finally, to realise the full benefits of RDA.

Against this background, our interest in actively exploring metadata futures was piqued in early 2025 by a combination of circumstances: attendance at an OCLC Webinar in which the team at West Virginia University Libraries demonstrated a shift from a focus on NACO (Name Authority Cooperative Program) records towards a relationship with OCLC and their linked data entity management system, Meridian (Fidelman and Washington, 2025); a presentation from Yale University Library, facilitated by OCLC, showcasing their Lux catalogue, and demonstrating the potential of linked data to revolutionise the user experience and the way in which we perceive collections (Sanderson and Thompson, 2025); and, finally, an attempt by the team at Southampton to complete NACO training, and our subsequent withdrawal when we realised that the required level of commitment didn't align with the workload of a small team.

As a team, we became curious and started to ask questions about the future of metadata.

As users of OCLC WMS, we had experienced the frustration of not being able to edit PCC (Program for Cooperative Cataloging) records or add new authorised 1xx/7xx headings. Having to request changes via OCLC gave us a sense of cataloguing FOMO (Fear Of Missing Out). Were we being left behind in the skills race by not completing NACO training?

Now, however, we started to ask if there will be space in the new linked data ecosystem for traditional authority control and, by association, NACO training.

At the same time, we became interested in exploring some of the issues raised by Richard Urban in his article, *Identity management beyond the LC/NACO Authority File*, namely, whether identity management platforms could achieve savings in terms of staff training and time, whilst being sufficiently reliable to rival LC/NACO authorities (Urban, 2023).

That FOMO we had about authority management began to shift to a fear of getting left behind in helping to shape the future of linked metadata.

Finally, the decision by OCLC to add entities to 1xx and 7xx fields in WMS/WorldCat records added to the sense that the 'flat' structure of MARC was on the way out in favour of a more connected metadata ecosystem designed for the semantic web. Was "Linky MARC" the first evolutionary step towards a future with linked data and, ultimately, BIBFRAME?

In this context a tentative idea emerged: Would OCLC grant us temporary access to their Meridian linked data entity management system to assess its functionality, test the impact on the user experience, present our findings and build expertise within the team?

Project planning and objectives

Our initial excitement when OCLC agreed to allow us access to Meridian for a month to work on (as yet undetermined) collections at Southampton was tempered by the realisation that we now had a defined time to build on our limited expertise in linked data and upskill the team.

Accordingly, there followed an intense period of training, underpinned by excellent support from OCLC, facilitated by the team in the UK and delivered by Anne Washington and the team in the US. This was supplemented by engaging with webinars, sharing articles and making use of our regular Cataloguing team meetings to inform and upskill staff working with metadata across three service areas.

The decision to work collaboratively across sites and teams is what led to the identification of two unique and distinctive collections that could benefit from the addition of linked data entities:

- a set of Visionaires (nominated by the team at Winchester School of Art Library), and
- a collection of Holocaust memoires, a subset of the Parkes Library for Jewish studies (identified by the Special Collections Team at Hartley Library).

More staff resource was assigned to the latter collection owing to its size and physical proximity to the main metadata team. As the bulk of the work was in creating and curating entities for the Parkes Library, we have also made it the main focus for this article.

Working across teams had the added benefit of enabling us to share the workload, maximise efficiency, and to make the most of specific expertise and subject knowledge.

In the course of our preliminary work, we synthesised the aims of the project to an assessment of the following:

- Linked data entities as a viable alternative to LC/NACO name authorities, simplifying the creation, curation and maintenance of authoritative data.
- The potential of linked data to reduce technical and training overheads without compromising on quality.
- The impact of linked data on the user experience and discovery.
- The functionality of Meridian as a product for managing linked data (feeding back findings to OCLC).
- The speed and ease with which the team could be upskilled to work with linked data entities, and the impact of entity management on efficiency and cataloguing workflows.

Methodology

With access to Meridian limited to one month we set out to maximise input opportunities by collecting and collating relevant data in advance of system switch-on. Accordingly, workflows were designed around entering entity data into a (substantial) spreadsheet, pending transfer to the system. This had the added benefit of removing the pressure to gather data and input it simultaneously, whilst also trying to get to grips with a new system.

We knew that the spreadsheet design was key to the success of the project. We also knew that it would have to be iterative, with scope for adding further columns or worksheets as we discovered the need for additional data. For example, new worksheets were later adopted for recording data about concentration camps and

ghettos, and an additional sheet was created to record related persons and organisations that did not feature in the data download, cross-referenced with works.

Core bibliographic data was exported from WMS into a spreadsheet (1xx, 7xx, 245, 264, plus call number data and barcode from the local holdings record). Authors, editors and organisations were then disaggregated to create a line per potential entity. Duplicates were flagged, to prevent double entry. Columns were added to reflect the content of OCLC entities – persons, organisations, birth and death dates and places, occupations, and relationships (with persons organisations, places, and works).

The final piece of 'pre-work' was a baseline assessment of existing OCLC entities for quality and content, which enabled us to categorise entities in the spreadsheet by fullness, identify duplicate entries, prioritise persons lacking an entity, and paste URIs into the spreadsheet for speed/ease of access.

Data collection and collation now began in earnest. Materials were collected in batches from their home in Special Collections and issued to the Metadata Team. Each member of the team was assigned a section of the spreadsheet. The initial objective was to populate the spreadsheet for as many entities as possible in the period leading up to Meridian switch-on. Data would be pasted directly from the spreadsheet into entity records. Any outstanding entities would be completed 'live', so that a comparison could be made between the two methods. With additional time, we could also have tested automated data ingestion from the spreadsheet to Meridian using an API.

Alongside extensive project planning, the team engaged with OCLC’s excellent training, which not only included system demonstrations but a number of Q&A sessions and regular liaison with Anne Washington and the team in the US.

Challenges around data collection

At this point, it would be great to say that we each quietly perused our collection of books and populated the spreadsheet with data, but this would not do justice to the process we adopted. In many cases, data was not readily available in the usual places we would look – front and back covers, title page and verso, contents pages – resulting in time consuming skim reading of materials. Frequently, key data was absent, which necessitated lengthy research online, making use of registries such as Wikidata, ISNI (International Standard Name Identifier) and VIAF (Virtual International Authority File), as well as Library of Congress name authorities, Wikipedia and the Yad Vashem Central Database of Shoah Victims' Names (for the collection of Holocaust memoirs). This was supplemented by other credible sources, including obituaries, newspaper articles, TV interviews, and exhibition catalogues and galleries (for the Visionaires collection). One of the team admitted to watching an online sermon presented by a particularly elusive author, in the hope of finding out more about him!

Our quest for authoritative data was often frustrating, and was most definitely time-consuming, owing to the number and variety of sources that were available, the discovery of conflicting data across different registries, and a lack of data quality and consistency. Through an iterative process, we finally opted for primary reliance on the item in hand, verified with external sources online as necessary.

We also faced decisions around the granularity of data input: how much data was the right amount? How could we strike a balance between sufficient data to disambiguate, and the time required to obtain information? The team was spurred on by the data fullness percentage score displayed at the top of each entity – the quest for the highest score led to some rather unhealthy rivalry.

Then there was what we would describe as "spider webs" of data – overlaps, intersections, relationships and connections. These are the very things that linked data is designed to surface, but we found ourselves having to create entities to add to other entities, which created a circular need for additional data.

Finally, we are a small team and, whilst we benefited from the assistance of staff from other areas of the service, there was nonetheless a tension between the project and the need to maintain business as usual. We were also aware of the time-limited nature of the project, with an urgency to complete as many entities as possible before access to Meridian was switched off.

Meridian

By the time access to Meridian was turned on in February 2026, we were in a position to input a substantial amount of data from the spreadsheet.

Meridian itself is well designed, generally intuitive, and easy to use. With practice, data input can be completed with reasonable speed, facilitated by drop-down menus and search boxes. We discovered a small number of glitches (for example, date format was designed to suit neither the US nor UK market) but the team at OCLC was commendably responsive, noting our feedback and allocating the staff resources to resolve any concerns.

The use of free text for entity descriptions led to some concern about consistency of data input, and there was initial confusion over how to categorise and describe concentration camps and ghettos. The team mitigated these by devising in-house data input standards and templates and documenting key decisions. Additionally, standards and data fullness across existing entities varied widely, and it appears that this was largely due to the source of the data in WorldCat Entities (data was obtained primarily from VIAF). If we were to move forward with permanent access to Meridian, we would determine minimum record content standards. The recently convened PCC Entity Management Cooperative program (EMCO) will draw on cataloguing community

expertise to address issues of data consistency and content across a number of registries including OCLC WorldCat Entities.

And what of the potential impact of this project on our users?

The immediate benefit is in the shift from string searching within the discovery layer to URI-driven discovery, which ensures greater accuracy. With new capacity to click on URIs to retrieve works by and about authors/editors, users are guaranteed an improved search experience. OCLC plans to extend this by making more of the data in entity records available to users in the form of a knowledge card that is visible in the discovery layer.

With this in mind, we would suggest that there is a particularly strong use case for entities for special and specialist collections. We can foresee a future where a single catalogue search could reveal complex, contextualised relationships across multiple library collections, offering users a rich discovery experience.

Entities transform catalogues into dynamic interconnected resources. This project brought to our attention some remarkable people, all of whom now have improved or new entities that reflect their lives, experiences, works and connections, which will be more readily surfaced as a result of our work.

Project conclusions

This project gave ample opportunity to assess the viability of Meridian as a product for the management of linked data entities, and we conclude that it is fit for purpose. Entity records are well structured and allow for rich data entry. URI generation is immediate. Meridian exceeded our expectations in terms of ease of use, requiring minimal training to acquire speed and accuracy of data input. In this sense, working with OCLC entities is a more inclusive process than engaging with LC/NACO authorities.

However, any gains in efficiency must be balanced against the time-consuming process of data acquisition. It is likely that this is no more impactful than the work required for LC/NACO records, but (for our small team at least) would necessitate selective creation and editing of entities. Should we choose to work with an entity management system in future, entities for special and specialist collections will be prioritised over those for standard reading list cataloguing – focusing work where it can be most impactful in surfacing collections and enhancing the user experience in an enriched discovery landscape.

In a linked data ecosystem, there is less impetus to train as PCC cataloguers, although work is necessary within the community to ensure the application of agreed standards and consistency in the management of entities. It is welcome news that this will be the work of EMCO going forward.

Through engagement with Meridian and OCLC Entities, the team has acquired new technical skills and an improved understanding of the potential benefits of linked data. We recently joined EMCO and are engaging with sector-wide projects relating to BIBFRAME and open linked data, taking us beyond the original objectives of the project.

We know that we will be working in a hybrid environment for the near future at least, but the next exciting step for us is the development of a broader strategic plan to build capacity for a future fundamental shift from MARC to BIBFRAME.

And this takes us back to that initial FOMO. Post-project, far from feeling that we are being left behind, we now feel part of part of something new, with the potential to facilitate fundamental changes to the metadata of the future and, in so doing, to revolutionise the user experience.

References

Edmunds, J. (2024) *BIBFRAME must die, Part III: A Brief History of the Future of Cataloging*. Available at: https://doi.org/10.26207/zn7t-0y56 [Accessed: 17 July 2026]

Fidelman, E., Washington, A. (2025) 'The Mothman properties: linked data for West Virginia Entities', *OCLC Cataloguing Community meeting*, 12 February 2025. Available at: https://community.oclc.org/t5/cataloging-and-metadata/oclc-cataloging-community-meeting-february-2025/ta-p/56869 (Members only) [Accessed: 17 July 2026]

James, K. (2025) 'Exploring AI uses in archives and special collections: integration, entities and addressing need', *Hanging together*, 14 October. Available at: https://hangingtogether.org/exploring-ai-uses-in-archives-and-special-collections-integration-entities-and-addressing-need/ [Accessed: 17 July 2026]

Mixter, J. (2024a) 'Introduction to linked data', *Next*, 26 March. Available at: https://blog.oclc.org/next/intro-to-library-linked-data/ [Accessed 14 July 2026]

Mixter, J. (2024b) 'How linked data can help libraries make more of an impact online', *Next*, 30 April. Available at: https://blog.oclc.org/next/linked-data-libraries-online-impact [Accessed: 14 July 2026]

OCLC (2024) *Linked data: the future of library cataloguing*. Available at: https://doi.org/10.25333/71w4-vq71 [Accessed: 17 July 2026]

OCLC (2026) *Meridian*. Available at: https://www.oclc.org/en/meridian.html [Accessed: 17 July 2026]

OCLC (2026) *WorldCat Entities*. Available at: https://entities.oclc.org/worldcat/entity [Accessed: 14 July 2026]

Sanderson, R., Thompson, T. (2025) 'E pluribus unum: aligning metadata across libraries, archives and museums at Yale', *OCLC Cataloguing Community meeting*, 18 June 2025. Available at: https://community.oclc.org/t5/cataloging-and-metadata/oclc-cataloging-community-meeting-june-2025/ta-p/58886 (Members only) [Accessed: 17 July 2026]

Saur-Games, M. (2024) 'OCLC's role in advancing library linked data', *Next*, 16 July. Available at: https://blog.oclc.org/next/oclc-advancing-linked-data/ [Accessed: 14 July 2026]

Tennant, R. (2002) 'MARC Must Die', *Library Journal*, October 15, pp. 26–27. Available at: https://www.libraryjournal.com/story/marc-must-die [Accessed: 14 July 2026]

Urban, R. (2023) 'Identity management beyond the LC/NACO Authority File', *Hanging together*, May 23. Available at: https://hangingtogether.org/identity-management-beyond-the-lc-naco-authority-file/ [Accessed: 17 July 2026]

Metadata & Discovery Group

Collaborative cataloguing now!

Amy Staniforth 0000-0001-7601-8133
Aberystwyth University & Freelance

Received: 28 August 2026 | Published: 16 September 2026

ABSTRACT

This article reflects on the "Collaborative Cataloguing Now!" workshop held at CILIP MDG's March 2026 conference held in Bristol.

The workshop aimed to both enhance the discovery of resources at a library that is delivering services without a cataloguer, and to test our professional capacity to create good quality bibliographic metadata, without specialist tools.

Together, we catalogued resources for the publicly accessible Drawing Room Library in Bermondsey, London and these can be found in the online catalogue. This was nice to do but, it makes only a small dent in the library's cataloguing to-do list. This article suggests, however, that the potential takeaways lie not so much in the results of this small intervention, but in its implications.

We catalogued these books using laptops, digital surrogates and access to a shared document drive. We created high-quality structured metadata that was easily imported into the Drawing Room's library management system.

In a context of significant change in our field – with Official RDA and Bibframe developments looming – this article asks us to strip back our understanding of bibliographic description. It argues that wherever we do it and whichever tools we use, we bring the experience and expertise.

KEYWORDS collaborative cataloguing; cataloguing expertise; source of bibliographic metadata

CONTACT Amy Staniforth mws@aber.ac.uk Aberystwyth University

The library

I was working in a small, well-lit art library with someone else's laptop and a pile of exhibition catalogues. The latter were diverse in age, size, shape and style. I was there to catalogue 500 of them into a basic library management system with no Z39.50 access.

I created a detailed MARC template and together with the gallery's librarian we came up with approaches to some of the specialist fields. The librarian is responsible for the whole service and she is always busy, running projects that support the exhibition programming and local communities. She has little time for cataloguing.

Figure 1: Drawing Room Library

A successful Paul Melllon funding bid[1] brought me (on a freelance contract) to Drawing Room's library to catalogue these items for improved discoverability and to develop guidance for volunteers to continue the work (see Figure 2 and Figure 3 for examples).

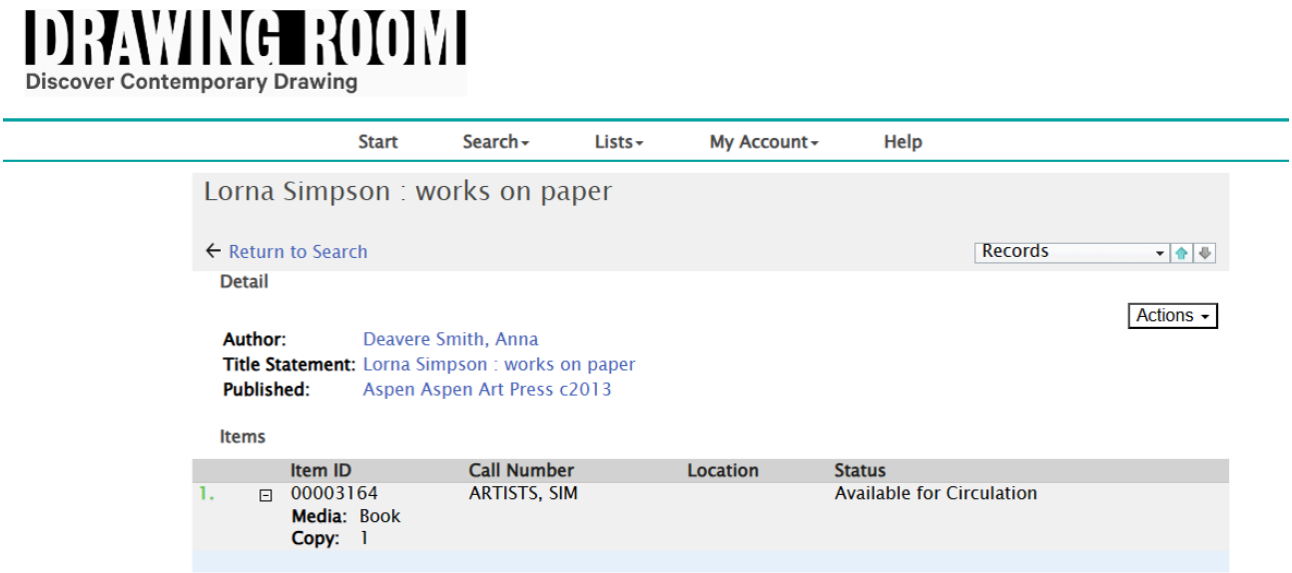


Figure 2: Example of basic metadata in user interface

I am used to cataloguing within a system that allows me to search for, and import, metadata and to check controlled vocabularies easily. Doing this manually was like stretching time. I'd find resources in Library Hub Discover and copy bits of metadata from different records where it proved useful. I typed a lot. I reformatted a lot. I

[1] For more information on the Paul Mellon funded cataloguing project see https://drawingroom.org.uk/event/exhibitions-publications-cataloguing-project/.

transcribed a lot. I searched through gallery websites and the art press for details. I used the open access end of Library of Congress Authorities for subjects and names of people, places, venues, publishers, prizes and events. Although most of the work was online it felt strikingly manual, clicking from site to database to system and back.

After a few weeks it struck me that I was enjoying myself. At times, I felt like I was combining the best of several records with information from the item-in-hand and relevant online sources… to make super records!

Figure 3: Example of enriched metadata in user interface

I felt privileged to have the time to find names, language information, and exhibition descriptions to enrich the records. It is satisfying to know that this metadata is owned by this library and will be openly available should they one day sign up to Jisc's National Bibliographic Knowledgebase.

It was manual and yet also somehow empowering. I realised I could get everything I needed – albeit slowly – from MARC21, LCSH, Library Hub Discover, Getty Thesaurus and the many art organisations and publications out there online.

On my train home to Wales each week I would idly wonder whether other cataloguers would enjoy this work. How long would it take us to catalogue the whole collection – about 5000 items – under a free-to-use licence?

This is where the idea for the Collaborative Cataloguing Now workshop came from. The workshop formed an interactive session at the CILIP Metadata and Discovery Group conference in March in Bristol where 11 delegates joined me for a cataloguer's 'busman's holiday'.

The Workshop

I had created packs of images, templates and guidance for Drawing Room's Library items in need of cataloguing. I numbered these packs on a Google Documents site that delegates could access from their laptops (see Figure 4 and Figure 5).

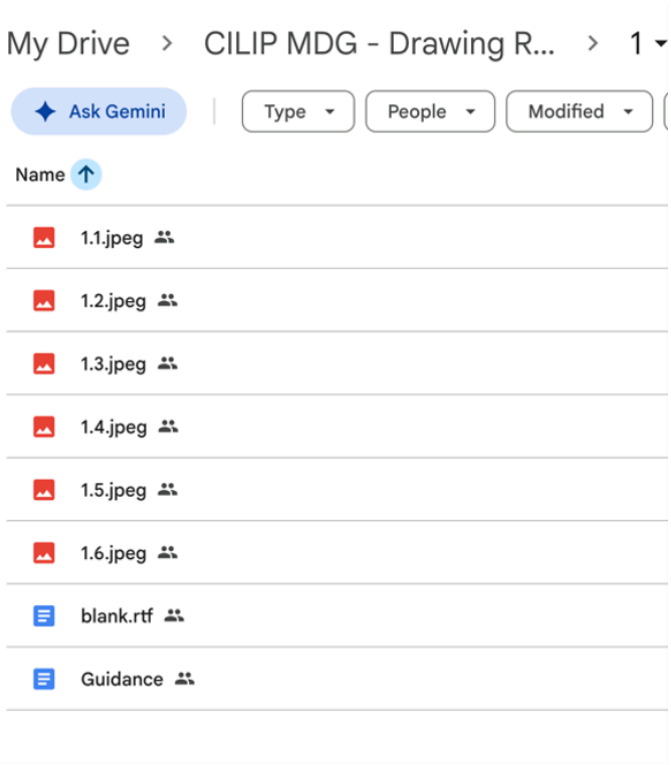


Figure 4: Pack 1 folder contents

Cataloguers worked together on each item in groups and created a MARC record in an .rtf (Rich Text Format) document. I later combined these in MarcEdit and sent the file to Mike, the library's LMS account manager, and he uploaded them to the catalogue for discovery.

We had an hour to do something quite unusual with people we didn't know. I'm grateful that everyone braced themselves and just went for it in spite of the strangeness, the technology making things tricky, and the brief time to get familiar with the task.

It was great to create records for Drawing Room (see below for examples to find in the catalogue) but the ultimate aim of the workshop was to test the idea of collaborative cataloguing itself and to see what other cataloguers think about it. The "real" results are best communicated via the answers to the below questions that I asked the delegates, in an email after the workshop.

```
=001
=003  UkLoDR
=008  xxxxxxsyyyy\\\\enka\\\\\cb\\\001\0\eng\d
=020  \\$a $q
=040  \\$aUkLoDR $beng $cUkLoDR $erda
=041  \\$aeng
=043  \\$a
=1xx  x\$a $e
=240  0\$a
=245  10$a : $b / $c.
=246  1x$a
=264  \1$a : $b , $c.
=300  \\$a : $b ; $ccm.
=336  \\$atext $btxt $2rdacontent
=336  \\$astill image $bsti $2rdacontent
=337  \\$aunmediated $bn $2rdamedia
=338  \\$avolume $bnc $2rdacarrier
=500  \\$aPublished to accompany the exhibition [x] held at [x], [place], [month
day-month-day, year].
=505  0\$a
=520  \\$a
=546  \\$aIn English.
=541  \\$a
=561  \\$a
=600  10$a $vExhibitions.
=650  \0$a
=651  \0$a
=655  \7$aExhibition catalogs $2aat.
=700  1\$a
=710  2\$a
```

Figure 5: Blank .rtf working template

Q.1 Were you surprised that you could do this work in such unusual circumstances? How might you feel if you were to do it again?

- No – it was fine. How might I feel if I were to do it again? Again – fine.
- Not really surprised but reminded that we can catalogue effectively with the knowledge we already have and fairly simple and free resources.
- I was surprised at how much I couldn't remember!
- Not particularly surprised – it was fine using scans to establish the information.
- I was surprised at how easy it was to start cataloguing together with a person whom I've never talked to nor met before. Although I am an English AFL [as foreign language, eds.] speaker, [MARC] is universal ;)
- The two books I was working on are actually available in my library. So, I was comfortable in cataloguing them. It was a bit slow because I "needed" to refer to the pictures in the folder. I am happy to do it again.

Q.2 What was it like seeing MARC fields in the .rtf document?

- Slightly strange at first but quickly got used to it.
- This was fine. Very like MarcEdit and easy to pick up the right formatting and spacing.
- I don't use MARC at all so it was mainly confusing; but the .rtf format wasn't the reason!
- Like being at home.
- Seeing [MARC] in the .rtf document did not pose a problem.
- I don't have experience with .rtf documents but it makes sense to me.

Q.3 You have contributed your skills to the catalogue of another organisation. How does that feel?

- Co-operative.
- It's similar to doing original cataloguing for my organisation and sharing the record – contributing something that will be useful to other organisations.
- Brilliant!
- Again, ultimately fine – but I did have a tiny tiny bit of ethical unease. [Presumably for copying and pasting, author's note] That said, it was curiously satisfying.
- Very good – it's nice to contribute to make data available when there isn't any!
- It was really exciting, and I wish that we would have had more time to do this at the conference. But continuing doing similar work at my ordinary work place would be a privilege.

Q.4 How has the experience made you reflect on any day-job cataloguing you do?

- Just that I wish that I [was] doing this kind of cataloguing as a part of my daily cataloguing routine.
- It hasn't.
- I feel lucky that we have Alma here [at own institution] so that we can download records from WorldCat or other libraries.
- I don't catalogue in my role; but I did reflect on how this could be made a public event to catalogue material that isn't catalogued.
- I should remember more.
- My day-job is largely managing cataloguing and I rarely actually catalogue individual items anymore. I have to look things up much more than I used to.

Q.5 The data we're contributing to Drawing Room's catalogue is unlicensed so it can be shared freely via any future aggregators (like the National Bibliographic Knowledgebase or via Z39.50). What are your thoughts on this?

- That is the way it should/ought to be.
- I would like to explore this.
- Bloody good!!
- That's the way it should be.
- This is great – I'm very happy for metadata I create to be shared freely (at least for non-profit purposes).
- Great idea – more shared data the better!

Q.6 Do feel free to elaborate on any other thoughts you've had about this exercise.

- Strangely enjoyable (maybe the format made it a bit novel…)
- I think I was the only person in the room who didn't use MARC (which I expected going in), so no doubt that had a detrimental impact on the others on my table! Despite this, I did enjoy taking part and I did learn about MARC a little too!
- The notes on cataloguing exhibition catalogues/art materials were useful because I rarely deal with materials like this. Art materials are an area where I'm aware there's specialist cataloguing knowledge that I don't have… I'd hope that over a longer session I'd get more confident with these materials.
- Prior cataloguing experience in MARC and RDA is really helpful.
- The other cataloguer I was paired up with was very experienced and we went right at and dove into it. Normally I very rarely catalogue together with someone else. Sure, we confer amongst each other on a weekly basis with questions on different details, but to catalogue like this was different. I also enjoyed contributing with a very tiny part in helping the collection becoming more visible and universal. More, more, more…

The source of bibliographic metadata

It should come as no surprise to us that metadata librarians carry with them the tacit knowledge underpinning resource description and discovery. What might this mean when many of us feel far away from where the tools and standards we use are designed?

Bibframe promises a model within which we can share much more easily. We have long had an infrastructure that allows us to share our metadata, however. We have used Z39.50 effectively but our institutional outsourcing has created a marketplace

that needs us all to buy the same metadata to cover costs of creating it. New technology is a tool not a solution.

As a profession we have a long history of collaborating beyond employer borders for the benefit of the sector. We do this at infrastructure, strategy and policy levels. If we recognise that metadata expertise resides in us first, rather than in existing and future technology, could we consider collaborating on a more practical level too? Could we consider ourselves to be the metadata "source" of choice?

I am lucky to sit on the Library Hub Community Advisory Board (LHCAB) where we discuss the UK's repository of shared bibliographic metadata. I'm involved in a Society of College, National and University Libraries (SCONUL) and Research Libraries UK (RLUK) initiative exploring open licencing for metadata. I've volunteered too for the UK subject heading work being led by CILIP's MDG. These are all national initiatives that reflect the combined contribution of many years of colleagues' working lives. This work shows that when we collaborate we shape our sector.

In theory, our acquisition overlap should mean that many new resources could be catalogued once by one of us and the record shared. We have the infrastructure to do this, from scratch, now. A shared cataloguing programme would both develop our experience of sharing and ensure that we have cataloguers left to catalogue in Bibframe when the time comes.

What is stopping us? Is it too much to expect libraries to share acquisition information? Is it too old-fashioned, too "MARC"? Is working out how much metadata we all currently buy in and how much of it we actually need, just too time consuming? All of these barriers are real but none seem insurmountable and none offer justification for not changing how we work. And for now, there are some of us around the country who could do more together than simultaneously oversee the import of the same metadata into each of our library catalogues.

We could team up across sectors and together catalogue the majority of our new material. Perhaps more excitingly, however, we can collaborate on projects, on the materials we know that could and should be more discoverable. We can team up to create super records for all sorts of subjects and resources. I still know little about cataloguing art resources. But thanks to all the good work already out there, I was able to bring together data that makes previously difficult-to-find resources much more discoverable with very little software.

With even less software and infrastructure, conference delegates turned up for an hour and did the same.

What else can we do?

Cataloguing Now books catalogued

The Drawing Room Library catalogue is available at https://drawingroom.org.uk/library-research/library-catalogue/

- **George Condo: ink drawings** / Skarstedt.
- **Lorna Simpson: ink** / [edited by Eve MacSweeney].
- **PissIng ink: 80 pages from the Miriam Books** / by Mathew Hale.
- **Shelf life : Neil Gall** / by Simon Groom, Johanna Malt and Charles Darwent.
- **David Hockney: travels with pen, pencil and ink** / introduction by Edmund Pillsbury.
- **The animal side** / Jean-Christophe Bailly, translated by Catherine Porter.
- **Beyond the line: the art of Diana Cooper** / edited by Margo A. Crutchfield with an interview by Barbara Pollack.
- **The ungentlemanly art : a history of American political cartoons** / Stephen Hess and Milton Kaplan.
- **The art of walking: a field guide** / edited by David Evans.
- **Italian drawing of the 20th century** / edited by Irina Zucca Alessandrelli with Antonello Negri.
- **The stage of drawing: gesture and act ; selected from the Tate collection** / selected by Avis Newman ; curated by Catherine de Zegher ; organized by the Drawing Center, Tate.
- **Based on paper: the Marzona collection : revolution in art, 1960-1975 = die Sammlung Marzona : Revolution der Kunst** / Michael Lailach, Andreas Schalhorn.
- **Contemporary art at Deutsche Bank London** / [editor and essays: Friedhelm Hütte ; translation, David Britt].
- **560 Broadway: a New York drawing collection at work, 1991-2006** / edited by Amy Eshoo.
- **The world on paper: Deutsche Bank collection** / [with texts by Elsy Lahner, Friedhelm Hütte, Katherine Stout and Lothar Muller, and translations by Burke Barrett.]
- **Pencil in hand: 20th century drawings** / Colecciones Fundación Mapfre.
- **The Judith Rothschild Foundation Contemporary Drawings Collection: catalogue raisonné** / [edited by Libby Hruska].
- **Dessins surréalistes: visions et techniques** / [edited by Béatrice Salmon, Viviane Tarenne and Rudolf Velhagen].
- **Donation Florence et Daniel Guerlain: dessins contemporains** / sous la direction de Jonas Storsve.

- **Oeuvres sur papier: acquisitions 1996-2001 : collections du Centre Georges Pompidou, Musée National d'art moderne, Cabinet d'Art graphique** / [catalogue coordinated by Claude Schweisguth].
- **Die Sammlung Ploner: Albertina, Belvedere, Neue Galerie Graz** / edited by Agnes Husslein-Arco, Peter Pakesch and Klaus Albrecht Schröder.

Metadata & Discovery Group

Metadata in motion

building datasets from library catalogues

Fran Frenzel 0009-0009-3890-5335
Metadata Analyst, The London School of Economics and Political Science (LSE) Library

Received: 26 August 2026 | Published: 16 September 2026

ABSTRACT

This article explores how library catalogue metadata can be transformed into computationally useful datasets through the lens of LSE Library's pilot project to publish metadata for its print PhD theses holdings. It outlines the rationale for approaching collections as data as part of digital scholarship support, describes practical decisions around scope, licensing, formats, documentation and publication, and reflects on the substantial work involved in cleaning, selecting and reconciling metadata. Drawing on lessons learned during the creation of the dataset, the article highlights the continuing value of traditional metadata expertise while showing how that expertise can be applied beyond catalogue discovery. It also considers the ethical responsibilities of dataset creation and stewardship, particularly in relation to transparency, bias, sustainability and AI use, and looks ahead to future collections as data work at LSE Library.

KEYWORDS collections as data; metadata datasets

CONTACT Fran Frenzel f.frenzel@lse.ac.uk The London School of Economics and Political Science

Introduction

Libraries have long been in the business of creating structured metadata to enable access to their collections, however increasing interest and activities in digital scholarship prompted them to look at what else library catalogues (and collections themselves) can do when presented as computationally useful datasets for research, teaching and other creative reuse. The idea of "collections as data" reached a wider audience with the "Always Ready Computational: Collections as data" project – running from 2016 to 2018 (Always Ready Computational, 2024) – and its successor "Collections as Data: Part to Whole" – running between 2018 and 2023 (Collections as Data, 2023) and libraries started experimenting with turning metadata and collections into datasets[1].

The Metadata Team at LSE is part of the Library's Digital Scholarship & Innovation group, whose mandate is to develop LSE's digital services and explore how the Library

[1] See e.g. the National Library of Scotland's Data Foundry (https://data.nls.uk/), the Library of Congress' LC Labs (https://labs.loc.gov/), datasets made available by the British Library (https://doi.org/10.23636/jyxw-xa90) and the German National Library (https://www.dnb.de/librarylabpractice).

can support research, learning and teaching in new ways in a digital environment. Thus, we set out to explore collections as data as a new offering within the Library's digital scholarship activities[2].

This article gives an overview of what Collections as Data means, describes our journey and lessons learned so far as well as looks ahead at our plans.

Can we do this?

Before launching a dataset service, we needed to look at the remit, workloads and skillsets within the team to assess our ability to deliver this. Changes within the team allowed us to re-purpose an existing role to a newly named Metadata Analyst post. This role includes exploring and establishing workflows for collections as data work. Next, we needed a trial dataset to experiment with the process, gain practical experience and use that work to inform future service development.

We chose the metadata for LSE print PhD theses holdings, because:

- it complemented existing Wikidata work we have done and are doing for LSE electronic PhD and MPhil theses[3],
- it only involved metadata created and owned by LSE, i.e. there would be no potential metadata copyright issues to navigate,
- it is of a manageable size for an experimental project (approximately 6,000 records).

Our objective was not to create a perfect dataset but to gain practical experience and develop a better understanding of the workflows, skills and infrastructure required to support a future dataset service.

What is "Collections as Data"?

In the final report of the Always Ready Computational: Collections as Data project the idea is described as

> "a conceptual orientation to collections that renders them as ordered information, stored digitally, that are inherently amenable to computation." (Padilla et al., 2019a, p. 7)

In plainer words: provide digital objects (digitised or born digital) and/or their metadata in a way that can be directly used by researchers, educators, learners or the public for computational processing such as text mining, machine learning, data visualisation and analysis.

[2] For more information on LSE Library's digital scholarship offerings see https://www.lse.ac.uk/library/research-support/digital-scholarship

[3] Helen has written about this work in various places: Williams, 2021, Williams, 2022, Williams and Elder, 2023, Williams and Elder, 2024, and LSEThesisProject, 2026.

Viewed from this perspective, catalogue metadata is more than infrastructure supporting discovery: it is an asset in its own right, that can be made available outside traditional library systems and support new forms of research.

Publishing a collection as data does not only mean that the data is in a useful, accessible format, it also means providing appropriate documentation and contextualisation. The Santa Barbara Statement (Padilla et al., 2019b) and Vancouver Statement (Padilla et al., 2023) – both outputs of the afore mentioned projects – provide a framework for this.

They emphasise that collections as data should be developed responsibly, acknowledging issues of bias, representation, ownership, access and sustainability. They also stress the importance of documentation, transparency and collaboration with user communities. This is important in the context of building a dataset from metadata as well, as library catalogue metadata is not neutral simply because it is structured. It carries the historical and cultural contexts of the standards, decisions, local practices and systems used to create it.

Building a metadata dataset also needs a little shift in perspective; away from what is best to support discoverability and towards asking how metadata might be reused, analysed, linked, visualised or combined with other datasets.

Building the LSE print PhD theses dataset[4]

Armed with a framework, an export of the theses records in MARC and some ideas gathered while looking at what others did, we set to work.

Initial decisions

Before we started work on the data itself, we made some initial decisions:

- The dataset will be openly accessible and published under a CC0 licence (see Creative Commons, no date).
- We will create a datasheet for documentation of and information about the dataset, that will be published together with it.
- The data will be published in CSV and JSON format[5].
- The dataset can be further developed in the future.

One question we would have liked to answer but could not at the time was where the dataset would be published. There was no precedent and, at the time, LSE did not have a data repository. We discussed options from hosting on GitHub or GitLab as well as third party repositories such as Zenodo or Figshare, to developing a dedicated site. By the time we were ready for publication however, LSE Library had just launched its data

[4] The dataset itself is available at https://doi.org/10.21953/researchonline.lse.ac.uk.00137571.

[5] CSV because it is platform agnostic and simple; JSON because it is better at representing nested structures common in bibliographic data.

repository, and an obvious hosting space presented itself. Once we have enough datasets, we would like to come back to the idea of a dedicated space as part of our website to showcase them.

Lesson 1: Think about where to publish, but don't let that get in the way of starting work. Options for long-term, stable hosting are out there, and one can revise later and implement something different[6].

Lesson 2: Creating a dataset and writing documentation, particularly for the first time, is time-consuming, and difficult to prioritise alongside the more immediate or urgent requirements of day-to-day work. In our case CRIS (Current Research Information System) and repository work supporting the REF and LSE's open research agenda, alongside a university-wide project which included the migration of LSE's repository, competed with time we had hoped to spend on datasets work, and our initial timescales had to be revised accordingly.

Weeding the records

The export from Alma included more records than we wanted for the dataset as LSE also has some holdings of non-LSE PhD theses as well as LSE Master's theses[7]. To find and isolate those in Alma would have needed the kind of detailed analysis I was going to run the records through anyway, so we opted for exporting all holdings and separating them later.

To get an overview of and feel for the data, I ran the MARC through field counts and validation in MarcEdit (Reese, 2026). This threw up a first round of conspicuous records: those with unexpected fields for thesis records, missing authors and invalid fixed-length field values, many of which turned out to be outside the scope of the dataset.

I then moved on to look at note field values, specifically 502 and 500, to identify any remaining out of scope records. This yielded some theses that needed a physical check as crucial data was missing or looked suspicious.

At this stage we also decided to exclude records representing publications submitted for a degree rather than a dedicated thesis[8].

Which data to include?

After we had our record set and filled the most pressing data gaps, Helen and I decided which data to take forward and which to leave behind.

[6] Choosing a hosting solution that allows creation and use of DOIs makes this especially easy and even hosting in more than one space will be fine. Also, using a collaboration space like GitHub and a service like Zenodo in tandem is rather common for research data and code.

[7] They are in the minority and mostly there for historic reasons. LSE Library does not routinely collect them.

[8] Thesis by published works (Wikipedia, 2025) is not current practice as LSE, but historic examples exist.

As mentioned before, the focus was on what is useful to include from a research and analysis perspective rather than a cataloguing and discovery one. E.g.:

- There was some interesting data of theses having been included in reading lists in some records, but they were so few that it was at best not useful and a worst misleading to include it in the dataset.
- We omitted language as this is the same for all our theses and thus of limited value and can be easily added should the dataset be combined with another.
- We decided against including the LSE departments theses were submitted to because the data was too patchy to be genuinely useful and we had no capacity to fill in the gaps and make it useful.
- Though we carried extent forward it did not make it into the current version of the dataset; the 300 data was messy and we decided against it. Our ongoing dataset work is causing us to consider releasing metadata "as is", so we may make a different decision were we embarking on the theses dataset now.

Data tidy

From here on all work was done in OpenRefine (Delpeuch et al., 2026). I exported the fields and subfields we wanted from MarcEdit into CSV and loaded that into OpenRefine[9].

The two areas that needed the most tidying work were titles, subtitles and dates. Usage of $b and $c was not consistent and for some records the years recorded in 008, 260 and 502 disagreed with each other. Where the disagreement presented as two consecutive years, we assumed them relating to the respective academic year and chose the later year for inclusion in the dataset.

Lesson 3: Although library metadata is generally highly structured, legacy records, (changing) local practice and changing descriptive standards mean it can be time-consuming to prepare it for use outside the catalogue environment. We had underestimated the amount of work needed to remedy this. Next time we aim to do more scoping work and be more open to releasing data "as is".

Author reconciliation

Then we got to the author and subject reconciliation and that did take a lot more time and effort than anticipated.

We didn't realise at the time that we could export Library of Congress Linked Data Service[10] URIs from Alma where headings are linked to NACO/SACO authorities. I

[9] It's worth thinking about the structure of this export: for the 245 tag for example one could export every subfield individually, or the whole thing. I overall prefer to go broad and split as part of the tidying work. This ensures no unexpected field usage is missed, and I can use the structure of the whole tag for data extraction.

[10] id.loc.gov

anticipated using MarcEdit's built in heading validation or the Linked Data Service API[11] using OpenRefine's URL fetching functionality, but this proved unworkable due to the amount of missed and erroneous matches[12]. Assessing multiple results from API responses via fetching URLs is also very difficult as there is no selection interface similar to OpenRefine's reconciliation functionality. Unfortunately, using this functionality was not an option as the Linked Data Service has no native SPARQL endpoint[13].

So, we decided to approach this from a different angle, leveraging existing Wikidata work: a significant proportion of LSE print theses have been digitised and added to LSE's repository and Wikidata[3]. Thus, by reconciling by title to those records, I was able to extract the linked author record and from this LCCNs and VIAF identifiers[14].

I then ran the rest of the authors through Wikidata reconciliation using OpenRefine's reconciliation functionality, auto-matching on candidates with high confidence scores. To increase the chances of this yielding the right person, I then fetched some additional data for each person from Wikidata:

- dates of birth and death,
- educated at,
- field of work and occupation, and
- employer.

With this additional data I could check the likelihood of correct matching. E.g. someone born after the theses was written is not the right person; someone being in their 80ies when the theses was written is also unlikely and warrants a closer look; someone who's occupation is recorded as biologist or basketballer is, of course not impossible, but unlikely enough to warrant closer inspection.

After that came the assessment and record selection for authors with more than one match. This and looking into conspicuous records identified from the additional data took a lot of time.

I repeated the process for unmatched authors trying a different name variant, e.g. expanding initials or omitting middle names, until I was confident that I had got everyone I could with the name data at hand. All still unmatched authors we assumed

[11] https://id.loc.gov/views/pages/swagger-api-docs/index.html MarcEdit's validation function uses this API as well.

[12] The API is not the easiest to navigate. If you want to know more about this, here are the notes I wrote up while investigating it: https://meriban.github.io/notes/loc_api_openrefine/.

[13] Jeff Chiu's conciliator extension (Chiu, 2025a) didn't help with this either. One can get Library of Congress authorities via VIAF, but I ran into the same problem with false matches and missed matches. On a related note, if you're interested in using conciliator and need more throughput than Jeff is happy to run via his server (Chiu, 2025b), here's a guide on how to set up your own server for it: https://meriban.github.io/notes/openrefine_conciliator/.

[14] Library of Congress Control Numbers (LCCN), the identifier used for NACO/SACO records in LoC's Linked Data Service URLs, and VIAF identifiers are present in many Wikidata person records.

to not have a Wikidata record. Finally, I pulled in LCCNs and VIAF identifiers for all authors reconciled to Wikidata.

With the remaining unmatched authors, I went back to the Linked Data Service API and ran their names and variants through it. This yielded only a few results, and it was quick to check them for accuracy.

Subject reconciliation

Reconciling the subject headings was more difficult and easier at the same time. Our headings are Library of Congress Subject Headings (LCSH) and the problem was not so much variation, but that not every conceivable subject heading is explicitly established in the Linked Data Service. E.g. a heading is established and a URI available for "Rural population", but not for "Rural population--Japan".

We decided to shorten the subject headings subdivision by subdivision until we had a reconcilable one with a URI. I could not think of a good way of doing this in OpenRefine, and scripted it in Python instead, bringing the results back into the OpenRefine project via a cross-table lookup based on the Alma system record number. This worked well for topical headings (650) and geographical headings, acceptably for personal (600) and corporate headings (610 and 611) and fell over entirely for title headings (630). Upon closer inspection of the latter, it became obvious that all of them were erroneous: they meant to represent a topic, and the cataloguer did not realise that they pulled in a title heading instead.

Of course, there were some headings that did not result in any match. I made some manual corrections, but in cases of the heading not being a currently or previously valid LCSH and/or where a judgement on a valid heading could not be made without consulting the resource, we decided to not interrogate further and drop them from the dataset.

We knew that this decision meant a loss of detail, and to counter this we decided to transform the LCSH into FAST headings to include alongside. OCLC provide a LCSH to FAST converter (OCLC, 2023) which takes MARC formatted text as input and outputs in the same format including $0 IDs that can be parsed into URIs. I scripted input and output processing and integrated the results into the OpenRefine project.

This process was also a bit lossy as the transformation didn't succeed for all LCSH. Again, we decided not to interrogate these further.

Lesson 4: Entity reconciliation is very time-consuming. It is not simply a matter of matching strings to identifiers. Reconciliation requires judgement, review, and an understanding of external vocabularies and authority files, as evidenced by the amount of manual, record-by-record work needed and the problems the pre-coordinated and modular nature of LCSH pose.

Transformation into CSV and JSON

The final transformation from OpenRefine project into CSV and JSON was simple: OpenRefine natively supports export to CSV and, via templating export, to JSON.

Documenting the dataset

In parallel to the data work I investigated dataset documentation approaches that would allow us to follow the Santa Barbara (Padilla et al., 2019b) and Vancouver Statements' (Padilla et al., 2023) recommendations. We needed a solution that would allow for:

- transparency in how, why and by whom the dataset was created,
- communicate absences, biases and/or other concerns regarding the data (e.g. sensitive information, representativeness), and
- state intended use, dependencies, risks and retention policies.

I looked at various templates and approaches. I found Microsoft's *Aether Data Documentation Template* (Microsoft's Aether Transparency Working Group, 2022), and its origin: the *Datasheets for Datasets* paper (Gebru et al., 2021), as well as Google's *Data Cards Playbook* (Pushkarna et al., 2022) particularly useful[15]. At first glance these might seem over the top for a dataset of theses metadata, but the intention was to develop a template we can reuse and build on for other projects. Some concepts and ideas are not needed or not relevant in our context, but the datasheet we published takes inspiration from each of the three as well as the Santa Barbara and Vancouver Statements.

Lesson 5: It pays to work on the documentation in parallel to the data work. It's easy to forget why certain decisions were made or what exactly one did. Writing it down there and then saves a lot of head-scratching and information hunting later.

AI, ethics and responsible stewardship

The project coincided with increasing interest in AI across libraries and higher education. As dataset creators and stewards we must think about how our data may be consumed by AI systems, how context may be lost, how biases may be amplified and how rights and responsibilities should be communicated to users.

While the encouragement and support of "responsible computational use of digitized and born digital collection" (Padilla et al., 2023, p. 2), ethical considerations such as "commitments to work against historic and contemporary inequities represented in collection scope, description, access, and use" (Padilla et al., 2019b, p. 3), and efforts towards lowering barriers to use, interoperability and transparency are

[15] Microsoft has also made a tool for datasheet generation available based on work by Roman et al. (2024): https://microsoft.github.io/opendatasheets

part of both the Santa Barbara and Vancouver Statement, the Vancouver Statement introduced some new principles clearly related to the rise of AI:

- sustainability and the consideration of the climate impact of collections as data work,
- the consideration of exploitative labour, and
- the consideration of "the ethical implications, including intellectual property and claims to data sovereignty" of data consumption by AI and a call to "develop adequate safeguards against improper usage, detrimental loss of context, and the amplification of biases through these technologies". (Padilla et al., 2023, p. 4)

In the current AI[16] hype it is tempting to position AI as a solution to labour-intensive processes such as entity reconciliation, but we must not lose sight or become blind to other, less environmentally detrimental and less ethically dubious technologies presenting an equally good or better approach. Just because we *can* use AI, must not mean that we *do* so indiscriminately and preferentially.

Metadata practitioners have long engaged with questions of authority, representation, bias, provenance and context – all ethical questions AI amplifies – and are thus in a good position to carry this work forward into collections as data work and our role as stewards. Collections as data work cannot be separated from questions of ethics, sustainability and responsible use.

Outlook and conclusion

The next phase of work is connected to The Women's Library[17]. A request from the curatorial team prompted the idea of compiling a metadata dataset of holdings published between 1920-1950. This offers us an opportunity to apply the learning from the print theses dataset to a flagship collection with significant research, teaching and public engagement potential. It is also a step in the right direction towards further dataset work arising from collaborations, researcher and user needs.

Apart from creating a useful dataset, we are also looking to involve the Metadata Team as a whole in the work. While there are aspects of the work that require advanced technical skills and familiarity with linked data, much of it is not that different from the work the team already do and many of the necessary skills already exist within metadata teams: authority control, quality assessment, data normalisation and workflow documentation as well as user-centred thinking and ethical reflection. This shows that, far from losing relevance, traditional metadata expertise has relevance far beyond the catalogue.

[16] And by AI, I mean mostly generative AI.
[17] https://www.lse.ac.uk/library/collection-highlights/the-womens-library

We are also hoping to be able to experiment further with machine learning and AI to help with the most time-consuming, manual tasks such as entity reconciliation. Though as stated before, this possibility should be approached critically.

The LSE print PhD theses dataset shows how a modest pilot can help a metadata team test capacity, develop skills and move towards a dataset service and that metadata teams can impact and support digital scholarship.

AI statement

I used Microsoft CoPilot to help me write this article, specifically to improve clarity and correct spelling.

References

Always Already Computational (2024) Available at: https://collectionsasdata.github.io/ [Accessed: 18 August 2026]

British Library (no date) *Researcher Format Datasets from Collection Metadata*. Available at: https://doi.org/10.23636/jyxw-xa90 [Accessed: 18 August 2026]

Chiu, J. (2025b) *Reconciliation Service for OpenRefine*. Available at: https://refine.codefork.com/ [Accessed: 18 August 2026]

Collections as Data (2023) Available at: https://collectionsasdata.github.io/part2whole/ [Accessed: 18 August 2026]

Creative Commons (no date) *CC0 1.0 Universal (CC0)*. Available at: https://creativecommons.org/publicdomain/zero/1.0/ [Accessed: 18 August 2026]

Deutsche National Bibliothek (2024) *DNBLab: Our data in practise*. Available at: https://www.dnb.de/librarylabpractice [Accessed: 18 August 2026]

Gebru, T. et al. (2021) *Datasheets for Datasets*. ArXiv. Available at: https://doi.org/10.48550/arXiv.1803.09010 [Accessed: 3 September 2026]

Frenzel, F. (2026a) *How to use the conciliator extension for OpenRefine and run your own server for it*. Available at: https://meriban.github.io/notes/openrefine_conciliator/

Frenzel, F. (2026b) *Using the id.loc.gov API for entity reconciliation or information retrieval in OpenRefine*. Available at: https://meriban.github.io/notes/loc_api_openrefine/

Frenzel, F. and Williams, H. K. R. (2026) 'LSE Print PhD theses 1905-2011'. Available at: https://doi.org/10.21953/researchonline.lse.ac.uk.00137571

Library of Congress (no date) *Labs*. Available at: https://labs.loc.gov/ [Accessed: 18 August 2026]

LSEThesisProject (2026) Available at: https://www.wikidata.org/wiki/Wikidata:WikiProject_LSEThesisProject [Accessed: 18 August 2026]

LSE Library (2026a) *Digital scholarship*. Available at: https://www.lse.ac.uk/library/research-support/digital-scholarship [Accessed: 18 August 2026]

LSE Library (2026b) *The Women's Library*. Available at: https://www.lse.ac.uk/library/collection-highlights/the-womens-library [Accessed: 18 August 2026]

Microsoft (no date) *Data Documentation*. Available at: https://web.archive.org/web/20260214043105/https://www.microsoft.com/en-us/research/project/datasheets-for-datasets/ [Accessed: 18 August 2026]

Microsoft's Aether Transparency Working Group (2022) *Aether Data Documentation Template (DRAFT 08/25/2022)*. Available at: https://www.microsoft.com/en-us/research/wp-content/uploads/2022/07/aether-datadoc-082522.pdf [Accessed: 18 August 2026]

National Library of Scotland (no date) *Data Foundry: Open data collections from the National Library of Scotland*. Available at: https://data.nls.uk/ [Accessed: 18 August 2026]

Padilla, T. et al. (2019a) *Always Already Computational: Collections as Data. Final Report*. Available at: https://doi.org/10.5281/zenodo.3152934 [Accessed: 18 August 2026]

Padilla, T. et al. (2019b) *Santa Barbara Statement on Collections as Data*. Available at: https://doi.org/10.5281/zenodo.3066209 [Accessed: 18 August 2026]

Padilla, T. et al. (2023) *Vancouver Statement on Collections as Data*. Available at: https://doi.org/10.5281/zenodo.8341519 [Accessed: 18 August 2026]

Padilla, T., Scates Kettler, H. and Shorish, Y. (2023) *Collections as Data: Part to Whole. Final report*. Available at: https://zenodo.org/records/10161976 [Accessed: 18 August 2026]

Roman, A. C. et al. (2024) *Open Datasheets: Machine-readable Documentation for Open Datasets and Responsible AI Assessments*. ArXiv. Available at: https://doi.org/10.48550/arXiv.2312.06153 [Accessed: 3 September 2026]

Wikipedia (2025) *Thesis as a collection of articles*. Available at: https://en.wikipedia.org/wiki/Thesis_as_a_collection_of_articles [Accessed: 18 August 2026]

Williams, H. K. R. (2021) 'Wikidata: what? why? how?', *Catalogue and Index*, 203, pp. 28–35. Available at: https://www.cilip.org.uk/media/zyzfxkln/catalogue_and_index_203.pdf [Accessed: 3 September 2026]

Williams, H. K. R. (2022) 'LSE's adventures in Wikidata-land: tears and triumphs down the rabbit hole'. *Catalogue and Index*, 206, pp. 2–6. Available at: https://www.cilip.org.uk/media/k3mixw30/catalogue_and_index_206.pdf [Accessed: 3 September 2026]

Williams, H. K. R. and Elder, R. (2023) *Wikidata Thesis Toolkit*. Available at: https://www.wikidata.org/wiki/Wikidata:WikiProject_Wikidata_Thesis_Toolkit [Accessed: 18 August 2026]

Williams, H. K. R. and Elder, R. (2024) 'Introducing the Wikidata Thesis Toolkit', *Catalogue and Index*, 208, pp. 26–26. Available at: https://journals.cilip.org.uk/catalogue-and-index/article/view/622 [Accessed: 18 August 2026]

Tools, software and services

Reese, T. (2026) *MarcEdit* [Software]. Available at: https://marcedit.reeset.net/downloads

Chiu, Jeff (2025a) *conciliator* [Software]. Available at: https://github.com/codeforkjeff/conciliator

Delpeuch, A. et al. (2025) *OpenRefine* [Software]. Available at: https://openrefine.org/download, https://github.com/OpenRefine/OpenRefine, https://doi.org/10.5281/zenodo.595996

Library of Congress (no date) *Linked Data Service*. Available at: https://id.loc.gov/

Library of Congress (no date) *Linked Data Service APIs*. Available at: https://id.loc.gov/views/pages/swagger-api-docs/index.html

Microsoft (2023a) *Open Datasheets*. Available at: https://microsoft.github.io/opendatasheets, https://github.com/microsoft/opendatasheets-framework Accessed: 18 August 2026]

OCLC (2023) *FAST Converter* [Software]. Available at: https://fast.oclc.org/lcsh2fast/

Pushkarna, M. et al. (2022) *The Data Cards Playbook*. Available at: https://sites.research.google/datacardsplaybook/, https://github.com/PAIR-code/datacardsplaybook

Metadata & Discovery Group

Book review: Information Science: History, Ideas, Applications

Reviewed by: **Sergio Alonso Mislata**
Rare Print Metadata Specialist, The University of Manchester Library

Received: 10 September 2026 | Published: 16 September 2026

Seadle, Michael (2025) *Information Science: History, Ideas, Applications*. London: Facet Publishing. ISBN 978-1-78330-694-7 (paperback)

CONTACT Sergio Alonso Mislata sergio.alonsomislata@manchester.ac.uk The University of Manchester Library and address

This book presents a multifaceted analysis of the idea of information science. Its scope is wide, moving from the history of information and of information science as a discipline to information behaviour, institutions working with information, language and culture, technology and artificial intelligence, data science, non-verbal information, information retrieval and information integrity. The width of its scope might be its strongest point, particularly in the way it attempts to bring all these apparently disparate areas together under the unifying idea of information.

In considering the notion of information itself, Seadle asks "what exactly is information in scholarly and philosophical terms?" A structured answer to that question is nowhere to be found in the book. Perhaps because it is considered to be evident enough, perhaps because it is expected that the efforts to define information science as a discipline would offer a good enough insight into the nature of its object. We are told right at the beginning that "information is ubiquitous", and that that is part of the problem when defining information science as a discipline. It is probably the case that this is also part of the problem when defining information itself. The lack of such a definition, however, leaves us wondering about the principles assumed by the author in his understanding of the discipline. And, subsequently, it leaves us wondering about the principles we are ourselves being asked to assume. In that sense, even if Seadle declares that "information science today does not and should not claim to be developing natural laws", and that "it is pragmatic rather than theory driven", his characterisation of the idea of information and related ideas throughout the book seem to effectively imply a certain number of ontological commitments. So, even if these commitments do not interfere with the declared pragmatism of the discipline, it would not be completely correct to present it as not theory driven.

In a key passage, for example, Seadle states that "pure information exists apart from culture. The law of gravity applies to objects in all cultural contexts. The laws of nature

exist outside of human culture, but the values that people assign to these facts vary". Elsewhere, he declares that "every aspect of the world implicitly contains information", that information itself is "older than any human awareness that information even exists", that "language may not be essential to information, but it helps when assigning meaning and sharing that meaning with others", and that information science's core subject is "fundamentally timeless", even if he later goes on to say that things only "formally have information content once thinking creatures assign meanings to them". If "every aspect of the world implicitly contains information", then we could feel justified in believing that information emanates from it without anybody's participation: individuals would simply actualise it by giving it meaning, that is, by having an understanding of it. This would, somehow, seem to suggest that production and communication are ultimately not necessary to information's core essence. Information at its purest would be there, beyond language and culture, waiting for us, and it would be accessible in different ways and at different levels depending on the context.

I think this point of view is clearly at work when Seadle, quoting Steiner, brings up Heidegger's idea that language "remains the master of man". He does so to highlight that words are "information carriers" that can have different meanings in different contexts and affect how information is understood. This seems to be missing something fundamental in Heidegger, something deeper: language does not simply transmit an already existing and intelligible reality that the human being would try to express with more or less mastery. It is the way we build our reality, we are born into it, we are it.

Apart from these conceptual disagreements, I had a couple of additional reservations about the book. China is repeatedly used as the main alternative or comparison to the European experience, while India, the Islamic world and other historically significant cultures receive much less sustained, and much more cursory, attention. Of course, no single volume could cover all of these, but more varied examples could have offered a different picture of what information meant, and how it worked in very different historical and cultural contexts. I was also surprised by Seadle's extensive referencing from Wikipedia. He himself warns us against relying on it instead of following its references to more authoritative sources, but the book's bibliography contains twenty-six Wikipedia entries. I find these two things difficult to reconcile.

None of these reservations prevented me from appreciating the book. As mentioned above, its greatest strength is perhaps the number and variety of questions it raises, and its presentation of information science as a field that can engage with a wide range of human activities and disciplines. It also constitutes a very interesting opportunity for those working in the information field to think about our contributions in a less insular, more connected way: as a shared effort towards a higher common goal.

CILIP Metadata & Discovery Group

Minutes 230th Committee Meeting 12 March 2026

1. **The meeting took place online between 12m and 12:45pm at the MDG Conference at the Engineer's House. Clifton, Bristol**. This was a short meeting, focused on upcoming activities. These minutes were accepted at the 231st Committee Meeting on 3 June 2026, where they were proposed by Georgiana Datcu and seconded by Karen Pierce.

2. **Attendees**: Will Peaden (Chair), Paul Cunnea (Vice Chair), Anne Welsh (Secretary), Gail Beadnell, Emma Booth, Jennie-Claire Crate, Fran Frenzel, Ceilan Hunter-Green, Helen Griffiths, Martin Kelleher, Sergio Alonso Mislata, Mary Mitchell, Karen Pierce, Colette Townend, Victoria Webb, Jenny Wright

   Apologies: Georgiana Datcu (Treasurer), Meg Fisher, Terrance Mann

3. **Welcome** The Chair welcomed everyone and shared the agenda, with apologies that one had not been circulated in advance, due to his focus on the conference and the Secretary's heavy workload in her day job. He asked if there were any other items of business and the Events Team asked for the costing of events to be added to the agenda.

4. **Sponsored places**

At the last CILIP Communities meeting, a member of another Special Interest Group (SIG) had queried the procedures around sponsored places for conferences and events. CILIP asked SIGs not to use capitation money for this purpose. Will said that going forward it will be important for MDG to be able to show in the accounts to CILIP each year that the money we use for bursaries is coming from other sources (such as donations / sponsorship and money made when our not-for-profit events do, in fact, make a profit).

a. ACTION: Leadership Team to look into ring-fencing and evidencing ring-fencing of any capitation money we receive from CILIP for committee actions

b. ACTION: Will and Paul to meet to discuss the MDGS approach and extract any lessons learned from this that MDG can adopt

5. Hours tracking

CILIP has asked us to track our hours.

a. ACTION: Leadership Team to set up an online form for us all to complete

b. ACTION: All to feedback any comments on this topic to the Chair and Secretary

6. Charges for Events

The Events Team shared that the last set of webinars was very well-subscribed, mostly from people whose institutions paid for their places. The Events Team feels that we could charge more without its causing a pain point for most attendees. It was suggested that if we raise our charges, we could stress that people who are unable to attend could email the Chair and Secretary to ask for a bursary place. As always, there are many options, and to avoid the restatement of a discussion we have already had in a meeting that could last a maximum of 45 minutes, the Chair guillotined the discussion until a future date. The Secretary asked the Events Team when they felt they would be able to present a full proposal and they opted for the December Meeting.

a. ACTION: Gail and Victoria to present their ideas for charges at the December Meeting for discussion by the whole committee. In the interim, the Events Team, Treasurer and Chair will continue to set charges in the way we agreed last time charges for events were discussed.

7. IFLA Satellite Meeting

There was a conference session the previous day which was an informal chat about IFLA and it was very well attended. Several members were very enthusiastic about the prospect of a satellite meeting focused on cataloguing topics. The committee was very pleased by the turnout and by the engagement from MDG members at the event. As a brief summary, those present who have IFLA experience (Jenny Wright, Paul Cunnea, Renate Behrens and Neil Murray) shared useful information both for those considering volunteering / attending WLIC 2027 and for MDG on how Satellite Meetings are usually organised and how we might go about letting relevant IFLA Sections know we are interested in hosting a Satellite. There are several IFLA Groups in the area of Cataloguing and Metadata and both Jenny Wright (MDG, UKCoR, RDA Board) and Renate Behrens (RSC) are active in these. The RSC has an official protocol with IFLA so if we partner with them Jenny and Renate have a very straightforward route in, as well as the official calls for Satellite Meetings that go forward. In the committee meeting, the MDG committee decided that we will partner with MDGS to propose a satellite meeting for the IFLA Conference. Subsequent to the meeting, IFLA issued its official call, and Will, Paul and Anne attended the CILIP online session for those interested. Following this, the Leadership Team has started to draft an application and to seek volunteers for a local Satellite subcommittee. Renate and Jenny

have approached several IFLA Sections who have warmly welcomed the idea of a Cataloguing Satellite, which CILIP informed us will be useful to include in our application form.

a. ACTION: Leadership Team to coordinate actions with MDGS to (i) submit and application to CILIP's call for Satellite hosts and (ii) organise a Satellite subcommittee.

8. Approval of Minutes of Meeting 229

Proposed by Fran and seconded by Karen.

9. MDG reports

a. Chair (**Will**): see items 4-7 above
b. Secretary (**Anne**): see items 4, 5 and 7 above
c. Treasurer (**Georgiana**): report received (Appendix 9.c)
d. CILIP Member Council (Memco): nothing to report
e. MDGS Metadata and Discovery Group in Scotland (**Paul**): MDGS volunteered to host an IFLA Satellite and would like to collaborate with MDG on this (Item 7)
f. WHELF (**Helen**): nothing to report

10. Reports from MDG representatives on other committees

a. ARLIS (**Mary**): report not received
b. IAML (**Meg**): report not received
c. RBSCG BSG (**Sergio**): report received (Appendix 10.c)
d. EURIG (**Anastasia**): report not received
e. UKCoR (**Anastasia**): report not received
f. DDC UK (**Terrance**): report not received
g. Mercian Metadata Group (**Will**): nothing to report
h. ALN (Academic Libraries North) Metadata CoP (**Ceilan**): report not received
i. MADSIG (Metadata and Discovery Southern Interest Group) (**Jennie-Claire**): Jennie believes that MADSIG has ceased to operate and so we can remove this report from the list
j. NAG (**Colette**): report received (Appendix 10.j)
k. BIC (**Will**): nothing to report

11. Reports of MDG events, Activities & Communications

a. Website Editor (**Fran**): nothing to report
b. Communications (**Emma**): report received (Appendix 11.b)
c. Marketing (**Jennie-Claire**): nothing to report
d. C&I (**Fran and Karen**): post hoc: Volume 214 was published on 23 March

e. Cataloguing Ethics Steering Committee (**James**): report not received
f. Events (**Gail and Vicky**): report received (Appendix 11.f)
g. Conference subcommittee: conference proceeding as planned; all present shared their many thanks to the subcommittee for their work
h. Metadata competencies (**Jenny**): report not received

12. Date of next meeting: 3 June 2026, 2:30-4:30pm

The Secretary will circulate a Teams Meeting request

13. Ongoing actions

(See Appendix 13)

APPENDIX

9. MDG Reports

9.c Treasurer

Prepared by Georgiana Datcu, February 2026

Financial Summary

1. *Charity Account*

- Balance: ÷£10,000
- Monthly Interest: ÷£39 (credited to the current account)

2. *Unity Trust Bank Account*

- Balance (at the start of 2026): £35,784.17
- Budget (to date) to make use of: £16,089;
- Set aside
  - £19,000 to cover 2026 Conference costs
    - Additional £8,500 put aside as received from sponsorship
      - Of which £3,772.5 spend on covering deposits for venues
  - £1,296 to cover Uni of Edinburgh Diamond (CILIP potentially to take this on board – pending!)
  - £2,500 to cover WorldCafe x3 events
  - £3,000 to cover Library Carpentry expenses
  - £3,000 to cover travel expenses for committee members in 2026
    - Of which £124.5 spent

Members Network annual funding grants - applied for £1,200 (received)

MDG has secured the grant in 2025 - this paid for a full bursary at MDG conference 2026.

The grant **will not continue into 2026** following CILIP's decision to cease capitation funding for all groups.

Income and Expenditure Updates

3. Upcoming Income

For details, refer to the Events Report.

4. Projected Expenditures

- Event speaker fees.
- Travel expenses for venues selected by the Conference Committee.

Administrative Updates

5. Lloyds Corporate Card Bank Account Access

The Treasurer is receiving correspondence about this account closing. Treasurer is investigating legitimacy and what is occurring. – **Lloyd's Credit Card closed**

6. Financial Document Retention Policy

The Treasurer seeks the MDG Committee's approval to confidentially shred any financial documentation older than five years. Upon approval, this material will be securely destroyed. – **pending discussion by Leadership Team**

10. Reports from MDG representatives on other committees[1]

10.d RBSCG BSC

Liaison to the MDG Committee report highlights

Highlights from the latest MDG Committee meeting presented.

[1] The editors are aware that some links in the reports are no longer functional or do no longer exist due to CILIP's website upgrade. However, we felt the minutes should be included as produced at the time.

RBSCG Committee report

Nothing to report from the RBSCG Committee as they are meeting the week after this RBSCG BS Committee meeting. Notes from that meeting will follow, to be disseminated among members of the Committee by email.

Liaison to the RBMS BS Committee report

The following points were presented:

From the MARC Advisory Committee liaison

There was a proposal which had support from RBMS and was approved without corrections, to add a 655 field to the MARC holdings format as a way of distinguishing between multiple copies held by an institution.

Another proposal, this one approved with minor revisions, consisted of adding a subfield z to the 245 field, to add a context note statement to a title. Its purpose is to meet the reparative description need for immediate contextualization of outdated, biased, prejudicial and/or hateful language as found in transcribed titles. Not so much a content warning as a note (between square brackets) explaining why this harmful language is present in the catalogue record (the basic option being that it is present because it has been transcribed as it appears on the original source), etc.

From the RBMS RDA Editorial Group

The RBMS Policy Statements (concise statements that provide guidance on the application of the options in the RDA Toolkit), which are separate from the DCRM RDA edition (DCRMR), have now been uploaded in the October 2025 official RDA Toolkit release.

From the OCLC liaison

OCLC have published a series of blog posts on AI and cataloguing. There is one blog post specifically on the use of AI in archives and Special Collections cataloguing.

From the DCRM editorial board

The DCRM editorial board has been working on including in DCRMR guidance for the cataloguing of graphic materials.

Once this is finished, it is feasible that the turn of either photographic or cartographic materials will come.

The editorial board has launched a survey to get input into what format users would like to be added to DCRMR next.

From the Library of Congress liaison

Marva and BIBFRAME

- LC cataloguers are now using Marva to create records, therefore using de facto BIBFRAME instead of MARC to create original records.
- LC library management system, FOLIO, converts the records into MARC to make them compatible with external library systems.
- To allow time for this transition into BIBFRAME to take place, LC has postponed its implementation of the official RDA and is still using the original RDA. The plan would be to switch over to the new RDA Toolkit sometime this year.

Dropping subfield v on 6XX fields

- LC has decided to discontinue the use of subfield v (for formats) in 6xx headings.
- Within the Rare Materials section, LC is still applying subfield v but coding the terms in that subfield as local headings and thereby treating its use as a local decision rather than a nationally supported standard. The terms recorded in subfield v are drawn from the legacy RBMS vocabularies rather than from the merged RBMS Controlled Vocabulary (RBMSCV).
- The official message from the Library of Congress is that other institutions are still welcome to continue adding it if they choose to do so, but new records downloaded from LC will no longer carry it.
- The recommended alternative is to add a separate 655 with the relevant term.

Authority field 375

- Field 375 (for gender) is being programmatically removed from LCC authority records in response to concerns about privacy.

Other news

- ALA has just announced the publication of recommendations for the creation of harmful metadata and/or harmful language statements.
- Iris O'Brien, from BL, has announced that the ESTC is now allowing institutions to request contributor accounts to record and update their ESTC holdings. ESTC has now migrated to the CERL website.

Committee engagement and public communication

Faced with the necessity of establishing whether the Committee wants to systematically and publicly engage with its members and interested public on issues of varied nature and to a variety of ends, it has been decided that this should be assumed by the Committee to fall within the scope of its duties.

It has been agreed that two main channels of communication should be used by the Committee to reach its audience:

- The Cataloguer's Corner section of the RBSCG newsletter: This feels like it might be the best place to communicate cataloguing-specific wider topics potentially of interest to the more general public. The newsletter would only be published three times a year, and a contribution to the cataloguer's corner consists of a 500-word piece of writing, with a picture if possible. The members of the Committee have been compelled to step up and regularly contribute to it. Katie Birkwood has volunteered to write about genre headings and prejudicial works for the next issue.
- The group's JiscMail listserv: To interact with fellow practitioners on more timesensitive matters, such as calls for feedback on the publication of a new standard, for example. Anybody who's interested or involved in the technical side of Special Collections cataloguing would be, in principle, subscribed to the group's JiscMail listserv and could be potentially reachable through it to raise awareness about and ask for feedback on relevant topics.

RBMS BS Committee Liaison

Sergio Alonso Mislata has stepped up to become the new liaison to the RBMS BS Committee.

Training and events for 2026

A training event has been programmed for April, and a number of members from the Committee have volunteered to organise cataloguing-related training sessions.

- Christine Megowan has already put a training session for the 15th of April in the calendar, an introduction to rare book cataloguing led by her, to take place at The University of Edinburgh, her institution.
- Sophie Floate has expressed a willingness to help organise and deliver an online Q&A session.
- Mary Clayton-Kastenholz has committed to arrange and deliver a training session at Lambeth Palace Library and Archive for summer-autumn. Ken Gibb will help with the preparation and delivery of the session.
- Sergio Alonso Mislata will enquire about the possibility of having a training session on rare book cataloguing at The University of Manchester.

Public discussion on the use of RBMS CVCRMC's "prejudicial works" terminology

Katie Birkwood has recently attended a meeting with Karen Pierce (University of Cardiff and member of the MDG Committee) and Alexandra Hill (Wellcome).

The three of them work with historical medical collections, which, to a large extent, can be considered to reflect medical theories that were inherently racist.

The aim of the meeting was to start thinking about ways to apply these terms effectively, usefully, and meaningfully to the collections under their care.

Now they are thinking about the possibility of setting some sort of e-discussion forum about it, open to contributions from anybody, where clear questions about possibilities for their use are posed.

To be publicised, if possible, under the RBSCG BS and/or MDG brand.

10.j National Acquisitions Group

Metadata standards for public libraries

It was raised at the last NAG executive committee meeting the need to progress on creating a national standard for metadata for public library use. NAG led by Emma Booth previously produced a survey and fielded expectations in the academic sector to develop a standards document to assist in procurement but also for an expectation of quality in catalogues.

With new issues arising in the public libraries sector, NAG are interested in exploring employing someone to undertake this project for public libraries. We would also like to hear from public library services and individuals in the public library sector interested in developing metadata networking and training resources for public library services. Anyone interested in an informal discussion about this can contact Colette CTownend2@lambeth.gov.uk or Jennie at NAG nag.office@nag.org.uk

Upcoming events

Coffee and chat (Wed 22nd April 1030-1130) on ILL and CONARLs – Tickets available for NAG Members at: https://nag.org.uk/event/chatapr26/

NAG Seminar and Forum in Nottingham this May

NAG Public Library Forum 20th May, includes talks on LEGO workshops, cocreating collections with children and updates for the NAG grant and award. Tickets and more information available at: https://nag.org.uk/event/forum-2026/

NAG Collection Development Seminar for Academic Librarians is both days 20th and 21st May (https://nag.org.uk/event/seminar-2026/). At the end of the day on the 20th there will be a ploughman's buffet (sponsored by Perlego) and a drinks reception

(sponsored by BDS) to enable further networking opportunities in a more informal atmosphere.

NAG strategy survey received over 80 responses which we are now collating and digesting to inform NAG policy going forward.

Library Carpentry Events

NAG have met with CILIP MDG colleagues to discuss co-hosting metadata skills workshops, potentially in Birmingham this summer. Details forthcoming.

Membership fee increases

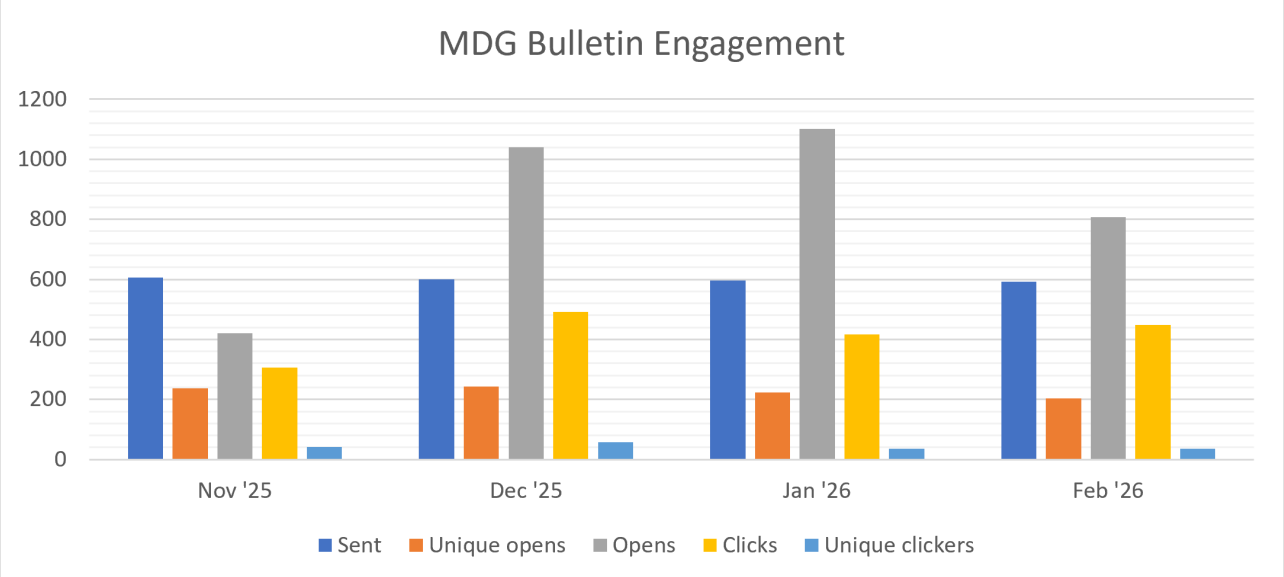
NAG membership fees have had to go up with rising costs, for commercial corporate members and for non-profit corporate but the current price has been protected for public libraries and individual members.

11. Reports of MDG Events, Activities & Communications

11.b Communications

EB 31/03/2026

MDG Bulletin



November 2025. Sent 24/11/25. 606 emails sent. 421 opens; 237 unique opens (56%); 306 clicks; 41 unique clickers (13%). Top 3 items:

- MDG Conference Programme – 12 clicks
- MDG Conference – 17 clicks

- Engineers' House webpage – 12 clicks

December 2025. Sent 18/12/25. 601 emails sent. 1041 opens; 242 unique opens (23%); 491 clicks; 57 unique clickers (12%). Top 3 items:

- C&I December 2025 issue (213) – 83 clicks
- C&I website; latest issue – 51 clicks
- C&I article: "Xenoglossophobia: taming the fear of all the jargon" – 45 clicks

January 2026. Sent 27/01/26. 596 emails sent. 1102 opens; 224 unique opens (20%); 416 clicks; 36 unique clickers (9%). Top 3 items:

- MDG Conference – 62 clicks
- MDG Conference Programme – 48 clicks
- IFLA website – 46 clicks

February 2026. Sent 25/02/26. 592 emails sent. 807 opens; 203 unique opens (25%); 448 clicks; 36 unique clickers (8%). Top 3 items:

- MDG Conference Programme – 80 clicks
- MDG Conference – 72 clicks
- MDG Website – 34 clicks

We still seem to be losing a few mailing list subscribers each month; unclear if this is CILIP purging lapsed members, or members choosing to unsubscribe.

Data suggests a lot of 'return readers' in the run-up to the Conference.

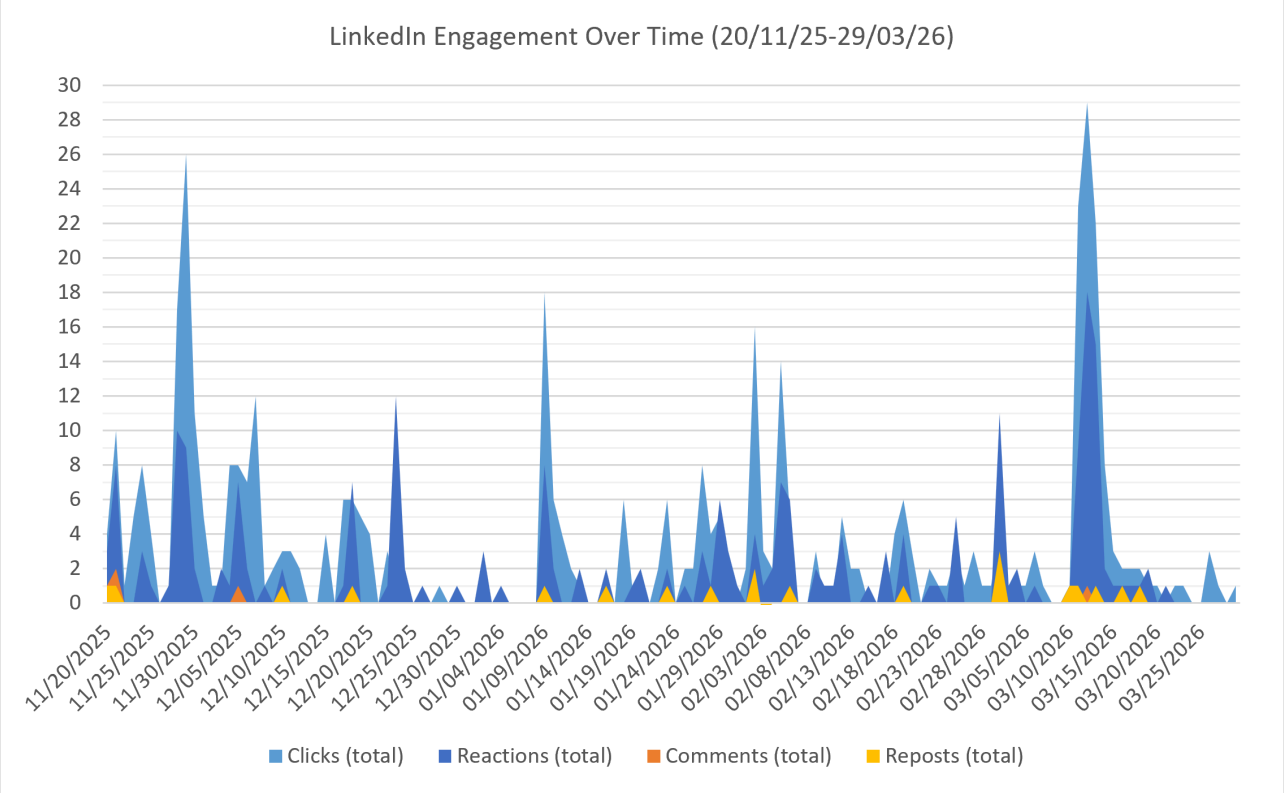
Emails sent to Listservs:

- 24/11/25 - MDG Bulletin November 2025
- 05/12/25 - Society of Indexers Coffee and Chat
- 18/12/25 - MDG Bulletin December 2025
- 09/01/26 - Society of Indexers Coffee and Chat
- 09/01/26 - MDG Conference & RDA Day Promotion
- 27/01/26 - MDG Bulletin January 2026
- 28/01/26 – International Metadata Recommendations Report Published
- 16/02/26 - Society of Indexers Coffee and Chat
- 25/02/26 - MDG Bulletin February 2026
- 02/03/26 - MDG Conference & RDA Day Promotion
- 17/03/26 - Society of Indexers Coffee and Chat

LinkedIn:

https://www.linkedin.com/company/cilip-mdg/ Created: 20/11/2024.
Total followers = 1,061
In the period 20/11/2025 - 29/03/2026:

- Page views = 279, Unique visitors = 130, Followers gained = 134,
- Our 40 original posts (i.e. not including reposts) have had 9893 Impressions/views, with 219 Reactions/Likes, 409 Clicks, 5 Comments, and 111 Reposts.
- Top 3 posts in terms Total Engagement (= Clicks + Reactions + Comments + Reposts):
  - 28/11/25 – "Webinar: Python" = 59 Clicks, 15 Reactions, 1 Comment, 4 Reposts.
  - 11/03/25 – "MDG Conference in Bristol" = 55 Clicks, 15 Reactions, 1 Comment,7 Reposts.
  - 29/01/26 – "MDG Conference Bursary" = 18 Clicks, 11 Reactions, 0 Comment, 11 Reposts.



X:

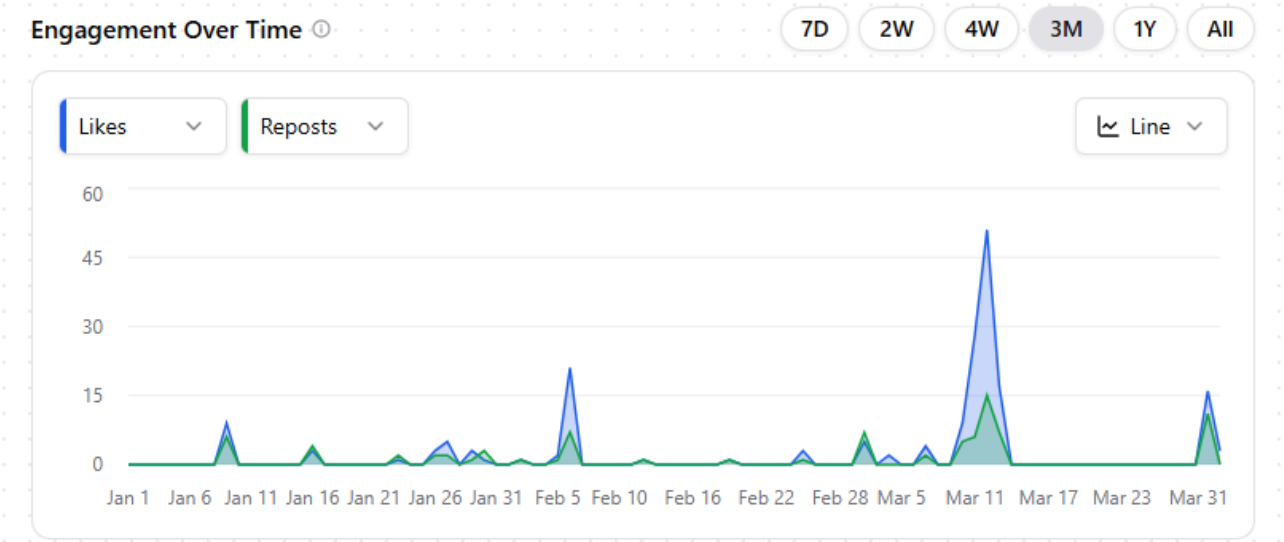
https://x.com/CilipMDG Followers: 1,512 (down from 1,551 in December 2025). Account mothballed as of 01/01/26

BlueSky:

https://bsky.app/profile/cilipmdg.bsky.social Created: 18/03/2025. **Total followers = 395**

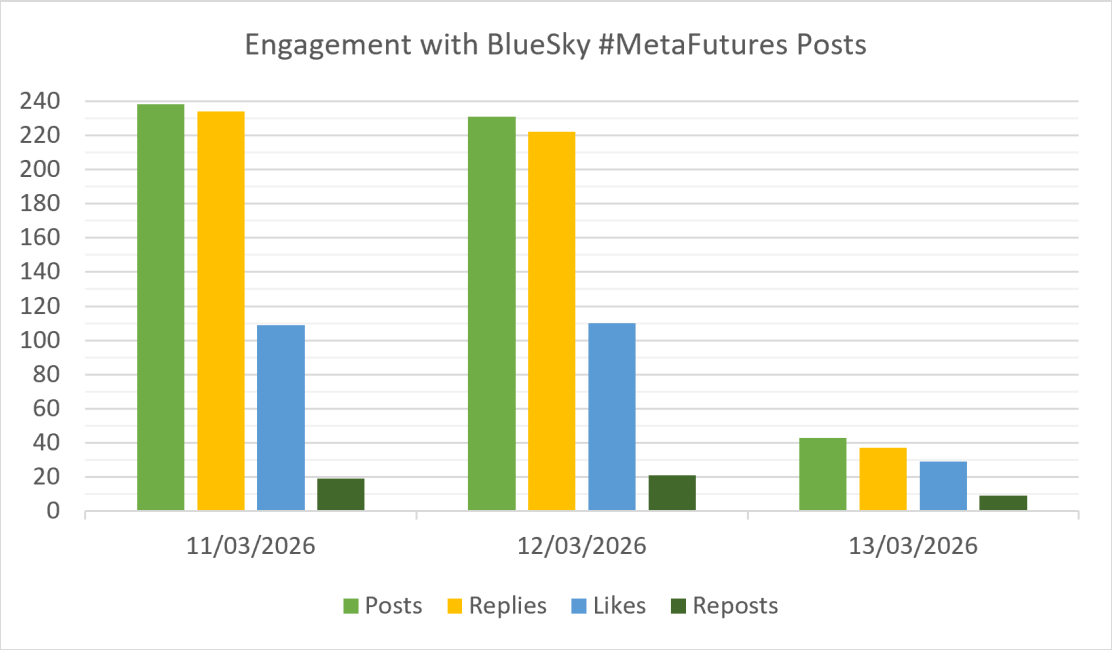
In the period 20/11/2025 - 29/03/2026:

- Followers gained = 63
- **643 posts have received 449 Likes, 577 Replies, and 131 Reposts.**
- Top 3 posts in terms of Total Engagement (= Likes + Reposts + Replies):
  - 11/03/26 – "We're here ready to welcome you to the MDG Conference" = 10 Likes, 2 Reposts, 1 Reply.
  - 09/01/26 – "We'll be posting highlights from the MDG Conference Programme" = 6 Likes, 5 Reposts, 1 Reply.
  - 12/03/26 – "Today's final paper: Fixing the Leaky Pipeline..." = 7 Likes, 3 Reposts, 1 Replies.



Focus on MDG Conference 2026 #MetaFutures Posts:

- **Wed 11/03/26**: 238 Posts with 109 Likes, 234 Replies, 19 Reposts
- **Thurs 12/03/26**: 231 Posts with 110 Likes, 222 Replies, 21 Reposts
- **Fri 13/03/26**: 43 Posts with 29 Likes, 37 Replies, 9 Reposts
- **Totals**: 512 Posts with 248 Likes, 493 Replies, 48 Reposts



11.f Events

Following a busy Autumn 2025, it has been a quieter time so far in 2026 with only the Society of Indexers coffee mornings being held. Discussions are ongoing to deliver two further World Cafes on Subject Analysis and the Library Carpentry sessions.

Unfortunately, Penny Curry, our Digital Design Assistant, withdrew from the voluntary role. We will look to generate interest in the role at the Conference and in the meantime the workload is manageable, shared between the 2 Events Coordinators.

Events held:

Society of Indexers coffee morning (online)

Two colleagues registered for the January session, and 5 colleagues signed up in February. While these numbers are low, it does keep communication open between MDG and the Society. As the current Chair Paula Clarke Bain, is presenting at the Conference, more interest may follow for future sessions.

Upcoming Events:

World Cafes on Subject Analysis (Cardiff, 21st April 2026; Liverpool, 29th April 2026)

Two regional sessions repeating the London-based workshop held in November 2025.

Visit to the Drawing Room (London, 20th May 2026)

Following librarian Yamuna Ravindran's presentation at the Autumn webinar series, she offered to host a visit for MDG members. The timing of this visit coincides with the Drawing Biennial 2026, offering attendees the opportunity to experience the exhibition after the visit. https://drawingroom.org.uk/

Library Carpentry (Aston University, June 2026, dates tbc)

Two sessions over two weeks focusing on Tidy data (spreadsheets), Working with data (RegEx) and OpenRefine.

13. Ongoing Actions

All other actions have been discharged.

a. **AGM, 2 September 2026**: International speaker to be agreed and invited; Secretary to coordinate (Meeting 228 Item 10.a)
b. **Ethics email address**: Fran to set up a Google email (Meeting 228 Item 14.a)
c. **Ethics event**: James to find resources for sharing local practices and coordinate with the Events team (Meeting 228 Item 14.b)
d. **Standing item: Secretary to arrange regular meetings of the Leadership Team** (Meeting 225 Item 9.3). The next Leadership Team meeting is scheduled for 15 May 2026 and information from it will be read into the minutes for meeting 231 as appropriate.
e. **Standing item: Affinity acquisition**: Fran to let the Treasurer know when she is ready to acquire an Affinity login and Treasurer to provide funds (Meeting 226 Item 226.5.2.1)
f. **Standing item: Bulk ingestion of online back issues of C&I**. Fran to continue to provide updates at each meeting until completed (Meeting 226 Item 5.4.2)
g. **Standing item: MAC Papers**: Will as MDG representative to the BIC LMG to circulate MAC papers and deadline for feedback as and when received from the BIC LMG Chair (Meeting 227 Item 14.k)

Catalogue & Index is electronically published by the Metadata & Discovery Group of the Chartered Institute of Library and Information Professionals (CILIP) (Charity No. 313014).

Submissions: Please follow the submission guidance on our website.

Book reviews: Please contact the editors.

Advertising rates: GBP 70.00 full-page; GBP 40.00 half-page. Prices quoted without VAT.

ISSN 2399-9667

MDG Website: www.cilip.org.uk/mdg

CILIP SPECIAL INTEREST GROUP

OPEN ACCESS