

Catalogue & Index

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



September 2025, Issue 212

ISSN 2399-9667

EDITORIAL

On 15th October 2002 Roy Tennant declared that “MARC must die” in the Library Journal, fifteen years later in 2017 he reflected on this article on the OCLC Research blog – Hanging Together – and described the ‘firestorm of criticism’ that had arisen in the wake of his original article. He also noted that during the intervening years, the idea had ceased to be controversial. The majority of us still work with MARC, despite the repeated assertions that it is on its way out, but replacements, such as BIBFRAME, are on the rise. We should also consider, however, that there are plenty of metadata practitioners who don’t use MARC at all – and for both these reasons therefore this issue of C&I is dedicated to exploring non-MARC cataloguing and metadata practices.

Helen K. R. Williams showcases an exciting Wikidata project undertaken by LSE Library’s Metadata team to enhance the discoverability of a set of oral history interviews about the British Suffrage Movement which were conducted between 1974 and 1981. Williams has previously contributed to C&I (206: 2-6) with a paper on the LSE Wikidata thesis project and we hope readers will be interested in seeing how LSE’s contribution to Wikidata continues and has enhanced staff skills alongside increasing the discoverability of their resources.

Ourania Karapasias’s article uses a case study approach to illustrate the process of converting MARC records to enriched TEI-XML files that form meaningful digital objects in Manchester Digital

- 1-2** Editorial
The Editors
- 3-11** **‘We are here not because we are metadata-breakers; we are here in our efforts to become metadata-makers’**
Helen K. R. Williams
- 12-25** **Envisioning Dante**
Ourania Karapasias
- 26-33** **Decolonising heritage collections**
Jessica Roberts
- 34-44** **Uncharted cells**
Carol Hunter
- 45-48** **The challenges of data ingest, transformation and aggregation at the National Bibliographic Knowledgebase**
Jennie-Claire Crate
- 49-58** **Non-MARC cataloguing**
Anne Welsh
- 59-62** **Book review: Ethics in Linked Data**
Elizabeth Cooper
- 63-65** **Book review: Records and information management**
Sarah Henning

Collections. The case study comes from the Dante Digital Library project which is digitising 99 editions of Dante Alighieri's Divine Comedy. Karapasias demonstrates the advantages of using TEI files to more accurately represent the descriptive complexity of digitised early printed books.

Jessica Roberts from the People's Collection Wales talks about their decolonisation toolkit which was released in 2025 and provides structured guidance for dealing with discriminatory and colonial language within the context of descriptive metadata on the digital heritage platform. A variety of case studies indicate the kind of issues that might be encountered and how they have been addressed. As contributions to PCW are welcomed from everyone, not just heritage staff, it is crucial to have clear guidelines in place.

Carol Hunter describes a project that arose out of necessity when the Covid pandemic led to her team at the National Library of Scotland working from home without access to their library management system. By utilising spreadsheets staff, with or without cataloguing experience, were able to input data from digitised HMSO catalogues, which was subsequently ingested by Alma. The success of the project has led to this process being used more extensively for routine metadata tasks.

Many of us are contributors to Jisc's National Bibliographic Knowledgebase, and in her article Jennie-Claire Crate talks about the challenges of ingesting data through different transfer methods, deduplication of records, and working with non-standard metadata.

Our final article by Anne Welsh discusses why non-MARC cataloguing systems are used, and offers some useful things to think about if you find yourself dealing with one.

We also have two book reviews: one of Ethics in Linked Data and one of Records and Information Management.

Our next issue in December will be looking at various aspects of RDA, if you wish to contribute an article on this topic please contact the editors at catalogueandindex@gmail.com.

Karen F. Pierce & Fran Frenzel, September 2025

‘We are here not because we are metadata-breakers; we are here in our efforts to become metadata-makers’

Helen K. R. Williams  0000-0003-1259-7097

Metadata Manager, The London School of Economics and Political Science

Received: 8 August 2025 | Published: 22 September 2025

ABSTRACT

This article outlines a Wikidata project undertaken by LSE Library’s Metadata team to enhance the discoverability of a unique set of oral history interviews about the British Suffrage Movement. The paper outlines each stage of creating linked open data to represent interviews, interviewees and associated entities, and reflects on challenges less familiar to traditional cataloguing practices in RDA and MARC21, including data protection impact assessments, data modelling, URL stability, and creation and enrichment of contextual linked data, including non-LCSH identifiers. The project explores new ways to interrogate and visualise metadata through SPARQL queries, and highlights opportunities to contribute to Wikipedia and Wikimedia Commons. Reflections on conflict of interest, community engagement, and the evolving role of metadata professionals are shared, alongside early insights into project impact. The article demonstrates how Wikidata can extend traditional metadata practice, offering new opportunities for collaboration, discovery, and representation.

KEYWORDS Wikidata; open linked data; discovery

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

Introduction

One hundred and seventeen years on I wonder what Emmeline Pankhurst would make of my appropriation of her famous words ([Pankhurst, 1908](#)) for an article about ‘breaking out’ of MARC metadata silos. As we venture beyond the traditional remit of library metadata teams and embrace new ways to make suffrage voices discoverable, I like to think she would cast her vote of confidence in our direction!

Many *Catalogue & Index* readers will already be familiar with LSE’s Wikidata thesis project¹ which successfully increased the reach and engagement of this content while also developing the skills of the Metadata team in contributing to the broader linked open data ecosystem ([Williams, 2022](#)).

¹ https://www.wikidata.org/wiki/Wikidata:WikiProject_LSEThesisProject

Building on that success, and with the aim of further expanding the team's Wikidata expertise, we explored further opportunities to contribute unique content to Wikimedia platforms and selected a set of oral history interviews about the British Suffrage Movement. The 205 interviews (with 183 individuals) were conducted by the historian Brian Harrison between 1974 and 1981, and were subsequently deposited with The Women's Library². The interviews (with surviving suffrage campaigners and their relatives and employees) also explore broader themes around women's organisations, employment and family life, the birth control movement, politics, peace activism, and trade unionism ([LSE Library, no date](#)).

Wikidata metadata creation and enhancement

Having not previously worked with oral history interviews, in Wikidata or any other context, I initially explored the Wikidata pilot pages of the Program for Cooperative Cataloging ([Wikidata, no date](#)) to find other universities working with similar materials. These examples informed my understanding of appropriate data models to support rich contextual linked data and resulted in the creation of local data models for collection and interview level records, and for entities represented in them.

Following the principle that producing 'something' 'good enough' is more valuable than aiming for perfection and potentially achieving nothing, I outlined a basic iterative and experimental project approach:

- Complete Data Protection Impact Assessment (DPIA) and create Wikidata project page.
- Automate creation of basic QIDs for names not yet represented in Wikidata.
- Enrich new and existing QIDs using local authorities and interview summaries.
- Create QIDs for interviews, using linked data to connect interviewees and named entities.
- Edit relevant Wikipedia pages for individuals/organisations.
- Explore creation of Wikipedia pages for notable individuals currently missing from Wikipedia.

Our project page³ ([Williams, no date](#)) demonstrates that we largely followed this structure, working out the detail as the project progressed, and adding a couple of extra steps as required.

Before beginning we explored data protection considerations and completed a DPIA. Interviews which do not have consent to be made public were excluded from the project. Interviews and related metadata for the rest of the collection are already publicly available, but the DPIA highlights the way in which the project will bring

² <https://www.lse.ac.uk/library/collection-highlights/the-womens-library>

³ https://www.wikidata.org/wiki/Wikidata:WikiProject_The_Women%E2%80%99s_Library_LSESuffrageInterviewsProject#Tasks

metadata together in new ways to increase discoverability in the semantic web environment, and assesses potential risks accordingly.

At this stage we also needed to address the issue of URL stability. The metadata records, with accompanying descriptions for each interview, are hosted on our archives catalogue, but an imminent system migration meant that existing URLs would change. The audio files are accessed via the LSE website but plans to re-digitise the collection meant that these URLs were also subject to change. Consequently, neither option was stable enough to provide reliable reference URLs on the hundreds of Wikidata entities we would create. Via Wikidata help pages⁴ ([Wikidata, no date](#)) I explored the option to use a *'stated in'* qualifier instead. This could be linked to the collection level QID⁵ for the project. If URLs changed then only one QID would need updating to deprecate the old URL and provide the current one.

With metadata for the interviews and related authority records extracted from our Archives system I was able to reconcile names with Wikidata via OpenRefine. Of 222 names 98 were already represented in Wikidata, and 124 were not. Information about these individuals was spread across interview metadata, authority records and interview summaries, so manual creation of full records would be time consuming. I prioritised a 'quick win' by focusing on missing names and uploading an easily assembled spreadsheet of basic information to OpenRefine, from where we could write a schema to batch upload to Wikidata. This included statements for label (entity name), description, alias, given name, family name, occupation, gender, dates of birth and death, and identifying that an oral history for the individual was held at LSE, all of which was standard metadata easily extracted from our Archives metadata exports. Had we been creating entities for living people I would not have presumed to assign such personal data as gender, but all names associated with the project were thought to be deceased, and the inclusion of Wikidata's sex or gender property was relevant to the historical context. The newly created QIDs were extracted along with existing ones, and a bulk update to add P5008 properties *'on focus list of Wikimedia project'*: [The_Women's_Library_LSESuffrageInterviewsProject](#)⁶ gathered them into one identifiable project set.

All names were then ready for manual enhancement, a step considered a worthwhile investment of time given the uniqueness of the content and the collection's importance as a key source for British women's history of the 20th century. The wider team used interview metadata and summaries to add further statements to the Wikidata entities, including country of citizenship, place of birth or death, employer, educated at, notable works, affiliation, spouse, and relatives. All enhancements were recorded on a spreadsheet which was later used to batch load all edits to Wikidata. Metadata for the interviews themselves required manual enhancements in three stages, each designed to create rich contextual linked data:

⁴ <https://www.wikidata.org/wiki/Help:Sources>

⁵ <https://www.wikidata.org/wiki/Q100380678>

⁶ <https://www.wikidata.org/wiki/Q117322976>

- **Significant people mentioned in interviews** (defined as those either already represented in Wikidata or identifiable through at least one verifiable attribute). For the latter, the team populated a 'missing people' spreadsheet with name, Wikidata properties to represent identifying information (such as employer or family relationship), and reference URLs to verify these statements. This was uploaded to Wikidata via OpenRefine, with a schema which also automated the addition of each name in a significant person statement, with the qualifier 'object of statement has role: subject' to the correct interview QID. Identifying significant names for each interview was a manual process that took place in 2023. Newly digitised audio files and accompanying transcriptions may now allow us to further enhance our linked data by using named entity recognition to generate fuller metadata and identify additional associated individuals.
- **Organisational names**, such as National Union of Women's Suffrage Societies (NUWSS), Women's Social and Political Union (WSPU), and others were added as subjects of interviews where summaries indicated significant discussion.
- **Significant places and events**, such as imprisonments, specific arson events, or named marches enhanced contextual links.

The work with significant people raised an unexpected conundrum - should all the people mentioned in the interviews be included on the project's P5008 focus list in the same way as the interviewees and the people who were their main subjects? For example, the interview titled *Adams, Mrs Dorothy re Anna Munro* is with Dorothy Adams, talking about her employer, Anna Munro. Although Munro herself was not interviewed, she is the focus of the interview and would be valuably included on the focus list of the project by adding a statement to her QID to indicate this. The interview mentions a number of other names as well, such as Munro's husband, her children, both of whom are also interviewed as part of the project, and other notable figures in the suffrage movement. Wikidata encourages the creation of meaningful connections, so in that spirit all these names were added as significant people on the QID for the interview, but their inclusion on the project's focus list was less straightforward. Readers may hold differing views on the 'correct' approach, but after internal discussion it was decided that in the interests of suffrage research it was valuable to record as many names as possible so that the project could potentially bring to light new connections or relationships, and widen research possibilities. Guidance was given to the team in assessing how substantive a 'mention' was in order to avoid making excessively time-consuming connections with misleading value.

Interconnections and visualisations

One of the key advantages of working with Wikidata is the ability to explore and interrogate our data in ways that were not previously possible. In a MARC environment, metadata feeds into the library catalogue, where metadata specialists and many users are familiar with established search techniques to locate relevant

content. Wikidata, by contrast, offers a more dynamic and flexible approach. Using SPARQL queries we can break the data down into a whole variety of meaningful categories, offering researchers more granular insights. The categorisation of individuals for this project is supported by the data model used for interview QIDs, and further details can be seen in the *Names data* section on the project page. Importantly, by contributing our metadata to Wikidata, we can link it to existing entities and generate visualisations of those connections, something not possible when metadata remains siloed within an archives catalogue. For example, we can now explore questions such as:

- who else holds archives, or oral histories, of people connected with the interviews?
- What family relationships exist between individuals?
- What awards did they receive?
- What spread of occupations is represented?
- What significant events are mentioned across the interviews?

A full list of the SPARQL queries being used to interrogate the collection can be seen on the project page⁷.

Contributing to the wider Wikimedia community

While contributing metadata to Wikidata has been the cornerstone of this project, we also explored the broader suite of Wikimedia platforms to further enhance access and visibility. This is another step expanding the traditional roles of MARC21 or institutional repository metadata creators. As a new editor I found it to be an area where, despite my best attempts at interpreting Wikimedia policies, I made genuine mistakes, was corrected by other editors, and needed to revise my approach.

Those of us who manage metadata teams are often used to being the institutional experts in our domain, and less accustomed to others critiquing our metadata than we might once have been! Entering the world of Wikimedia and transparently seeking to make our content accessible to a wider global audience, in the right way, has been a steep learning curve, where I have needed to learn from the generosity of more experienced editors.

As an LSE employee I needed to declare this affiliation on my user talk page and avoid edits which could be construed as conflict-of-interest. This can perhaps be a particularly confusing area for university staff. Making unique content, with research potential, available for a global audience does not feel like an exercise in self-promotion in the same way as it might for a commercial entity, and yet the same policies and risks apply. In my case, I misunderstood guidance around external links and references to my employer in article text. Once alerted, I revised edits accordingly

⁷ https://www.wikidata.org/wiki/Wikidata:WikiProject_The_Women%E2%80%99s_Library_LSESuffrageInterviewsProject#SPARQL_queries

and engaged in discussion with fellow editors. I have also engaged in other areas of Wikipedia editing, rather than just those related to content from my institution, both to be, and demonstrate willingness to be, a 'good citizen' in this global knowledge space.

Perseverance and willingness to learn has enabled me to contribute to regularly viewed biographical and organisational Wikipedia pages and receive thanks from other editors. The Women's Institute article, for example, is viewed nearly 42,000 times a month ([Wikipedia, 2025](#)) at the time of writing, while the National Union of Women's Suffrage Societies sees over 14,000 visitors ([Wikipedia, 2025](#)). Biographical pages tend to attract lower page views than subject pages, though the actress Sybil Thorndike, one of the interviewees, receives nearly 51,000 monthly views ([Wikipedia, 2025](#)). Among other interviewees, the socialist politician Margaret Cole receives over 8,000 page views ([Wikipedia, 2025](#)), and the suffragette Leonora Cohen, over 4,500 a month ([Wikipedia, 2025](#)). We have also been able to add some images to Wikimedia Commons, which can now be re-used in Wikidata and Wikipedia, and in SPARQL generated image grids related to the project.

The global Wikimedia community enables collaboration beyond institutional boundaries, particularly on topics supported by established WikiProjects. The *Women In Red* project⁸ seeks to reduce systemic gender bias on Wikipedia by increasing the visibility of notable women who lack articles ([Wikipedia, no date](#)). The project maintains a list of women suffragists who do not yet have a Wikipedia article⁹ to which we have been able to contribute a number of names discovered through the Suffrage Interviews project. There may be future opportunities to create biographical pages for these under-represented women in collaboration with LSE curators or *Women in Red* volunteers, and a colleague has suggested these 'hidden' women could be valuable candidates for inclusion in the *Dictionary of National Biography*, offering another avenue for recognition and scholarly engagement.

Impact

As we draw closer to the end of the project, demonstrable qualitative impacts have included:

- Strengthening Wikimedia skills within the team.
- Sharing the Suffrage Interviews metadata for re-use on a globally accessible, open platform.
- Creating linked data connections between LSE metadata and existing Wikidata entities.
- Enabling new ways to explore the collection through SPARQL queries and visualisations.

⁸ https://en.wikipedia.org/wiki/Wikipedia:WikiProject_Women_in_Red

⁹ https://en.wikipedia.org/wiki/Wikipedia:WikiProject_Women_in_Red/Missing_articles_by_occupation/Suffragists

- Enhancing discovery for potential users through Wikimedia platforms.
- Sharing project outputs via our new Digital Scholarship webpage¹⁰

For the conflict-of-interest reasons outlined above the project did not centre on active promotion of LSE content via Wikipedia, so investigating quantitative outcomes has not been the main focus of assessing project impact. While it is useful to explore whether any early conclusions can be drawn, several important caveats apply:

- Due to Google Analytics changes we are only able to access usage data for the main interviews page back to April 2023 and for interview downloads back to June 2023. We began the project in April 2023 so 'before and after' comparisons are difficult.
- Initial basic metadata was uploaded to Wikidata in automated batches, but ongoing metadata enhancement of those QIDs took place across subsequent and overlapping months, again making comparisons difficult without a longer period of analytics data.
- Wider ongoing project work is still in progress as of summer 2025 so it is too early to draw firm conclusions from Analytics without longer term usage data from the completed project.
- Use of our systems tends to follow the academic year with drops during vacation periods.
- The collection is small and specialised.

This means that the analytics we currently have cannot be considered reliably indicative of long-term trends. Nevertheless, as an interim assessment, they suggest small but significant positive gains from the work:

- Active users over the last 8 months, compared to the 8 months during which we carried out initial data uploads and enhancements to Wikidata, have increased by 41%, suggesting more users have been attracted to the site and engaged with it in some way on arrival.
- Traffic coming to the suffrage interviews homepage from Wikipedia has increased by 179% in the last 12 months compared to the previous 12 months.

The Wikidata Suffrage Interviews project has opened up new possibilities for representing and connecting the Library's unique and distinctive collections. It illustrates the way in which moving beyond traditional and familiar cataloguing workflows can enable the exploration of innovative approaches to modelling, creating, and enriching metadata, exposing connections between related entities, and contributing metadata to a global linked data ecosystem. The project's iterative and experimental approach, on a collaborative platform, has resulted in access to, and visualisation of, the collection in ways that were not previously possible. It

¹⁰ <https://www.lse.ac.uk/library/research-support/digital-scholarship>

demonstrates how we can extend our metadata skillsets to support discovery and re-use, not only within our own systems but across the wider web of data.

AI statement

I used Microsoft Copilot to proof-read my final draft of this article, incorporating some of its suggestions to improve the clarity of particular paragraphs. I also prompted it to help me find quotes which could be appropriated for the article title.

References

- LSE Library (no date) *The Suffrage Interviews*. Available at: <https://www.lse.ac.uk/library/collection-highlights/the-suffrage-interviews> [Accessed 8 July 2025]
- Pankhurst, E. (1908) 'We are here not because we are law breakers; we are here in our efforts to become law makers'. *Women's suffrage*. Available at: <https://www.npg.org.uk/schools-hub/womens-suffrage-deeds-not-words> [Accessed 8 July 2025]
- Wikidata (no date) *Help: Sources*. Available at: <https://www.wikidata.org/wiki/Help:Sources> [Accessed 8 July 2025]
- Wikidata (no date) *Oral evidence on the suffragette and suffragist movements: the Brian Harrison interviews*. Available at: <https://www.wikidata.org/wiki/Q100380678> [Accessed 8 July 2025]
- Wikidata (no date) *The Women's Library LSESuffrageInterviewsProject*. Available at: <https://www.wikidata.org/wiki/Q117322976> [Accessed 8 July 2025]
- Wikidata (no date). *WikiProject_PCC_Wikidata_Pilot*. Available at https://www.wikidata.org/wiki/Wikidata:WikiProject_PCC_Wikidata_Pilot [Accessed 8 July 2025]
- Wikipedia (2025) *Page views analysis: Leonora Cohen*. Available at: https://pageviews.wmcloud.org/pageviews/?project=en.wikipedia.org&platform=all-access&agent=user&redirects=0&start=2024-07&end=2025-06&pages=Leonora_Cohen [Accessed 8 July 2025]
- Wikipedia (2025) *Page views analysis: Margaret Cole*. Available at: https://pageviews.wmcloud.org/pageviews/?project=en.wikipedia.org&platform=all-access&agent=user&redirects=0&start=2024-07&end=2025-06&pages=Margaret_Cole [Accessed 8 July 2025]
- Wikipedia (2025) *Page views analysis: National Union of Women's Suffrage Societies*. Available at: https://pageviews.wmcloud.org/pageviews/?project=en.wikipedia.org&platform=all-access&agent=user&redirects=0&start=2024-07&end=2025-06&pages=National_Union_of_Women%27s_Suffrage_Societies [Accessed 8 July 2025]
- Wikipedia (2025) *Page views analysis: Sybil Thorndike*. Available at: https://pageviews.wmcloud.org/pageviews/?project=en.wikipedia.org&platform=all-access&agent=user&redirects=0&start=2024-07&end=2025-06&pages=Sybil_Thorndike [Accessed 8 July 2025]
- Wikipedia (2025) *Page views analysis: Women's Institute*. Available at: https://pageviews.wmcloud.org/pageviews/?project=en.wikipedia.org&platform=all-access&agent=user&redirects=0&start=2024-07&end=2025-06&pages=Women%27s_Institute [Accessed 8 July 2025]
- Wikipedia (no date) *Wikipedia:WikiProject_Women_in_Red*. Available at: https://en.wikipedia.org/wiki/Wikipedia:WikiProject_Women_in_Red [Accessed 8 July 2025]

- Wikipedia (no date) *Wikipedia:WikiProject_Women_in_Red/Missing articles by occupation/Suffragists*. Available at: https://en.wikipedia.org/wiki/Wikipedia:WikiProject_Women_in_Red/Missing_articles_by_occupation/Suffragists [Accessed 8 July 2025]
- 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://eprints.lse.ac.uk/114976/> [Accessed 8 July 2025]
- Williams, H. K. R. (no date) Wikidata:WikiProject LSEThesisProject. Available at: https://www.wikidata.org/wiki/Wikidata:WikiProject_LSEThesisProject [Accessed 8 July 2025]
- Williams, H. K. R. (no date) *Wikidata:WikiProject_The_Women's_Library_LSESuffrageInterviewsProject*. Available at: https://www.wikidata.org/wiki/Wikidata:WikiProject_The_Women's_Library_LSESuffrageInterviewsProject [Accessed 8 July 2025]

Envisioning Dante

a metadata journey from Inferno to Paradiso

Case study: *Le terze rime di Dante*: from Alma MARC (18873) to MDC TEI (PR-ALDI-18873)

Ourania Karapasias

Digital Metadata Specialist, Metadata & Discovery, Collection Strategies Directorate, The University of Manchester Library

Received: 11 August 2025 | Published: 22 September 2025

ABSTRACT

Through the case study of *Le terze rime di Dante* (PR-ALDI-18873), this article charts the progression from a MARC bibliographic record in Ex Libris Alma | Primo to an enriched TEI-XML file that forms a meaningful digital object in Manchester Digital Collections (MDC). The case study outlines the automated and editorial stages of the workflow, highlights key modelling decisions and challenges, and reflects on TEI's value as a flexible, expressive framework for describing digitised early printed books whose complexity exceeds the limits of standard bibliographic control. It demonstrates that while automation enables efficiency, human-led editorial intervention remains essential to produce accurate, semantically rich, and reusable metadata for sustainable digital collections.

KEYWORDS metadata transformations; early printed books; TEI-XML metadata; MARC to TEI; Dante Alighieri

CONTACT Ourania Karapasias ✉ ourania.karapasias@manchester.ac.uk 📍 The University of Manchester Library

On 29 May 2025, the first instalment of the Dante Digital Library¹ was published on Manchester Digital Collections (MDC)². This major digital resource features some of the rarest and most significant early printed editions of Dante Alighieri's *Divine Comedy*, many of which are being made available digitally for the first time. The release coincided with a two-day academic conference held at The John Rylands Research Institute and Library, home to the original volumes now featured in the digital collection.

The Dante Digital Library forms a central output of *Envisioning Dante, c. 1472–c. 1630: Seeing and Reading the Early Printed Page*, a digital humanities research project funded by the Arts and Humanities Research Council (AHRC) and based at The University of Manchester. Encompassing the digitisation of 99 editions printed between 1472 and

¹ <https://www.digitalcollections.manchester.ac.uk/collections/dante>

² <https://www.digitalcollections.manchester.ac.uk/>

1629, the project represents both a major scholarly resource and a significant advancement in metadata practices for early printed materials in digital collections. The first release features 20 editions, with further material scheduled for release after 2025.

Beyond its scholarly contributions, the Dante Early Printed collection³ within the Dante Digital Library also served as a platform for experimentation and refinement in metadata transformation. At the heart of this work lies the transformation of MARC 21 records into structured TEI-conformant files, aligning with FAIR principles to support metadata that is *findable, accessible, interoperable* and *reusable*.

This article uses *Le terze rime di Dante* (PR-ALDI-18873) as a case study and examines that transformation in detail. It considers the technical workflow, editorial modelling, and key TEI decisions, and reflects on the challenges, outcomes, and possibilities for further development. Through this example, it demonstrates how TEI can serve as a flexible and semantically rich framework for representing the descriptive complexity of digitised early printed books beyond the reach of conventional bibliographic control.

From MARC to TEI: overview of the workflow

The process begins with a MARC 21 record catalogued in Alma, which is exported to MARCXML using MarcEdit. An automated script then generates a preliminary TEI-conformant file, based on a template aligned with MDC conventions. This TEI file is committed to the mdc-contents Bitbucket repository via GitHub Desktop, enabling version control and collaborative editing. A Python script enriches the file with references to digital images, utilising `<facsimile>`, `<surface>`, and `<graphic>`. The result, a complete digital object (facsimile files plus TEI metadata), is then uploaded to the MDC test environment for internal review. At this point, the TEI file represents the initial structured output of the conversion workflow. Following conversion, the TEI file undergoes editorial refinement in Oxygen XML Editor. A spreadsheet-based tool is used to review the output and apply targeted encoding adjustments in line with MDC's TEI guidelines, to produce a publication-ready file.

For brevity, from here on MARC 21 is referred to as MARC, and TEI-XML and TEI-conformant file(s) as TEI file(s).

Stage 1: Conversion

From Primo | Alma MARC to MDC | TEI: mapping metadata

The conversion stage involves the automated mapping of MARC fields to corresponding TEI elements. While MARC already limits semantic expression, the conversion process further flattens distinctions between types of bibliographic description and contributor roles. The following examples illustrate how semantic detail is lost or obscured in practice.

³ <https://www.digitalcollections.manchester.ac.uk/collections/danteearlyprinted>

General notes | MARC 500 → TEI <note>

In MARC, separate 500 fields are used to record discrete bibliographic observations. For example, collation, typographic features, or editorial responsibility, each appear “independently” in Alma and in Primo’s public interface⁴.

Description Signatures: a-z⁸ A-F⁸ G¹². Leaf l2 is a blank.
 A Lyonese counterfeit of the Aldine edition of 1502, omitting Aldus' preface.
 No named printer, printer's device, date or colophon.
 Imprint from UCLA Online Catalog.
 Printed in italic type; capital spaces with guide letters at the beginning of the three parts; no catchwords.
 Without preface or notes.
 On title-page verso: Lo'nferno e'l Purgatorio e'l Paradiso di Dante Alaghieri.
 Edited by Pietro Bembo.

500 __ |a Signatures: a-z⁸ A-F⁸ G¹². Leaf l2 is a blank.
500 __ |a A Lyonese counterfeit of the Aldine edition of 1502, omitting Aldus' preface.
500 __ |a No named printer, printer's device, date or colophon.
500 __ |a Imprint from UCLA Online Catalog.
500 __ |a Printed in italic type; capital spaces with guide letters at the beginning of the three parts; no catchwords.
500 __ |a Without preface or notes.
500 __ |a On title-page verso: Lo'nferno e'l Purgatorio e'l Paradiso di Dante Alaghieri. owner. |5 UkMaJRU
500 __ |a Edited by Pietro Bembo.

During conversion, these fields are merged into a single composite structure. In the TEI output, all notes are rendered within a single <note>, structured only by multiple undifferentiated <p>.

• **Note(s):**

Signatures: a-z⁸ A-F⁸ G¹². Leaf l2 is a blank.
 A Lyonese counterfeit of the Aldine edition of 1502, omitting Aldus' preface.
 No named printer, printer's device, date or colophon.
 Imprint from UCLA Online Catalog.
 Printed in italic type; capital spaces with guide letters at the beginning of the three parts; no catchwords.
 Without preface or notes.
 On title-page verso: Lo'nferno e'l Purgatorio e'l Paradiso di Dante Alaghieri.
 Edited by Pietro Bembo.

⁴ https://www.librarysearch.manchester.ac.uk/permalink/44MAN_INST/1lr7mpn/alma9912616334401631 The fonts used in Alma are not encoded in XML. They pertain to the visual presentation layer and are not part of the structured metadata. As such, they are not retained or represented in the XML outputs, which focus solely on the underlying bibliographic data.


```

<note>
<p>Signatures: a-z8 A-F8 G12. Leaf 12 is a blank.</p>
<p>A Lyonese counterfeit of the Aldine edition of 1502, omitting Aldus'
preface.</p>
<p>No named printer, printer's device, date or colophon.</p>
<p>Imprint from UCLA Online Catalog.</p>
<p>Printed in italic type; capital spaces with guide letters at the
beginning of the three parts; no catchwords.</p>
<p>Without preface or notes.</p>
<p>On title-page verso: Lo'nferno e'l Purgatorio e'l Paradiso di Dante
Alaghieri.</p>
<p>Edited by Pietro Bembo.</p>
</note>

```

Although syntactically valid, this structure reproduces the MARC display logic without reintroducing any semantic differentiation. Collation, publication context, and descriptive notes are flattened into generic prose, limiting interpretability and complicating downstream processes such as filtering, indexing. These notes require disaggregation and more detailed encoding in Stage 2.

Personal names and roles | MARC 100, 700 → TEI <author>|<editor>

In MARC, field 100 records the primary creator of a work, typically an author, while field 700 records additional individuals who have contributed in other capacities. Relator terms in subfield #e, such as #e editor, #e printer, or #e former owner, specify the nature of each individual's role. Subfield #5 indicates that the information applies only to a specific institutional copy. Together, these fields and subfields help identify who created, contributed to, or owned a given copy of a work.

Name	Dante Alighieri, 1265-1321, author. Manuzio, Aldo, 1449 or 1450-1515, associated name. Bembo, Pietro, 1470-1547, editor. Gabiano, Balthazard de, -approximately 1517, printer. Mequignon, Cavaliere, former owner. Soma. Luigi, Magioniere. former owner. Christie, Richard Copley, 1830-1901, former owner. Spencer, George John Spencer, Earl, 1758-1834, former owner. Fréret, Nicolas, 1688-1749, former owner. Nicolet, Jean Baptiste Thomas. former owner.
-------------	---

```
100 0_ |a Dante Alighieri, |d 1265-1321, |e author.
```

```
700 1_ |a Manuzio, Aldo, |d 1449 or 1450-1515, |e associated name.
```

```
700 1_ |a Bembo, Pietro, |d 1470-1547, |e editor.
```

```
700 1_ |a Mequignon, |c Cavaliere, |e former owner. |5 UkMaJRU
```

```
700 1_ |a Soma. Luigi, |c Magioniere. |e former owner. |5 UkMaJRU
```

700 1_ |a Christie, Richard Copley, |d 1830-1901, |e former owner. |5 UkMaJRU

700 1_ |a Spencer, George John Spencer, |c Earl, |d 1758-1834, |e former owner. |5 UkMaJRU

700 1_ |a Fréret, Nicolas, |d 1688-1749, |e former owner. |5 UkMaJRU

700 1_ |a Nicolet, Jean Baptiste Thomas. |e former owner. |5 UkMaJRU

However, MARC's structure limits copy-level specificity. When multiple former owners are recorded using subfield †5, such as †5 UkMaJRU, it implies that all are associated with the same institutional copy. If more than one copy of the work exists, MARC cannot reliably indicate which name corresponds to which copy, as in the case of copy R213785.

In TEI the **<author>** is reserved for primary authorship derived from MARC field 100. All other contributors from field 700 are rendered as **<editor>**, with **@role** preserving the relator term through its corresponding relator code: **role="prt"** for printer, **role="fmo"** for former owner.

- **Author(s):** Dante Alighieri, 1265-1321
- **Editor(s):** Bembo, Pietro, 1470-1547
- **Printer(s):** Gabiano, Balthazard de, -approximately 1517

—

- **Former Owner(s):** Mequignon, Cavaliere; Soma. Luigi, Magioniere.; Christie, Richard Copley, 1830-1901; Spencer, George John Spencer, Earl, 1758-1834; Fréret, Nicolas, 1688-1749; Nicolet, Jean Baptiste Thomas.

—

- **Associated Person(s):** Manuzio, Aldo, 1449 or 1450-1515; Gabiano, Balthazard de, -approximately 1517

```
<author key="person_v00000000" ref=" http://viaf.org/viaf/00000000">Dante Alighieri, 1265-1321, author.</author>
<editor role="asn">Manuzio, Aldo, 1449 or 1450-1515</editor>
<editor role="edt">Bembo, Pietro, 1470-1547</editor>
<editor role="prt">Gabiano, Balthazard de, -approximately 1517</editor>
<editor role="prt">Gabiano, Balthazard de, -approximately 1517</editor>
<editor role="fmo">Mequignon, Cavaliere</editor>
<editor role="fmo">Soma. Luigi, Magioniere.</editor>
<editor role="fmo">Christie, Richard Copley, 1830-1901</editor>
<editor role="fmo">Spencer, George John Spencer, Earl, 1758-1834</editor>
<editor role="fmo">Fréret, Nicolas, 1688-1749</editor>
<editor role="fmo">Nicolet, Jean Baptiste Thomas.</editor>
```

This results in a semantic collapse: editorial, physical production, and provenance roles are conflated under `<editor>`, an element conventionally associated with editorial responsibility. While encoding printers as `<editor role="prt">` is justifiable in some contexts, extending the same structure to former owners and associated names conflates fundamentally distinct roles.

Although accepted in Stage 1 as a pragmatic interim measure, this approach has clear limitations. In Stage 2, only roles relevant to this copy are retained. Former owners will be encoded using more appropriate elements, such as `<provenance>` within `<history>`, while references to other copies will be removed.

The absence of authority-controlled identifiers, such as VIAF IDs, limits the file's capacity for disambiguation and interoperability. Where available, they will be added in later stages to support consistent, reusable metadata.

These two examples exemplify broader challenges in automated conversion: flattened descriptive content, ambiguous role encoding, and copy-specific metadata collapse into generic structures. Related MARC fields such as **561**, **562**, and **563** are mapped to high-level TEI wrapper elements such as `<provenance>`, `<physDesc>`, and `<bindingDesc>`, each containing generic `<p>`.

From syntax to semantics

At first glance, the TEI file derived from the MARC record appears to be a successful conversion. The XML is valid, well-formed, and legible. However, this surface-level success conceals a deeper issue: the file conforms to XML syntax but reproduces the logic of MARC rather than engaging TEI's modelling potential. Descriptive richness is often lost in translation.

In effect, the TEI file reproduces the structure of the Alma catalogue record. It presents MARC-encoded metadata reformatted using TEI elements, without engaging TEI's underlying model. Although the XML is technically sound, it lacks the interpretative structure that gives TEI its descriptive power. XML provides structural rules; TEI offers a modelling vocabulary that defines how parts of a resource function and relate. This supports meaningful, contextual representation.

TEI is not just a structural wrapper for metadata. It is a modelling language designed to express not only what a source is, but also how and why it holds meaning in historical, material, and intellectual terms. The automated conversion follows the letter of the standard but not its intent. Valid XML does not, in itself, constitute valid TEI in any intellectually rigorous sense. The challenge lies not in conversion but in interpretation. It is not a matter of syntax, but of semantics.

Stage 2: Transformation

TEI editorial refinement and semantic modelling

The transformation stage introduces both standardisation and interpretive encoding. At this point, MARC-derived metadata is aligned with TEI not only structurally but also conceptually. The focus shifts from syntax to semantics, from automated output to expert-led modelling.

Editorial refinement addresses structure, terminology, and descriptive content to ensure clarity, consistency, and accuracy. It corrects misaligned mappings, applies normalisation rules, and disaggregates descriptive, structural, and copy-specific information, ensuring that each element conveys precisely the relationship it is meant to express.

A key challenge identified in Stage 1 was the generic mapping of MARC fields to high-level TEI wrapper elements, which prompted the remapping of fields to more semantically appropriate TEI elements and the re-encoding of their content to accurately reflect the structure and meaning of the source. This work was underpinned by a spreadsheet-based mapping tool. It captures element-level markup, editorial rationale, and proposed schema enhancements, while supporting metadata enrichment through authority control and linked data.

The following examples revisit the same cases presented in Stage 1, repeated here to illustrate how they are structurally and semantically transformed during Stage 2.

Personal names and roles | MARC 100, 700 → TEI <author>|<editor>|<provenance>

Stage 2 of the editorial workflow introduces a clearer, semantically grounded treatment of names imported from MARC records. In pre-editorial TEI, personal names, regardless of their role, were rendered as <editor>, distinguished only by relator codes such as **edt** (editor), **prt** (printer), **asn** (associated name), or **fmo** (former owner). This flattening obscured meaningful distinctions and often introduced inconsistencies, including duplication.

Contributor roles are disaggregated by assigning each name to the appropriate TEI element based on function: the author is re-encoded using <author>, while only printers and editors are encoded using <editor> but are distinguished using specific **@role** values. VIAF identifiers are added via **@key** to support authority control and disambiguation.

- **Author(s):** Dante Alighieri, 1265-1321
 - **Editor(s):** Bembo, Pietro, 1470-1547
 - **Printer(s):** Gabiano, Balthazard de, -approximately 1517
-
- **Former Owner(s):** Nicolet, Jean Baptiste Thomas; Fréret, Nicolas, 1688-1749; Spencer, George John Spencer, Earl, 1758-1834; Spencer, John Poyntz Spencer, Earl, 1835-1910; Rylands, Enriqueta, 1843-1908
-
- **Associated Person(s):** Manuzio, Aldo, 1449 or 1450-1515
-
- **Provenance:**
 - Jean Baptiste Thomas Nicolet
Bookplate on the inner back cover: *Ex Libris Joannis Baptistæ Thomæ Nicolet.*
 - Nicolas Fréret (1688-1749), historian and linguist
Possibly Nicolas Fréret. Manuscript inscription on the upper right corner of a1r: *N. Freret | # | 1710.*

```

<author key="viaf:97105654">Dante Alighieri, 1265-1321</author>
<editor key="viaf:54144140" role="edt">Bembo, Pietro, 1470-1547</
editor>
<editor role="prt">Gabiano, Balthazar de, -approximately 1517</editor>
<editor key="viaf:96325760" role="asn">Manuzio, Aldo, 1449 or 1450-
1515</editor>
<provenance>
<p> <name role="fmo"><persName type="display">Jean Baptiste Thomas
Nicolet</persName><persName type="standard">Nicolet, Jean Baptiste
Thomas</persName></name></p>
<p>Bookplate on the <locus from="Inner_back_cover" to="Inner_back_
cover">inner back cover</locus>: <hi rend="italic">Ex Libris Joannis
Baptistæ Thomæ Nicolet</hi>.</p>
</provenance>
<provenance>
<p><name key="viaf:34456779" role="fmo"><persName
type="display">Nicolas Fréret (1688-1749), historian and linguist</
persName><persName type="standard">Fréret, Nicolas, 1688-1749</
persName></name></p>
<p>Possibly Nicolas Fréret. Manuscript inscription on the upper right
corner of <locus from="a1r" to="a1r">a1r</locus>: <hi rend="italic">Ex
N. Freret | # | 1710 </hi>.</p>
</provenance>

```

—

Former owners, previously conflated with contributors, are now correctly encoded in **<provenance>**, using **<name role="fmo">** with structured **<persName>** content.

Physical traces, such as bookplates, inscriptions, and annotations, are anchored via `<locus>` and `<hi>`, allowing evidence of ownership to be linked precisely to locations within the item. For example, Jean Baptiste Thomas Nicolet is associated with a bookplate on the inner back cover, while Nicolas Fréret is linked to a marginal inscription on folio a1r. These granular encodings preserve the material history of the object without distorting intellectual attribution.

General notes | MARC 500 → TEI `<note>` | `<physDesc>` | `<additions>`

A broad range of information, such as collation, edition history, typographic features, and editorial attributions, was previously grouped within a single `<note>` and embedded in `<physDesc>`, expressed through multiple `<p>`. This content is now distributed and re-encoded across semantically appropriate TEI elements.

General remarks about the edition's character, such as its identification as a Lyonese counterfeit of the Aldine edition, are retained in `<note>`, but only when they serve as interpretive or contextual commentary rather than formal description.

- **Note(s):**

A Lyonese counterfeit of the Aldine edition of 1502, omitting Aldus' preface.

```
<note>
```

```
<p>A Lyonese counterfeit of the Aldine edition of 1502, omitting Aldus' preface.</p>
```

```
<note>
```

Within `<physDesc>`, collation details such as signature sequences and blank leaves are encoded to `<collation>`; layout and line count are expressed using `<layoutDesc>`; and typographic features, including typeface and guide letters, are captured in `<typeDesc>` and `<decoDesc>`. Where applicable, controlled vocabulary terms, such as material types from the Getty AAT identifiers, are added via `@key` to improve semantic precision and support interoperability.

- **Format:** Codex

- **Material(s):** Paper

- **Extent:** 244 leaves

- **Collation:** Signatures: a-z⁸ A-F⁸ G¹²
Leaf l2 is blank.

- **Layout:**

Printed in one column, 28-30 lines.

- **Typeface:**

- Printed in *italic type*.

- **Decoration:**

Blank spaces for initials are marked with printed guide letters.


```

<physDesc>
<objectDesc form="codex">
<supportDesc material="paper">
<support><material key="aat:300014109">Paper</material>, folded in
<measure type="folding">inner 8vo (octavo)</measure>.</support>
<extent><measure quantity="244" type="leaf">244</measure> leaves</
extent>
<collation>Signatures: <signatures>a-z<hi rend="superscript">8</hi> A-
F<hi rend="superscript">8</hi> G<hi rend="superscript">12</hi></
signatures>
<p>Leaf <locus from="Blank_folio_[l2r]" to="Blank_folio_[l2r]">l2</
locus> is blank.</p>
</collation>
</supportDesc>
<layoutDesc>
<layout columns="1">
<p>Printed in one column, 28-30 lines.</p>
</layout>
</layoutDesc>
</objectDesc>
<typeDesc>
<typeNote scope="sole">Printed in <term>italic type</term>.</typeNote>
</typeDesc>
<decoDesc>
<decoNote type="initial">
<p>Blank spaces for initials are marked with printed guide letters.</
p>
</decoNote>
</decoDesc>

```

Copy-specific features, such as manuscript inscriptions and pencil annotations, previously embedded as multiple **<p>** within generic **<physDesc>** content, are now encoded in **<additions>**. The **<locus>** indicates their physical location within the item, while **<hi>** is used to preserve stylistic or visual emphasis.

- **Additions:**

Manuscript pencil note on flyleaf 1v: *Character of G. Huyon A | 1520?*.

Various manuscript pencil prices[?] on flyleaf 1v: £2.2.0 | 0 | 3/[?] | 1/4/2[?].

Manuscript inscription, mostly illegible, on the lower margin of a1r: [?]De[?]dras[?].

```

<additions>
<p>Manuscript pencil note on <locus from="Flyleaf_1v" to="Flyleaf_
1v">flyleaf 1v</locus>: <hi rend="italic">Character of G. Huyon A |
1520?</hi>.</p>

```

```
<p>Various manuscript pencil prices[?] on <locus from="Flyleaf_1v" to="Flyleaf_1v">flyleaf 1v</locus>: <hi rend="italic">£2.2.0 | 0 | 3/[?] | 1/4/2[?]</hi>.</p>
<p>Manuscript inscription, mostly illegible, on the lower margin of <locus from="a1r" to="a1r">a1r</locus>: <hi rend="italic">[?]De[?]dras[?]</hi>.</p>
</additions>
```

Through these interventions, the TEI file moves beyond formal validity to become a meaningful digital object in MDC, representing the physical item through metadata, displayed alongside its digital facsimile.

Reflections on metadata conversion and transformation

The *Envisioning Dante* project provided a timely opportunity to trial and refine a metadata workflow not bound by the structural constraints of MARC. Drawing on both automation and editorial expertise, it explored scalable methods for producing semantically rich, structurally coherent TEI files for publication in MDC as digital objects.

In Stage 1, an automated conversion model was used to generate a baseline TEI from MARC metadata. This reliable and repeatable workflow reduced the need for manual intervention and enabled efficient early-stage processing of a large number of files.

In Stage 2, the spreadsheet-based mapping tool guided encoding decisions and established a uniform encoding practice, ensuring consistency across approximately one hundred TEI files. Used collaboratively by team members, it proved effective, sustainable, and adaptable within a distributed editorial workflow. Its successful application demonstrated that it supported a practical, repeatable process, well-suited to both current and future projects. Beyond refining individual files, the tool also contributed to the development of the MDC TEI guidelines, both informing and being shaped by editorial decisions grounded in practical encoding challenges.

TEI files were further enriched with persistent identifiers (e.g. VIAF, Getty TGN, and AAT), refined encoding of names and roles, and additional copy-specific descriptive features such as bindings, provenance, and annotations. These enhancements strengthened authority control and improved interoperability with linked data frameworks.

As part of this work, additional TEI elements were introduced into our practice to support the description of printed materials. Elements such as **<typeDesc>** and **<typeName>** were used to encode and display as *'Typeface'*, in parallel with **<handDesc>** and **<handNote>** for manuscripts displaying as *'Script'*.

A customised `<note type="printersDevice">` was added to encode and display as *'Printer's Device'*, enabling the description of devices used by printers in early printed books.

To support the structured encoding of bibliographic formats such as *'folded in 8vo (octavo)'*, which is currently recorded using `<measure type="folding">` within `<support>`, a proposal was submitted to the TEI community to extend the schema to allow `<objectType>` as a valid child of `<objectDesc>`. This change should enable more precise representation of an item's functional or semantic category, along with its physical form.

In addition, the customised element `<note type="publisher">`, introduced earlier in local practice, was used to encode and display as *'Publication'*, supporting the representation of publication, distribution, and related information (imprint), as recorded in MARC field 260 or 264.

The `<filiation>` was also employed to record bibliographic relationships between related printed editions. In this case, linking to the Aldine copy R213785 to reflect shared printing history or edition lineage.

Despite these successes, significant challenges emerged when working with MARC-derived metadata. Records created under varying cataloguing standards and levels of detail often exhibited semantic ambiguity, inconsistent field usage, and limited granularity. These issues frequently hindered automation and required careful editorial judgement to harmonise descriptive content and ensure alignment with MDC's standards for display and discovery.

This became particularly evident during the processing of the case study file, *PR-ALDI-18873*⁵. As the first TEI file generated, it included minimal descriptive content, making it relatively straightforward to convert and helping to establish a practical benchmark for the workflow. By contrast, later files presented richer but more ambiguous content that could not always be confidently interpreted. This highlighted the value of working with colleagues who have expertise in early printed books. Future projects would benefit from deeper cross-disciplinary engagement to address uncertainty and support the creation of more consistent, higher-quality TEI files.

Due to its bibliographic orientation, MARC's flat, field-based structure often lacks the contextual depth required for the kind of expressive, semantically rich markup supported by TEI. Such a structure obscures relationships between entities, restricts item-level attribution, and excludes interpretative elements such as meaningful summaries or tables of contents. These limitations reinforce the need for human-led editorial intervention to produce metadata that is semantically precise, structurally coherent, and optimised for digital presentation and discovery.

⁵ <https://www.digitalcollections.manchester.ac.uk/view/PR-ALDI-18873/1>

A hybrid approach, combining automation, editorial expertise, and collaborative practice, can deliver high-quality, non-MARC metadata at scale. The editorial model developed in this project offers a practical foundation for future initiatives aiming to produce TEI metadata that meets both structural and descriptive standards and supports long-term discovery and use of early printed books in digital collections.

Conclusion

Through the lens of a case study, this article has examined the evolving potential of TEI to enhance and contextualise MARC-based metadata for digitised early printed books. TEI's semantic flexibility, extensibility, and use of structured XML designed for human interpretation make it particularly well suited to describing materials that exceed the descriptive and structural limits of MARC.

The Library's public-facing interfaces, Primo and MDC, manifest the pivot from MARC's catalogue-centred logic to TEI's semantically driven model. Primo draws on MARC records to support catalogue search, inventory management, and authority control. These are tasks aligned with traditional library workflows. However, MARC is not designed to deliver the layered, interpretative access required for digitised early printed books.

Transforming MARC into TEI entails more than field mapping. It requires interpretation, structural modelling, and content enrichment to ensure that the resulting metadata is accurate, interoperable, and suitable for digital presentation and discovery. In short, MARC provides bibliographic data, whereas TEI enables it to be expressed as meaningful, structured, and contextualised digital object metadata.

The MDC Dante items, comprising both facsimile images and their corresponding TEI files, now benefit from richer, semantically enhanced descriptions. This pairing of object and metadata reflects MDC's core definition of a digital object: a unified entity that enables access, interpretation, and future reuse.

Working across the full set of records revealed the limits of automation and the importance of flexibility in metadata creation. While defined workflows are necessary, they must be adaptable enough to accommodate the descriptive complexity of early printed books. Moving beyond MARC requires more than conversion. It is a process shaped by interpretation, editorial judgement, and contextual understanding.

Emerging AI tools are beginning to reshape the metadata landscape. Steven Hartshorne (2025) recently conducted a small-scale experiment that used ChatGPT to enhance existing MARC records. The trial suggests that AI can streamline certain metadata upgrade tasks by providing lightweight suggestions that still require human input. Crucially, the semantic depth, structural complexity and editorial nuance of TEI continue to demand interpretive decisions that current AI cannot match.

Our metadata journey echoes Dante's ascent from *Inferno to Paradiso*: moving from inherited constraints toward a realm of expressive possibility. The progression from initial conversion to enriched transformation was shaped by human judgement, editorial expertise and care. At its core, the process remains interpretative and constructive and, for the foreseeable future, "*unmistakably*" human.

Both PR-ALDI-18873 (<https://www.digitalcollections.manchester.ac.uk/view/PR-ALDI-18873/1>) and PR-ALDI-R-00002-13785 (<https://www.digitalcollections.manchester.ac.uk/view/PR-ALDI-R-00002-13785/1>) are now accessible for viewing on MDC.

References

Hartshorne, Steven (2025) Manipulating Rare Print Metadata with ChatGPT. *Catalogue & Index*, 211, pp. 9-19. Available at: <https://journals.cilip.org.uk/catalogue-and-index/article/view/748>

Decolonising heritage collections

The People's Collection Wales Toolkit

Jessica Roberts

People's Collection Wales

Received: 19 August 2025 | Published: 22 September 2025

ABSTRACT

This article presents the Decolonisation Toolkit created by People's Collection Wales (PCW), a community-driven digital heritage platform. Released in 2025, the Toolkit provides structured guidance for identifying, contextualising, and addressing discriminatory and colonial language in both English and Welsh metadata. Core features include website-level and item-level content warnings, transparent audit trails, and practical workflows for reviewing legacy material and preparing new uploads. Drawing on case studies – including historic newspapers, Penrhyn Castle, minstrel show photographs, and community contributions – the article explores how the Toolkit supports inclusive description while maintaining accuracy and public trust. It situates PCW's approach within wider debates about non-MARC metadata, critical description, and community participation.

KEYWORDS decolonisation; metadata; inclusive description; community archives; digital heritage

CONTACT Jessica Roberts ✉ jessica.roberts@museumwales.ac.uk 📍 People's Collection Wales

Introduction

People's Collection Wales (PCW) is a bilingual platform enabling individuals, local groups, and institutions to upload material and describe it with metadata. This openness enriches the national record but introduces challenges: harmful terms appear in historic collections, and contributors risk reproducing outdated language in new uploads.

To address this, PCW developed the Decolonisation Toolkit with input from the Cultural Heritage Terminology Network and a Task and Finish Group of heritage professionals. The Toolkit aligns with PCW's Charter for Decolonising the Collection ([People's Collection Wales, 2025](#)) and the Welsh Government's Anti-Racist Wales Action Plan ([Welsh Government, 2022](#)), ensuring that descriptive practice responds to ethical and cultural responsibilities.

This article will examine the key principles and practical guidance of the PCW Decolonisation Toolkit¹, using specific case studies to demonstrate how a community-driven approach can effectively address harmful metadata in a digital heritage context.

Key Principles

The Toolkit defines decolonisation as a diagnostic and dismantling process distinct from general equality and diversity agendas. It involves acknowledging colonial legacies in heritage collections, recognising that metadata choices reflect power and perspective, and shifting descriptive practices to centre historically marginalised voices ([People's Collection Wales, 2025](#)). Decolonisation is positioned as a continuous effort rather than a one-off corrective, reflecting the evolving nature of language, identity, and social awareness.

Toolkit Guidance

The Toolkit offers a comprehensive set of measures to support contributors and moderators. It begins by introducing a site-wide content warning that acknowledges the likelihood of encountering harmful language, supported by item-level warnings that alert users to offensive terms or imagery in specific records. These warnings are not designed to censor but to prepare audiences for what they may see or read, enabling informed engagement ([People's Collection Wales, 2025](#)).

Equally important is the audit trail system, which ensures that changes to metadata are fully documented. Every alteration is recorded with details of what was changed, when it happened, and why. This not only enhances transparency but also demonstrates the iterative process of learning, revising, and improving. Rather than erase outdated interpretations, PCW encourages users to preserve them within the audit trail to retain their historical value while also correcting or contextualising them for modern readers.

In practice, contributors are encouraged to examine their own items carefully. Harmful terminology can be contextualised with additional explanation or replaced with language that respects dignity and accuracy. The Toolkit points to the Inclusive Terminology Glossary² as an invaluable resource in this process (Cultural Heritage Terminology Network, 2024). New uploads are subject to the same expectations: contributors are asked to reflect on provenance, review their descriptive text, and consult relevant communities when necessary. Because PCW is bilingual, the Toolkit highlights the need for Welsh-language equivalents to ensure parity across both languages, acknowledging that linguistic nuances may affect how sensitive terms are translated or reframed.

¹ <https://www.peoplescollection.wales/sites/default/files/documents/Final%20Eng%20Decolonisation%20Toolkit%202025.pdf>

² <https://culturalheritageterminology.co.uk/>

Finally, the Toolkit reinforces the collaborative nature of PCW. Contributors and users are not passive actors but active participants in shaping an inclusive collection. They are invited to report offensive content, revisit their own contributions, and reuse items in ways that tell more inclusive and accurate stories.

Case Studies

The Toolkit sets out its approach through a series of practical examples that show how inclusive description works in practice. One early example is a 1918 newspaper clipping containing the derogatory term “Chinaman” as seen in Figure 1.

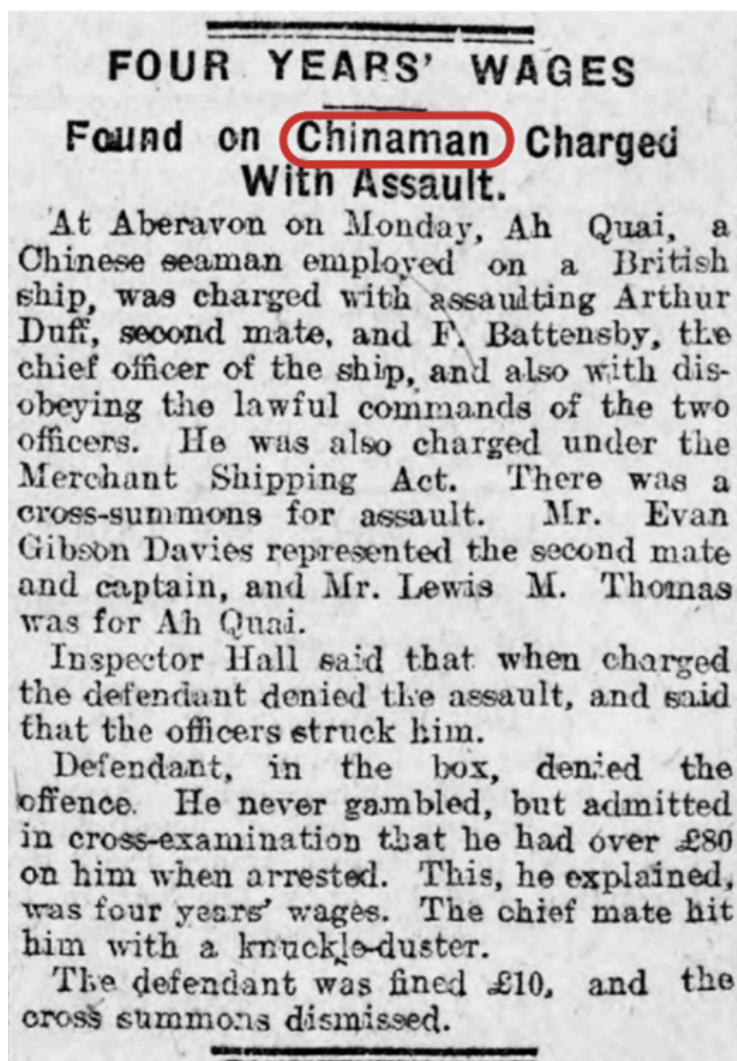


Figure 1. “Four Years’ Wages” (The Cambria Daily Leader, 1918).

Rather than remove the item, PCW retained it for accuracy but added a clear content warning and a contextual explanation from the Inclusive Terminology Glossary noting that the “term ‘Chinaman’ is an archaic 18th/19th century term for Chinese people, which is widely considered derogatory today...” ([People’s Collection Wales, 2025](#)). This ensures that the historic record remained accessible, while also preparing readers to recognise and understand the problematic language. In this case, staff flagged the item, consulted with contributors, and reviewed glossary entries before agreeing the final approach.

The decision - to retain but contextualise - was recorded in the audit trail. Researchers and users now encounter not only the historic source but also a linked explanation of racial terminology in the early twentieth century.

A second example extends this approach by showing how entire descriptions can be revised to centre lived experience. For the item relating to Penrhyn Castle, contributed by the Royal Commission on the Ancient and Historical Monuments of Wales, the original description focused on the building's architectural features. Its historical account referred only briefly to how Richard Pennant profited from the enslavement of Caribbeans:

“...Pennant himself had married into the Penrhyn family and had subsequently made his fortune through slate quarrying industries in north Wales and slavery in Jamaica” ([People's Collection Wales, 2025](#), p.21).

The use of the term *slavery* is problematic because it abstracts and dehumanises those affected. This vague phrasing was replaced with terminology that acknowledges people directly - “he enslaved” - and the description was expanded to explain Pennant's role in the transatlantic slave trade. The new entry also links to the National Trust's wider research on Penrhyn Castle³ and includes a note documenting the changes. An item-level content warning was added at the same time, preparing readers for potentially harmful terminology and signalling that the record contains sensitive material. The revised version reads:

“...Pennant himself had married into the Penrhyn family and had subsequently made his fortune through slate quarrying industries in north Wales and sugar plantations in Jamaica. He enslaved nearly 1,000 people across his four plantations. Hay-Dawkins Pennant was also an owner of enslaved people and received £14,683 from the government on abolition. (For more information on how wealth gained from the transatlantic slave trade was used to build Penrhyn Castle, visit the National Trust's website: <https://www.nationaltrust.org.uk/visit/wales/penrhyn-castle-andgarden/penrhyn-castle-and-slave-trade-history>.)” ([People's Collection Wales, 2025](#), p.22)

The corresponding audit trail note reads:

“This description was updated in March 2024. As per the Inclusive Terminology Glossary (1.1. African American History and the Atlantic Slave Trade: <https://docs.google.com/document/d/1JaJ8VchUCbtg7jPmhwizOQYsabBqKLxZ7n69urQS8VM>), the word ‘slavery’ was replaced by a sentence detailing how and where Pennant and his cousin enslaved people to accrue wealth, and a link to the National Trust website was

³ <https://www.nationaltrust.org.uk/visit/wales/penrhyn-castle-and-garden/penrhyn-castle-and-slave-trade-history>

added for further information about the Castle's colonialist history."
([People's Collection Wales, 2025](#), p.23)

This fuller treatment aligned with best practice in inclusive metadata. It shifted the focus away from abstract institutions and towards the people whose labour and freedom were taken, embedding historical accountability directly into the record. The combined use of item-level content warning and audit trail note ensured transparency while also protecting users from unexpected exposure to harmful terms.



Figure 2. "Churchill's Minstrels" (Conwy Archive Service, 1920s)

A third example demonstrates how visual records can be reinterpreted. As shown in Figure 2, Conwy Archive Service had contributed a photograph of Churchill's Minstrels at the Happy Valley Theatre, Llandudno, in the 1920s that was part of a 'Holidays in Llandudno' slideshow with the description:

"Happy Valley had a succession of theatres of increasing size that held huge audiences. Here we see Churchill's Minstrels in the 1920s." ([People's Collection Wales, 2025](#), p.25)

While factually correct, this description did not address its problematic content. PCW worked with the archive to revise the record so that it acknowledges both the historical popularity of minstrel shows and their reliance on racist stereotypes and blackface. The revised entry reads:

"This image is from the 1920s. It shows the Churchill's Minstrels, a minstrel group set up by Will Churchill in c.1906-07. Minstrel shows were a form of popular entertainment from the early 19th century until well into the 1970s. These shows mostly involved white male performers in blackface. While they had a distinct impact on popular music, dance and

other aspects of popular culture, they were founded on the comic enactment of racist stereotypes and are now considered to be exploitative and racially offensive.” ([People’s Collection Wales, 2025](#), p.25)

By adding historical explanation and critical framing, the new description educates readers about the wider history of minstrel shows while also signalling how perspectives have changed. The phrase “now considered to be exploitative and racially offensive” makes clear that, while these performances were accepted by white audiences of the time, they are understood very differently today. To prepare this revision, PCW staff researched best practice, consulted reference sources, and tested language choices carefully, ensuring the record reflected both historical context and ethical responsibility.

Together, these case studies show the Toolkit in action: not simply substituting words, but reshaping records to balance historical accuracy with inclusivity, transparency, and accountability.

Expanded Guidance in Practice

The Toolkit goes further by outlining detailed workflows to support this work. For item titles, where wording reflects an original title or historic language, it recommends placing harmful terms in quotation marks with replacements in brackets. This allows the original phrasing to remain searchable while clarifying its meaning for contemporary audiences. Item descriptions are expanded with contextual paragraphs that explain problematic language or imagery, often supplemented with references to external sources such as the Inclusive Terminology Glossary. By encouraging contributors to link to trusted resources, PCW promotes ongoing learning beyond the record itself. Equally, the Toolkit stresses that bilingual consistency is essential. Updates in English must also appear in Welsh, and translation should go beyond direct equivalence to reflect cultural context and nuance.

Challenges

Despite its structured framework, the Toolkit faces challenges. Some audiences view decolonisation as intrusive interpretation or even ‘censorship,’ while others have reacted negatively on social media. PCW acknowledges these criticisms but stresses that updating records is an established part of archival practice. The aim is not to erase history but to make it more accurate, transparent, and inclusive, while continuing to challenge racism and prejudice within the Collection. The need is clear: the 2018–19 National Survey for Wales found that 76% of respondents from ethnic minority backgrounds did not take part in arts, culture, or heritage activities, highlighting the barriers that remain ([Welsh Government, 2022](#)).

Implications for Metadata Practice

The PCW model, which relies on Dublin Core with free-text fields, demonstrates that decolonising practices can be embedded outside MARC-based systems. The Toolkit shows the importance of participatory approaches that empower communities and contributors to influence metadata standards. It also highlights the value of inclusive glossaries, which guide terminology decisions in an area where language continues to evolve ([Cultural Heritage Terminology Network, 2024](#)). Audit trails are a further addition, providing a method to record changes while preserving interpretive history. The use of website- and item-level content warnings has increased trust among users by making the risks of harmful content explicit. Together, these practices position PCW within the broader landscape of critical metadata and community archiving.

Future Plans

Looking ahead, PCW intends to expand the training offered to contributors, volunteers, and partner institutions to ensure the Toolkit is widely understood and effectively applied. PCW also aims to publish new case studies, sharing lessons with other organisations in Wales and further afield. Finally, evaluation will become a priority: tracking search success, user engagement, and discoverability to assess how inclusive metadata affects access and use of the collection.

Conclusion

The Decolonisation Toolkit provides a structured and replicable model for community-driven decolonisation of metadata. By embedding content warnings, audit trails, glossary support, and community involvement into routine description practice, PCW shows that inclusive records can be achieved without compromising transparency or historical accuracy. The Toolkit demonstrates that addressing harmful language is not simply a technical adjustment but a cultural and ethical responsibility. In doing so, PCW contributes to a more responsible, accountable, and equitable record of Wales's heritage.

References

- The Cambria Daily Leader (1918) *"Four Years' Wages"*, 19 February, p. 2. Available at: <https://www.peoplescollection.wales/items/1184456> [Accessed: 18 August 2025]
- Churchill's Minstrels, Happy Valley theatre, Llandudno* (ca. 1920-1929) Available at: <https://www.peoplescollection.wales/items/1862571> [Accessed: 19 August 2025]
- Collections Trust (2023) *Spectrum Standard*. Available at: <https://collectionstrust.org.uk/spectrum/> [Accessed: 19 August 2025]
- Cultural Heritage Terminology Network (2024) *Inclusive Terminology Glossary*. Available at: <https://culturalheritageterminology.co.uk/> [Accessed: 18 August 2025]
- People's Collection Wales (2025) *Charter for Decolonising People's Collection Wales*. Available at: <https://www.peoplescollection.wales/content/charter-decolonising-peoples-collection-wales> [Accessed: 18 August 2025]

People's Collection Wales (2025) *Decolonisation Toolkit*. Available at: <https://www.peoplescollection.wales/sites/default/files/documents/Final%20Eng%20Decolonisation%20Toolkit%202025.pdf> [Accessed: 18 August 2025]

Welsh Government (2022) *Anti-Racist Wales Action Plan*. Available at: <https://www.gov.wales/anti-racist-wales-action-plan> [Accessed: 18 August 2025]

Unchartered cells

Cataloguing with Excel in Covid and beyond

Carol Hunter

Metadata and Authority Control Manager, National Library of Scotland

Received: 31 August 2025 | Published: 22 September 2025

ABSTRACT

Cataloguer training takes time, patience and a lot of learning – from different standards and procedures depending on the material to grappling with different Library Management Systems and all the jargon associated.

When Covid hit, the recently digitised collection of HMSO catalogues became an ideal working from home project. Each listing had adequate information to make a catalogue record to allow these partially hidden items to be discoverable. The problem was that cataloguers at the National Library of Scotland only had their personal devices to begin with and no access to our usual software. Plus, work was also required for Library staff who were not trained cataloguers.

The solution – Excel spreadsheets. A friendly and familiar format where data entered in an organised fashion could then be imported into Alma (our Library Management System) to make items available. Users would be able to create Metadata quickly with minimal training and no knowledge required about MARC or RDA standards.

What started as a Covid solution has now become common place for specific routine metadata projects for staff without in-depth cataloguing training.

KEYWORDS data entry; spreadsheets; delimited data to MARC; Alma; normalisation rules

CONTACT Carol Hunter  c.hunter@nls.uk  National Library of Scotland

Cataloguing Training

Learning to catalogue takes time and can be an overwhelming process if rushed. There are various standards and procedures to follow, from international to in-house, plus different material types and becoming familiar with complex Library Management Systems. It takes time and effort to train someone to the point where they can be let loose to describe Library collections.

But what happens when we cannot do this?

Adapting to Covid - Working from Home

When the Covid pandemic began in 2020, our cataloguers had never worked from home before. We were all based onsite with PCs and predominantly worked with physical items in hand. When we initially had to adapt to Working from Home, most staff just had their own personal devices and were not able to access our usual cataloguing systems.

So – how could we continue to work?

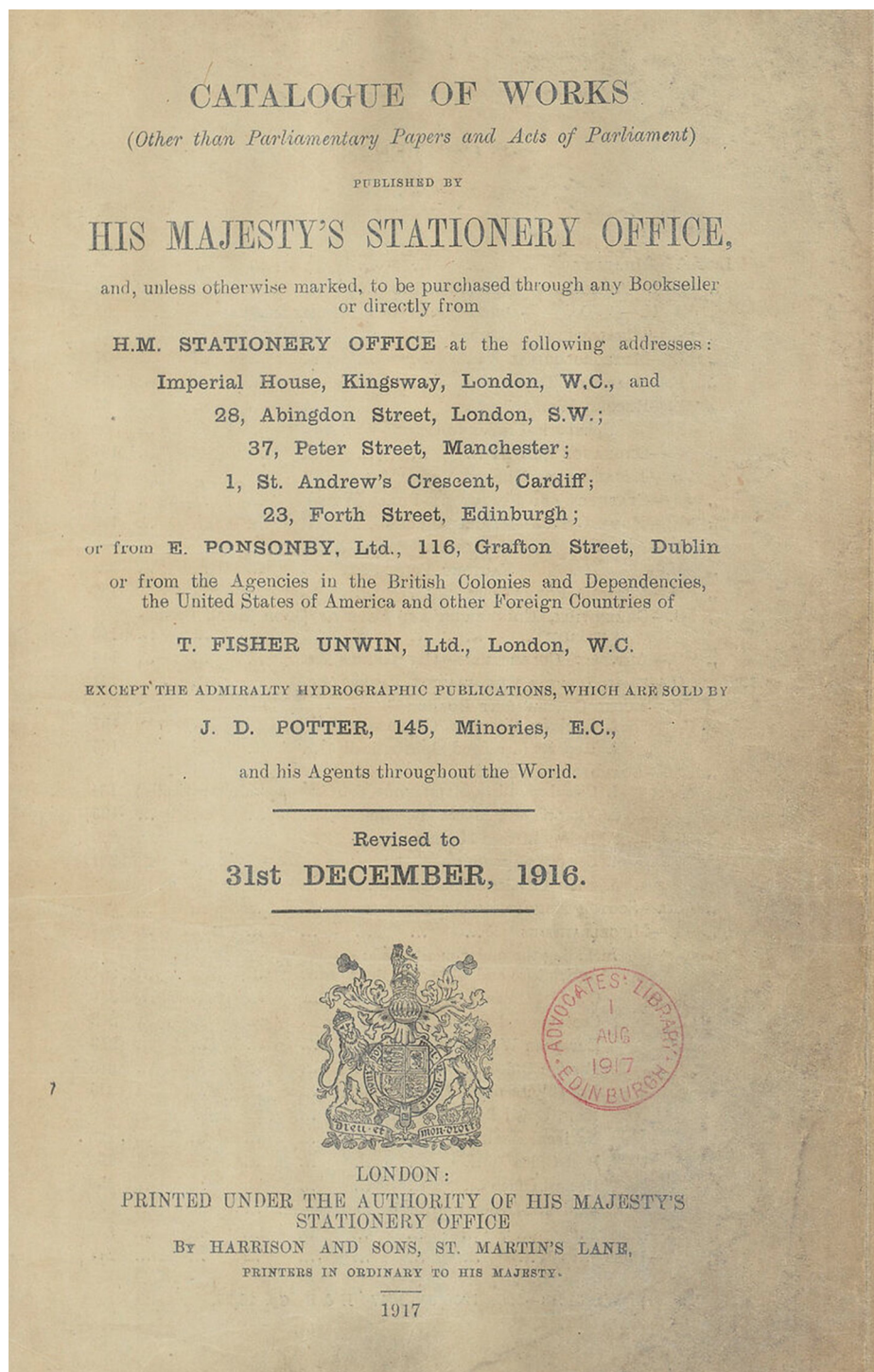


Figure 1: Title page from HMSO catalogue of 1917 (HMSO ©1917)

HMSO Catalogues

Prior to the pandemic, the National Library of Scotland had digitised their thirty-nine volumes of Her/His Majesty's Stationery Office (HMSO) catalogues which give bibliographical informational about HMSO publications from 1916 to 1980.

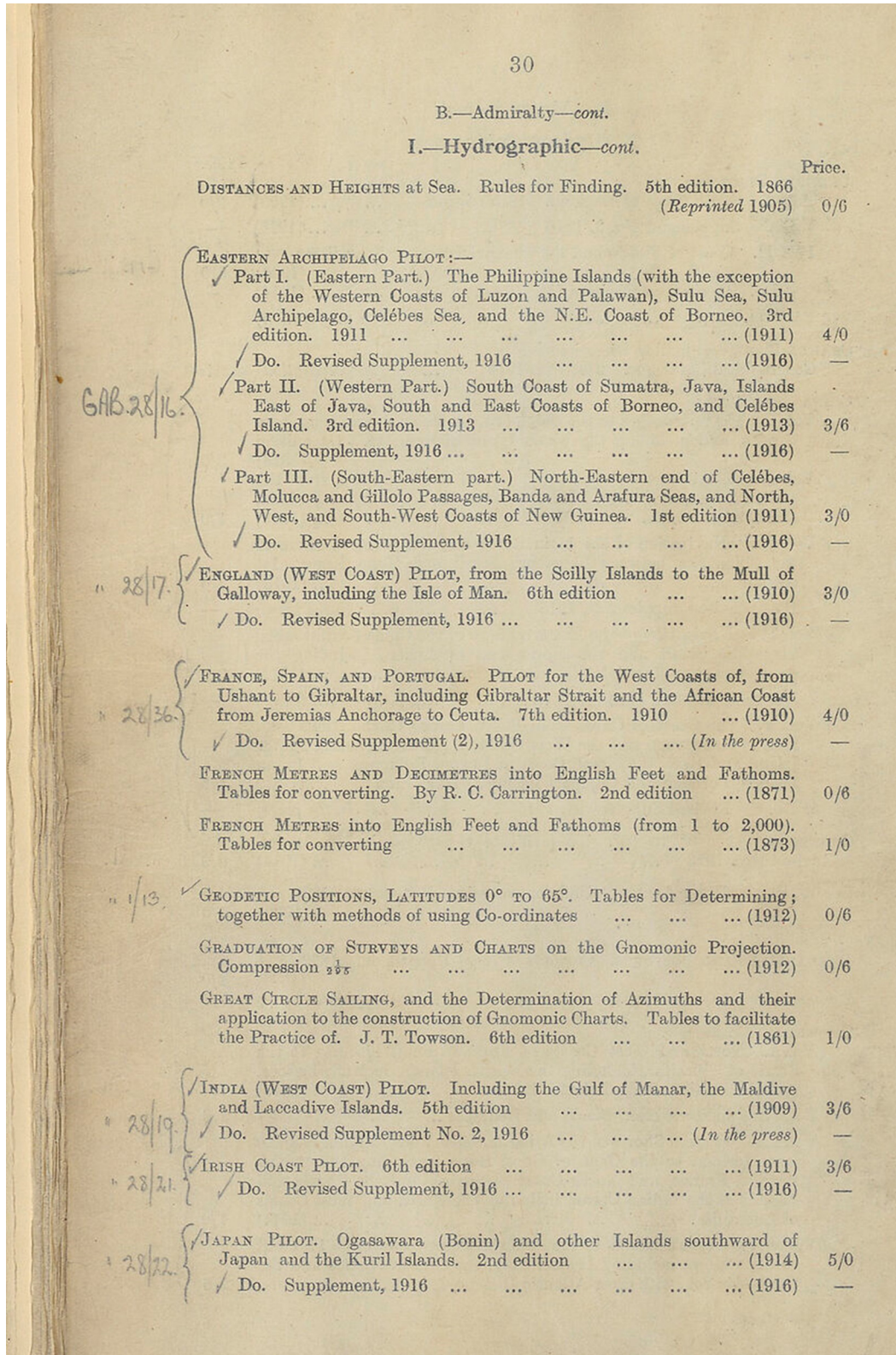


Figure 2: sample page from HMSO catalogue, with shelfmarks pencilled in (HMSO ©1917)

Historically the items listed were not individually catalogued – a shelf mark would be handwritten in each catalogue against each item held by the National Library of Scotland (NLS). These catalogues were only available in the Reading Rooms to be consulted, and items consequently requested.

While having digitised images of the HMSO catalogue available could be useful to readers, making individual records for each publication in our own catalogue would make these items much more accessible. This would allow readers to request specific items before being onsite, which became more crucial with limited Reading Room access due to Covid restrictions.

From each listing in the HMSO catalogue, there was enough information to create an adequate catalogue record without the item in hand. There was a publisher (HMSO), place of publication (London), the department that had issued the publication (from the section title of the catalogue) as well as the title and date of publication. A unique extract code could also be added to all these records so they could be further enhanced in the future if required.

Use of spreadsheets

The next challenge was overcoming the issue of staff having a variety of electronic devices without access to our usual Library Management system, Alma. The solution was for everyone to use spreadsheets instead – Alma allows the import of specially formatted spreadsheets¹ to create live catalogue records with associated holdings and items.

Two types of spreadsheet template were created – one for simple metadata creation that any member of staff could use and one for staff with prior cataloguing experience who felt comfortable adding additional metadata included in the HMSO entry.

	A	D	E	F	G	H	I	J	K	N	P	U
1	999\$a	24500\$a	24500\$b	24504\$a	24504\$b	24502\$a	24502\$b	24503\$a	24503\$b	264 1\$c	7301 \$b	NOTES
2	SHELFMARK	TITLE:	SUBTITLE:	TITLE STARTING THE:	SUBTITLE:	TITLE STARTING A:	SUBTITLE:	TITLE STARTING AN:	SUBTITLE:	YEAR:	DEPARTMENT:	*Author, series, note to add to record
3		TITLE:		TITLE STARTING THE:		TITLE STARTING A:		TITLE STARTING AN:		YEAR:	DEPARTMENT:	*Author, series, note to add to record
4												
5												
6												
7												

Figure 3: Simple spreadsheet template, with example data in rows 2 and 3

Each item is described in an individual row with columns filled in as required. As the spreadsheet is designed to map into Alma, the column headings translate to specific MARC fields with indicators and subfields:

- 999\$a – becomes the shelfmark in Alma
- 24500\$a / 24500\$b – Title and subtitle with no initial article
- 24504\$a / 24504\$b – Title and subtitle with initial article ‘the’

¹ [https://knowledge.exlibrisgroup.com/Alma/Product_Documentation/010Alma_Online_Help_\(English\)/040Resource_Management/060Record_Import/075Importing_Records_with_CSV_or_Excel_Files](https://knowledge.exlibrisgroup.com/Alma/Product_Documentation/010Alma_Online_Help_(English)/040Resource_Management/060Record_Import/075Importing_Records_with_CSV_or_Excel_Files)

- 24502\$a / 24502\$b – Title and subtitle with initial article ‘a’
- 24503\$a / 24503\$b – Title and subtitle with initial article ‘an’
- 264 1\$c – Year of publication
- 7101 \$b – Government Department

However, how would someone with no cataloguing background know that the column 24500\$a would mean this is the place to enter the title? Moreover, the data had to be strictly entered to create the correct fields with specific punctuation to import correctly into Alma.

We therefore needed to make the instructions as clear as possible, with the recognition that some staff were completely new to the jargon of cataloguing. The spreadsheet was colour coded and matched to guidance which explained each column with example data. Any data which was consistent for each item (such as the place of publication - London) was added in bulk to the spreadsheet and then hidden. This minimised manual data entry and allowed staff to concentrate solely on adding unique data.

The trickiest section of the spreadsheet was inputting the title – staff needed to think about if there was a title and subtitle or just a title, and whether this title started with an initial article or not. These elements can change how the title displays to our end users – and with no subjects or keywords, the accuracy of the titles was vital for the discoverability of the records!

Consequently, the spreadsheet had different ‘sets’ of 245 columns. Red for no initial articles, blue for a title starting with the, and so on. It was vital that only one ‘kind’ of 245 was filled out for each row – Alma would not be happy with a record which tried to add multiple main titles! (Of course, we did learn later how to automatically do this in Alma, but that came many months into the project.)

To make this as straightforward as possible, online training sessions were held with all staff on the project as well as a Teams chat to share tips and ask questions. People could also add notes to each item they added to the spreadsheet if they required a second opinion. Very quickly, the more experienced cataloguers were able to advise others and double-checked spreadsheets before import.

Moreover, as confidence and knowledge grew, more staff switched to using the advanced Excel template where they could use additional columns to add extra data supplied from the HMSO catalogue. This included edition statements and series information.

W	X	Y	Z
500\$a	250\$a	4900 \$a	4900 \$v
Amended 1921 version. Illustrated.			
Includes errata slip.	Revised edition.	Special report ; Special report ;	no. 1 4
		Technical paper ;	5
		Safety in mines research board ;	Paper no. 2

Figure 4: Columns in advanced spreadsheet

	A	D	E	F	G	H	I	L	N
1	999\$a	24500\$a	24500\$b	24504\$a	24504\$b	24503\$a	24503\$b	264 15c	7101 \$b
2	GWB.3/25	Combatant commissions as Second-Lieute	non-commissioned officers of the Regular Army.					1922.	War Office,
3	GWB.3/25	Commissions in H.M. Regular Army :	allowances, choice of regiment, outfit, &c., and directions as to regulations which should be consulted.					1922.	War Office,
4	GWB.3/24	Indian Empire :	for the use of soldiers proceeding to India.					1922.	War Office,
5	GSA-VII.6	Report for the years 1920, 1921.						[1922]	Fuel Research Board,
6	GSA-VII.1	Pulverised coal systems in America.						[1922]	Fuel Research Board,
7	GSA-VII.1	Tests on ranges and cooking appliances.						[1922]	Fuel Research Board,
8	GSA-VII.8					gravity of gases in small quantities.		[1922]	Fuel Research Board,
9	GMC.8	Research Committee on explosives for use in fiery and dusty mines, and the methods of testing them.						[1923]	Mines Department,
10	GMC.8			The application of stone dust in coal mines.				[1923]	Mines Department,

Figure 5: Completed spreadsheet ready for import

Importing spreadsheets to Alma

The next stage was to create a specific import profile² for the project in Alma. This would bulk upload the Excel spreadsheets and create MARC records with dedicated holdings and items. To further enhance this data, a set of normalisation rules³ was created to run alongside the ingest. This added in some specific RDA elements (such as 336-338 fields), added the project extract code HMSO to each record and converted fields so they were MARC compliant.

² [https://knowledge.exlibrisgroup.com/Alma/Product_Documentation/010Alma_Online_Help_\(English\)/040Resource_Management/060Record_Import/020Managing_Import_Profiles](https://knowledge.exlibrisgroup.com/Alma/Product_Documentation/010Alma_Online_Help_(English)/040Resource_Management/060Record_Import/020Managing_Import_Profiles)

³ [https://knowledge.exlibrisgroup.com/Alma/Product_Documentation/010Alma_Online_Help_\(English\)/Metadata_Management/016Working_with_Rules/020Working_with_Normalization_Rules](https://knowledge.exlibrisgroup.com/Alma/Product_Documentation/010Alma_Online_Help_(English)/Metadata_Management/016Working_with_Rules/020Working_with_Normalization_Rules)

```

HMSO rules
Rule Normalization Drool
rule "HMSO rules"
when
TRUE
then
removeField "999"
addField "040.{-,-}.a.StEdNL"
addSubField "040.b.eng"
addSubField "040.e.rda"
addSubField "040.c.StEdNL"
addField "300.{-,-}.a.1 volume" if (not exists "300.a")
addField "336.{-,-}.a.text"
addSubField "336.2.rdacontent"
addField "337.{-,-}.a.unmediated"
addSubField "337.2.rdamedia"
addField "338.{-,-}.a.volume"
addSubField "338.2.rdacarrier"
changeField "910" to "710"
changeField "911" to "710"
changeField "912" to "710"
changeField "701" to "700"
addField "919.{-,-}.a.HMSO"
end
    
```

Figure 6: Initial normalisation Rules to enhance spreadsheet data, see end of article for the current version

One quirk with Excel import to Alma is that each column number needed to be unique – if more than one column had the same configuration, these would become merged in the Alma record. A common practice was to have more than one corporate body, thus more than one 710 field. We therefore used random field numbers in Excel and converted these when the record was created.


After some trial and error, we eventually ended up with a smooth routine and created standard records which allowed these items to be individually retrieved.

	N	O	P	Q	R	S
	7101 \$b	7101 \$e	9101 \$a	9101 \$b	9111 \$a	9111 \$b
	Standing Advisory Committee on Trunk Road Assessment.	issuing body.	Great Britain.	Her Majesty's Stationery Office.	Great Britain.	Department of Transport.

Figure 7: Spreadsheet with multiple corporate bodies using different field numbers

LDR	00526nam a22001697i 4500
001	99116105443604341
005	20201207092527.0
008	201207s1921 enk f000 0 eng d
040	__ a StEdNL b eng e rda c StEdNL
245	00 a Handbook of bayonet training for His Majesty's Fleet.
264	_1 a London : b His Majesty's Stationery Office, c [1921?]
300	__ a 1 volume
336	__ a text 2 rdacontent
337	__ a unmediated 2 rdamedia
338	__ a volume 2 rdacarrier
710	1_ a Great Britain. b Admiralty, e issuing body.
919	__ a HMSO

Figure 8: Ingested record in Alma



Book

[Handbook of bayonet training for His Majesty's Fleet.](#)

Great Britain. Admiralty, issuing body.
London ; His Majesty's Stationery Office; 1921?

[Request >](#)


Top


Request


Details


Links


Request


Export RIS


Export to Excel


Permalink


Email


Citation



Print

Figure 9: Record in Library Search for readers to request

Lessons Learned

It was a steep learning curve for all staff, maybe even more so for staff who were used to cataloguing directly into Alma rather than being one step removed! Everyone had different comfort levels with Excel spreadsheets. A common irritant was the size of the spreadsheets as this work required a lot of columns and was easier for some people to navigate than others depending on the screen size of their device.

Another frustration was not being able to access the item at the shelf to double check any confusion over the description in the HMSO catalogue. Luckily, we were able to flag such items so they could later be checked when back onsite.

One of the most complicated issues to overcome was the import routine to Alma. Using the Alma Sandbox, spreadsheets were initially loaded, and many were coming back with partial or full errors. We uncovered the main reasons for failure were having blank columns in the middle of the spreadsheet and rows where multiple 245 titles had accidentally been added. This eventually became a core task in the double checking of spreadsheets (as well as for typos and punctuation) before being sent for ingest.

Successes

The project quickly grew to include staff from across the Library who worked over many months to make over 30,000 items available. It is a project that quite possibly would never have been feasible without the sudden need for everyone having to work from home.

Not only has it widened the cataloguing experience of those who previously worked on mainly physical items, it gave staff across the Library some practical metadata experience. It helped to demystify how items end up on the catalogue and the standards that are used.

Ongoing Excel projects

The most significant success has been continuing to use Excel for other projects, even when back onsite and using Alma. These spreadsheets are great for projects where there are many items with a similar layout, so the data entry is routine (such as with reports from public bodies). This consequently frees up the time of trained cataloguers to focus on more complicated cataloguing tasks which don't necessarily lend themselves well for spreadsheets.

As time has progressed, the normalisation rules employed on these projects have also become more sophisticated (see below). We are now able to detect if there are articles in the title field (therefore only one 245 column is needed in Excel) as well as add the correct punctuation on import. This means that the spreadsheet users can spend more time entering the data without these concerns, making the workflow more efficient.

Most importantly, having all cataloguing staff comfortable with using Excel for metadata work means we have an immediate backup plan for any network issues planned or unplanned. We can now carry on cataloguing, whatever happens!

Improved normalisation rule(s)

```
rule "HMSO rules"
when
TRUE
then
removeField "999"
addField "040.{-,-}.a.StEdNL"
addSubField "040.b.eng"
addSubField "040.e.rda"
addSubField "040.c.StEdNL"
changeSecondIndicator "245" to "4" if (exists "245.{*,*}.a.The *")
changeSecondIndicator "245" to "3" if (exists "245.{*,*}.a.An *")
changeSecondIndicator "245" to "2" if (exists "245.{*,*}.a.A *")
addField "300.{-,-}.a.1 volume" if (not exists "300.a")
addField "336.{-,-}.a.text"
addSubField "336.2.rdacontent"
addField "337.{-,-}.a.unmediated"
addSubField "337.2.rdamedia"
addField "338.{-,-}.a.volume"
addSubField "338.2.rdacarrier"
changeField "910" to "710"
changeField "911" to "710"
changeField "912" to "710"
changeField "701" to "700"
addField "919.{-,-}.a.HMSO"
end
```

```
rule "Add punctuation"
when
TRUE
then
suffix "100.a" with "." if (not exists "100.a.*\\\\"")
suffix "700.a" with "." if (not exists "700.a.*\\\\"")
suffix "710.a" with "." if (not exists "710.a.*\\\\"")
suffix "500.a" with "." if (not exists "500.a.*\\\\"")
suffix "264.a" with " :" if (not exists "264.a.*:")
suffix "264.b" with "," if (not exists "264.b.*,")
suffix "264.c" with "." if (not exists "264.c.*\\\\"")
end
```

```
rule "Fix punctuation"
when
  (exists "245.a.*" ) and (exists "245.b.*")
then
  suffix "245.a" with " :" if (not exists "245.a.* :")
  suffix "245.b" with "." if (not exists "245.b.*\\\\".)
end
```

```
rule "Fix punctuation 2"
when
  (exists "245.a.*") and (not exists "245.b.*")
then
  suffix "245.a" with "." if (not exists "245.a.*\\\\".)
end
```

```
rule "Change 245 first indicator"
when
  (exists "100.a.*")
then
  changeFirstIndicator "245" to "1"
end
```

The challenges of data ingest, transformation and aggregation at the National Bibliographic Knowledgebase

Jennie-Claire Crate  0000-0002-6594-8017

Product Manager - Library Hub, Jisc

Received: 08 September 2025 | Published: 22 September 2025

ABSTRACT

Large-scale library data aggregations provide additional discovery opportunities for users and benefit collection management, metadata and interlending work in libraries. However, they present significant delivery challenges around the matching and deduplication of records and require continuous improvement and maintenance to keep pace with changes in the way the sector creates and shares its metadata.

This article describes how contributors to Jisc's National Bibliographic Knowledgebase (NBK) supply their data and how the Jisc Library Hub team matches and deduplicates the data for use via Library Hub Discover, Library Hub Cataloguing, and Library Hub Compare services. It will also explore some of the challenges faced with data transfer, metadata formats, non-standard metadata, and future developments.

This article is based on a paper given at CILIP's Rare Books & Special Collections Group conference in September 2025.

KEYWORDS NBK; metadata transformation; metadata deduplication

CONTACT Jennie-Claire Crate  jennie-claire.crate@jisc.ac.uk  Jisc

Background

Jisc's Library Hub services were launched in the summer of 2019, replacing the Copac and Copac Collections Management tools created by MIMAS and EDINA's SUNCAT serials catalogue. Underpinning the three services (Library Hub Discover¹, Library Hub Cataloguing², and Library Hub Compare³) is the National Bibliographic Knowledgebase (known as the NBK). The NBK brings together catalogue data from 205 contributors including national libraries, legal deposit libraries, academic, specialist, and museum libraries. Ingesting, storing and transforming the data supplied by our contributors is a challenging process that requires continual monitoring and adaptation in order to keep up with demand on the service.

¹ <https://discover.libraryhub.jisc.ac.uk/>

² <https://cataloguing.libraryhub.jisc.ac.uk/>

³ <https://compare.libraryhub.jisc.ac.uk/>

Data ingest

One of the chief strengths of the NBK is the breadth of its coverage, and making sure that smaller, specialist libraries are able to contribute their data is crucial to achieving this. Smaller libraries generally have fewer staff members and therefore are sometimes restricted in how they can create and contribute their data in terms of both staff time and technical expertise. This means that the Library Hub team needs to be able to accept data via a variety of methods and support contributors in using them.

The majority of contributor data is sent to the NBK directly via SFTP, but increasingly this method presents problems for libraries. Local IT security measures can prevent implementation and mean that we need to seek alternative methods for data transfer. Additionally, a lack of technical expertise or confidence at the contributing institution can sometimes hamper the setup of SFTP, with the Library Hub team only being able to offer help remotely. In terms of managing the flow of data into the NBK, being able to harvest data from local systems using OAI-PMH would be ideal, but again local cyber security measures now mean that this is increasingly not supported in our contributor's institutions. Cyber security concerns also mean that WebDAV is no longer considered to be an optimal way to transfer data onto our servers, so what other options are there?

It is possible to transfer files by less automated means, such as via online file sharing service like Dropbox or as an email attachment. Whilst seeming to be a solution, these file transfer methods present the Library Hub team with workflow difficulties as we need to manually access, download and then reupload the data to the correct area in our file transfer server. These time-consuming options introduce the possibility of user error on both sides of the transfer, and as a result we cannot support these methods.

To offer a new and more user-friendly option to NBK contributors, the use of Amazon Simple Storage (abbreviated to S3) buckets is currently being trialled with a group of contributing libraries that have experienced difficulties with alternative file transfer methods. If successful, this could work alongside SFTP as one of our main data submission options as it allows the automated collection and processing of data files and therefore fits in with our existing workflows.

Matching and deduplication for Library Hub Discover

Once the data has been received, it is processed and deposited across multiple databases and data stores depending on its intended use. The NBK is just one part of a complex data flow that has to take into account the needs of both users of the three services and those of the Library Hub team and Jisc for monitoring and reporting.

Library Hub Discover, whilst being the most heavily used of the services (with over 10 million searches carried out last year) is perhaps the most straightforward use of the data. As with standard library discovery layers, Library Hub Discover surfaces the metadata held in the underlying database and allows search parameters to be set by

the user, with pre-filtering by data facets such as author, title, format and publication date. However, the scale of Library Hub Discover means that the data needs to be deduplicated to create aggregated records where multiple contributing libraries hold copies of the same item.

Discover's deduplication is carried out using scripts which pass the data through a series of challenges to match and merge identical items. New entries into the NBK are filed alphabetically by title and are then tested against multiple items on either side of their place in the alphabetical index to see whether there are matches in title and multiple standard number fields (ISBN, ISSN, ISMN, ESTC and others). If this doesn't result in a match, metadata relating to edition, creator, date of publication and eventually pagination, score type and map scale are compared. If no match is made, the new entry into the database is treated as a unique item and will be displayed in Library Hub Discover as being held in a single NBK contributing library.

Unmatched records are not usually caused by any kind of issue with the match scripts, but more usually by the variation that is found in the metadata submitted by contributing libraries. In order to enable all sizes and types of library to contribute to the NBK, Library Hub accepts data in almost any digital format from MARC21 to spreadsheets and Word tables. The minimum level of description set for a record is also very low to enable maximum inclusion, and only a record ID number and item title are mandated. What this means in practice is that these less-full, lower quality descriptions have far fewer data points against which to match, making it much more likely that they will be assessed to be "unique" even where they are not. In these cases, users of Library Hub Discover will find multiple entries for what appear to be identical publications which can cause frustration when searching for the nearest copy of a book for researchers and uncertainty for interlibrary loan teams looking for an item for their patrons.

Matching and deduplication for Library Hub Compare

Deduplication issues also have an impact on users of Library Hub Compare, Jisc's collection management tool. Library Hub Compare uses the same deduplicated index as Library Hub Discover, allowing library teams to compare their local holdings with other NBK contributor institutions. This collaborative collections management approach is about to come to the fore for libraries in the UK with the launch of the UK Print Book Collection (UK PBC), an initiative created to ensure that a minimum of seven print copies of books published in or before 2010 is retained in the UK whilst enabling libraries to make evidence-based decisions when managing their physical collections. UK PBC will launch in October 2025, and is using Library Hub Compare to show libraries which items in their print collection might need to be retained to avoid losing one of the final seven remaining copies. The benefit of this work to libraries is that they can safely manage down their print holdings and repurpose physical space in their buildings to meet the needs of their users more effectively, and for library users it means that access to the print items they need will be preserved. Just as with Library

Hub Discover though, poor or sparse metadata means fewer matches being made and therefore more library holdings being reported as 'unique'. This results in lower confidence in the Compare reports for UK PBC and can cause confusion for collections management teams. All collection analytics tools tend to over-report item rarity due to deliberately conservative matching algorithms, which prioritise avoiding false positives over identifying duplicates. Making incorrect matches is even more undesirable than incorrectly labelling an item as rare or unique, but clearly neither is giving a completely accurate picture.

An additional challenge for Library Hub Compare is the currency of the dataset. Contributors choose how often to update their holdings in the NBK. Most opt to do this weekly or monthly but longer intervals can be caused by lack of staff capacity or changes of key team members in contributing libraries, or by the disruption caused by a change of library management system. Data currency is important for collections management analytics as it ensures decisions are being made based on the most recent available data, bringing confidence to stock management and editing activity. This is another strong argument for preferring automated data transfer methods, as once they have been set up and scheduled they can be carried out frequently with minimum staff intervention, ensuring data currency and workflow efficiency.

Ways forward

Libraries have historically tailored metadata practices to meet local needs, system requirements, and workflows. In addition to this, the hybrid environment created by changes in standards that are not universally adopted means that it can be difficult to define what "good" metadata looks like in the NBK and adds another layer of complexity to deduplication in the database. Revisiting and reassessing local metadata practices that vary from the standard and improving minimal legacy records could go a long way to addressing issues with data matching in large aggregations but of course comes with a cost to libraries. The benefits of this work, and the retrospective application of the new standards, can be difficult to articulate in a business case but could bring financial benefits in the long term when it becomes possible to download and ingest data from the NBK without having to edit it to match local practice. Moving toward more standards-based approaches can facilitate collaborative description and collections management, simplify system migrations, and ease the adoption of new tools and evolving standards. While innovations like linked data can be slow to take root, the long-term benefits in enhancing discoverability via the semantic web and for accessibility are substantial for both collections teams and users.

Non-MARC cataloguing

What you (and your organisation) need to know

Anne Welsh  0000-0002-5621-7490

Beginning Cataloguing

Received: 17 September 2025 | Published: 22 September 2025

ABSTRACT

A checklist to use as a starting point if and when you begin cataloguing in a system that does not use MARC.

KEYWORDS non-MARC systems

CONTACT Anne Welsh  anne@beginningcataloguing.com  Beginning Cataloguing

MARC, The Whole MARC, And Nothing But The MARC?

If your background is entirely in national and academic libraries, you might be forgiven for thinking that all library cataloguing is created in MARC 21 ([Library of Congress Network Development and MARC Standards Office, 1999-2025](#)), the latest English language iteration of the Machine Readable Cataloguing format, which was introduced in 1999 ([Library of Congress Network Development and MARC Standards Office, 1998](#)) and superseded UKMARC (1969-1999), USMARC (1965-1999) and CANMARC (1974-1999), the Australian library community having already moved from AUSMARC (which began in 1973) to USMARC in 1996 ([Chapman, 2005](#)). Indeed, the National Acquisitions Group's report *Quality of Shelf-ready Metadata* ([Booth, 2020](#)) and its following *NAG Servicing Guidelines* ([National Acquisitions Group, 2020](#)) and *Metadata Profiles* ([National Acquisitions Group and Southern Universities Purchasing Consortium, 2021](#)) focus on MARC.

As Emma Booth described in her Executive Summary, "The National Acquisitions Group Quality of Shelf-Ready Metadata Survey collected data from 50 Higher Education libraries in the UK and Ireland" with the majority reporting "that they receive shelf-ready MARC records for print and / or e-books from multiple Framework suppliers, rather than using one supplier. 70% receive records for print books and 90% for e-books" with lower percentages for other formats ([Booth, 2020](#), p. 5). Eric Jackson explained, "As pressures on academic libraries increase, both in terms of staffing and budget, they have become ever more reliant upon the acquisition of 'shelf-ready' materials" ([National Acquisitions Group, 2020](#), Introduction). Explaining the rationale behind the *NAG Servicing Guidelines*, he highlighted the cost reductions that follow

from libraries adopting the same processing requirements, asserting that “Complete or partial adoption of the standards can also feed into more streamlined internal workflows, freeing up library staff time for other duties.” With regard specifically to metadata, Booth described “a need for clearer standards regarding shelf-ready MARC records to be established, so that suppliers can work with libraries to ensure that the metadata in the supply chain is functional for a variety of discovery purposes, and *does not require each library to perform manual checking, correction or enrichment tasks*” (Booth, 2020, p. 5, my italics).

Share and Share Alike?

Indeed, this cuts right to the essence of why so many libraries have adopted MARC. If you share data, it is currently still the main show in town, though great strides are being made in the development of its successor BIBFRAME, both by the Library of Congress (Library of Congress, 2025) and others, including the Share Family¹, of which the British Library is a member (British Library, 2023). The National Library of Sweden is the first to transition entirely to BIBFRAME (Breeding, 2024), having begun the move in 2018, using open access software VuFind and FOLIO.

As described in their introduction, “The MARC formats are standards for the representation and communication of bibliographic and related information in machine-readable form” (American Library Association ALCTS/LITA/RUSA Machine-Readable Bibliographic Information Committee and Library of Congress Network Development and MARC Standards Office, 1996, 1.1). Or in other words, “The MARC 21 formats are communication formats, *primarily designed to provide specifications for the exchange of bibliographic and related information between systems*” (American Library Association ALCTS/LITA/RUSA Machine-Readable Bibliographic Information Committee and Library of Congress Network Development and MARC Standards Office, 1996, 2.1, my italics). They have become seemingly ubiquitous because institutions who wish to ingest, output and share their metadata can do so easily. Library suppliers provide metadata in MARC, so those who wish to purchase new materials shelf-ready can receive metadata at the same time. And despite the growth of ONIX as the standard within the publishing industry, we see more and more publishers providing MARC alongside their new publications.

Why Not MARC?

However, not all libraries have a clear and present business need to share their metadata. Sometimes cataloguers can be surprised that their organisation does not recognise a business case for MARC cataloguing, but managers may be correct in assessing their needs in this way:

1. MARC cataloguing presupposes a level of knowledge that staff already have and / or need training to acquire / keep up-to-date, so there is an inherent cost.

¹ <https://www.share-family.org/>

2. MARC is an exchange format. If your organization does not share metadata (neither importing nor exporting it), it is not making use of the main advantage of MARC.
3. MARC (and subsequent exchange formats, like BIBFRAME) is designed to allow for sophisticated nuances to differentiate between different editions, different translations, and different issues of serials. Smaller collections simply may not have the variations in stock that require such nuance.
4. Some libraries' holdings are mainly discovered via the library catalogue / discovery layer, but others are mainly discovered by browsing. Most school libraries, for example, are organised to be used mainly by readers who browse the shelves – they need a catalogue for inventory purposes and borrowing, but their users don't tend to Search, Find and Retrieve – they browse as their main form of discovery. Similarly, some law firms have embedded collections within different legal teams whose members get to know so intimately that cataloguing is really about inventory and stock control rather than discovery. There are other situations in which Search and Retrieval is not the primary task for users. (Remember, IFLA's "user tasks" are "generic" and but not *all* user tasks are).
5. Sometimes the library is not the biggest curatorial activity. An archive with a small book collection is more likely to acquire software designed primarily for archival description. A museum with a small book collection is likely to acquire software designed for museum description. Even where these have a "library module" or "book module" they may not use MARC. They may have been designed for use not by librarians but by archivists or museum curators who have to deal with books, and who may, therefore, be using terminology and even a mindset that is a little different from a library cataloguer.
6. Some organisations may believe that library management systems with MARC cataloguing are more expensive. It is true that the purchase price may be lower, but there may be hidden costs if you are locked into a proprietary cataloguing system that doesn't use a well-known exchange format (like MARC or one of the archival or museum exchange formats).

What does Your Organisation Need?

If you work in an organisation that neither imports nor exports metadata into a shared system (or from metadata vendors), there are four things it will benefit from you doing:

1. Being consistent in how you use the fields.
2. Checking how things display on the screen your users see, and in any reports they run.

3. Being aware of the potential for importing metadata should a need be seen to do so in future.
4. Being aware of how you will export your metadata if and when you need to move to another system.

This checklist, which I originally wrote as a free download to celebrate my fifth anniversary trading as Beginning Cataloguing, is designed to be helpful across these four needs.

What Do You Need?

In arriving in a situation in which you are not cataloguing in MARC, you will be one of four types of person:

1. Someone with a background in cataloguing, including MARC cataloguing
2. Someone with a library background, but no practical cataloguing experience
3. Someone with a background in a related profession (Archives / Museums / Galleries) with documentation experience (for manuscripts, digital archives, objects, digital objects) but no book cataloguing experience
4. Someone with no background in documentation or cataloguing

You are also likely to fall into one of three categories of ambition for your next job:

1. You want to move into or continue in metadata, including library metadata, and therefore including MARC cataloguing
2. You want to move into a position in which you manage someone who does the library cataloguing
3. You don't want anything to do with library cataloguing ever again – you just want to ensure you are doing the best you can for your organisation

The type of person you are and your category of ambition will determine the level of awareness of MARC cataloguing you need to acquire. For example, if you are an experienced MARC cataloguer, you would be best advised to maintain your level of awareness of MARC (and BIBFRAME) for your future career, so at entry point, you might want to ensure your organisation will support your training to do so (even though they don't see a need for MARC themselves, they do have a duty of care to you). However, if you have no MARC experience and no desire to be a MARC cataloguer, you might want to ensure your organisation will support you by ensuring someone with MARC knowledge is involved if you ever need to export or import your data.

The Checklist covers both these ends of the spectrum and everything in between, so choose from the Checklist based on what you know and what you want to know.

Note: This checklist shares some of the things that I check when I am working with a new client or an existing client obtains different software. It cannot be comprehensive of every small thing I have learned over 30+ years. It simply provides a basis for you to get started. DO add your own items based on your own experience.

Non-MARC Cataloguing Checklist

Item	Answer	Don't Know
Why doesn't your organisation use MARC?		
Is your management team aware of how they will export metadata from the system if they need to? (Some managers will say that they will never leave the system they have, but any system can cease to exist, so you always want to be able to get your metadata out, even if only in an emergency).		
Will your organisation support you in maintaining (or even acquiring) MARC knowledge, even though they don't catalogue in MARC?		
Will your organisation support you in acquiring (or maintaining) archival / museum documentation skills (e.g. if you are cataloguing a small book collection in a bigger archive or museum)?		
What (else) will your organisation do to support you in being ready for your next career move?		

Item	Yes	No	Don't Know
Is it an XML system?			
If Yes, does it use MARC/XML?			
If No, does it use an XML standard from Archives / Museums?			
Do the field names in your input screen use library terminology?			
If No, do they use terminology from Archives / Museums or from another general standard like DublinCore?			
Can you add new fields to the input screen?			
If yes, do they show up on the display that catalogue users see?			
If yes, do they look as you expect?			
Is there a Name Authority File built into the system (i.e. a picklist of names of authors, etc.)?			
If yes, what rules does it follow? (Archival, Museums, Library)?			
If it's set up using Archival or Museum style headings, will your adding library style headings mess up what the archival / museum documentation module(s) do(es)?			
Can you add a new name heading?			
If yes, how does it display in the catalogue display your users see?			
What happens if you enter a title starting with the definite ("The") or indefinite ("A(n)") article?			
Does the title display in the correct place alphabetically in the catalogue your users see?			
Does the title display in the correct place alphabetically in reports you and your colleagues may run (e.g. reading lists)?			
What field do you use to record where the book sits on the shelf?			

Item	Yes	No	Don't Know
If you use a Classification Scheme (e.g. Dewey Decimal Classification, Library of Congress Classification) does it appear on the display your users see in the way you expect / would like?			
If you use a Classification Scheme, can you run a report in the order of the Classification Scheme?			
If yes, does it display in the order you expect / would like?			
If you use shelf marks, do they appear on the display your users see in the way you expect / would like?			
If you use shelf marks, can you run a report in shelf mark order?			
If yes, does it display in the way you expect / would like?			
Whether you are using a Classification Scheme or shelf marks, do your entries interfere in any way with those input in the archival / museum module? (i.e. will it be clear to catalogue users that this item is a book and shelved with the books and not in the archive / museum)?			
Are there any fields governed by pick lists (e.g. publisher, place names, series)?			
If yes, can you use the same ones the archive / museum documentation uses without any issues?			
Can you add to these pick lists?			
Can you add series information?			
If yes, do these appear in the display your users see in the way you expect / would like?			
Can you add edition information?			
If yes, does it appear in the display your users see in the way you expect / would like?			
If you keep old editions when the new one arrives, are you expected to create a new metadata set for each edition, or one metadata set that includes a list of all the older editions you hold in a holdings and / or notes field?			

Item	Yes	No	Don't Know
Can you add extent information and physical description in the way you expect / would like?			
	If yes, does it appear in the display your users see in the way you expect / would like?		
How are you expected to deal with serials? One metadata set with a start date (and a finish date if needed)?			
Is there a separate serials module or acquisitions module where you check in individual issues of serials?			
Is there a function to record when serials are expected and that flags up if one doesn't arrive?			
If you are cataloguing laws and treaties, are there special rules to follow?			
Can something have more than one title? (Some systems built for archives allow you to repeat the title field, rather than having a separate alternative title field as we do in libraries).			
Is there a subtitle field? (Some systems just have one field for all the title information)			
If there is more than one creator, do you repeat the field, or are there separate fields for co-creators (co-authors; added entries)?			
Are there separate notes fields for different types of note (e.g. binding, bound with, change of name of serial, etc.)?			
Are there any other fields specific to the book collection that you need to test? (ISBN? ISSN? Subject Headings? Abstract?)			
Can you export a set of metadata?			
	If yes, does it look as you expect?		
	If yes, can you put it into MarcEdit ² ?		
	If yes, does it look as you expect / would like?		
Can you export a range of metadata?			
	If yes, does it look as you expect?		
	If yes, can you put it into MarcEdit ² ?		
	If yes, does it look as you expect / would like?		

² <https://marcedit.reeset.net/downloads>

Item	Yes	No	Don't Know
Can you export all your metadata?			
If yes, does it look as you expect?			
If yes, can you put it into MarcEdit ² ?			
If yes, does it look as you expect / would like?			
Can you import a set of metadata? (You can use MarcEdit ² to manipulate data and then import it).			
If yes, does it look as you expect, and does the display your catalogue users see look as you expect?			
Does metadata created in MarcEdit ² and imported look better than metadata input directly into your system?			

References

American Library Association ALCTS/LITA/RUSA Machine-Readable Bibliographic Information Committee and Library of Congress Network Development and MARC Standards Office (1996) *The MARC 21 Formats: Background and Principles*. Revised November 1996. Available at: <https://www.loc.gov/marc/96principi.html> [Accessed: 17 September 2025]

Booth, Emma (2020) *Quality of Shelf-ready Metadata: Analysis of Survey Responses and Recommendations for Suppliers*. Available at: https://nag.org.uk/wp-content/uploads/2022/03/NAG-Quality-of-Shelf-Ready-Metadata-Survey-Analysis-and-Recommendations_2021Corrected.pdf [Accessed: 17 September 2025]

Breeding, Marshall (2024) 'The National Library of Sweden Makes Strategic Decision to Implement a New Library Service Platform', Library Technology Guides Current News Service and Archive, 15 February. Available at: <https://librarytechnology.org/pr/29785/the-national-library-of-sweden-makes-strategic-decision-to-implement-a-new-library-service-platform> [Accessed: 17 September 2025]

British Library (2023) 'Share Family: British National Bibliography (Beta) service is live', *British Library Digital Scholarship Blog*, 14 July Available at: <https://blogs.bl.uk/digital-scholarship/2023/07/share-family-british-national-bibliography.html> [Accessed: 17 September 2025]

Chapman, Ann (2005) 'MARC Cataloguing Formats Worldwide', in *Bibliographic Management Factfile*. UKOLN. Available at: <https://www.ukoln.ac.uk/bib-man/factfile/cataloguing-formats/othermarc/index.html> [Accessed: 17 September 2025]

Library of Congress (2025) *Bibliographic Framework Initiative*. Available at: <https://www.loc.gov/bibframe/> [Accessed: 17 September 2025]

Library of Congress Network Development and MARC Standards Office (1998) *MARC 21: Harmonized USMARC and CAN/MARC*. Washington, D.C.: Library of Congress. Available at: <https://www.loc.gov/marc/annmarc21.html> [Accessed: 17 September 2025]

- Library of Congress Network Development and MARC Standards Office (1999-2025) *MARC 21 Format for Bibliographic Data*. Washington, D.C.: Library of Congress. Available at: <https://www.loc.gov/marc/bibliographic> [Accessed: 17 September 2025]
- National Acquisitions Group (2020) *NAG Servicing Guidelines: Best Practice for Academic Libraries*. 2nd edition. Available at: <https://nag.org.uk/wp-content/uploads/2021/04/Academic-Servicing-Guidelines-V2-Nov-2020.pdf> [Accessed: 17 September 2025]
- National Acquisitions Group and Southern Universities Purchasing Consortium (2021) *Metadata Profiles: MARC21 Records for Print and Electronic Books*. Available at: <https://nag.org.uk/wp-content/uploads/2021/07/NAG-SUPC-Metadata-Profiles-MARC21-Records-for-Print-Electronic-Books-v2.pdf> [Accessed: 17 September 2025]

Book review: Ethics in Linked Data

Reviewed by: **Elizabeth Cooper**

University of Bristol

Received: 5 August 2025 | Published: 22 September 2025

Watson, B. M., Provo, Alexandra and Burlingame, Kathleen (eds) (2023) *Ethics in Linked Data*. Sacramento, CA: Library Juice Press. ISBN 978-1-63400-133-5

CONTACT Elizabeth Cooper  liz.cooper@bristol.ac.uk  University of Bristol

Ethics in Linked Data, edited by B.M. Watson, Alexandra Provo, and Kathleen Burlingame, and conceived by the LD4 Ethics in Linked Data Affinity Group, is an incredibly engaging, thought-provoking and ambitious work. Contributing authors present a wide range of ethical linked data project case studies from across libraries, museums and archives in North America.

There is a consensus among contributing authors that an ethical framework should be the foundation and guiding principle of any linked data project, from its technological underpinnings through to its conceptual models, vocabularies, workflows, and overall practice. This book positions linked data as a vehicle for bringing the work of marginalised communities and creators to a wider, worldwide audience, through increasing the visibility, discoverability, and connectivity of creators and works. Use cases to achieve this aim include converting datasets to linked open data; repurposing traditional metadata into linked data ontologies; creating linked open data from scratch to render a collection of works more visible; and enhancing existing authority files with linked data entities. There is an emphasis on collaborating with the communities represented, particularly with living creators: gaining consent to create or re-use legacy metadata; returning agency and sovereignty to communities through collaborative linked data workflows. Other common threads are working slowly and carefully to create accurate metadata that adheres to ethical policies and procedures; and working at local level with critical reflection and feedback mechanisms in place at the start. Overall, this is a book about creating a community of practice with ethical consideration at its forefront.

Ethics in Linked Data is divided into three main sections: the first applies ethical theories to linked data ontologies and schemas; the second challenges the legacy of hegemonic power structures inherent in the knowledge organisation systems that inform linked data frameworks and workflows; and the third examines the outdated naming and representation of identity in traditional authority files and controlled vocabularies that get perpetuated in linked data entities.

In the first section, notable chapters look at the development of linked data technologies over the decades by wealthy organisations in an inequitable way, emphasising that linked data technology must lower barriers to entry and place end-users at the centre of its aims if it's to have relevant application in cash-strapped GLAM institutions. James Kalwara and Erik Radio challenge the very structure of RDF triples, arguing that this syntactical structure is based on Western language syntax (subject/object/predicate), an expression of the dominant languages of the Global North. They argue that RDF syntax excludes participants from nations whose language and ways of knowing/organising data are radically different. The authors' Marxist critique of RDF triple stores made for interesting reading but did not really add much weight to their argument. Practical recommendations to alter the already-established RDF structure, other than examining linked data software with a critical eye, were not forthcoming.

The second section looks at historic dominant power structures and institutions, including the Library of Congress, and the continuing prevalence of the white, male, Christian, cisgender worldview in our knowledge organisation systems. Chapters document reparative work undertaken through linked open data projects to redress harm done, particularly through colonialism.

Many indigenous cultures were erased through the collection (and stealing) of cultural artefacts by institutions linked to the Library of Congress in the eighteenth and nineteenth centuries. Systematic oppression eradicated indigenous culture as it sought to replicate national heritage collections like the anthropological collections of museums in Europe. This practice imposed inaccurate or false descriptions and narratives on Native American culture, artefacts, and history. Authors posit that the harmful practice of collecting artefacts continues in large-scale metadata harvesting via APIs and completionist practices, perpetuating harm. Outdated descriptions and inappropriate naming also affect findability. The conflation of indigenous identity and nationhood with land in Wikidata means that tribes not associated with a specific geographical area, such as the Metis community, have remained invisible. Authors highlight that these injustices in linked data need to be addressed at a local level with community collaboration and careful metadata practices.

Another aspect of identity misrepresentation occurs through the problematic assignment of gender to person entities in Wikidata, particularly for people identifying as non-binary or trans. Daniele Metilli and Chiara Paolini have undertaken the first qualitative and quantitative study of gender modelling in Wikidata since its inception in 2012. Their findings highlight that the Wikidata knowledgebase is far from inclusive of non-binary identities. This issue is sometimes amplified in data augmentation by bots and by the 'anyone can edit anything at any time' (AAA) ethos of Wikidata (with its overreliance on assumed goodwill of Wikidata editors). There is still much work to be done to ensure identities are described accurately and fairly in open knowledgebases like Wikidata.

The third section expands on the second section focusing on case studies looking at creating equitable entries in authority files, linking traditional authority files to more flexible diachronous linked data entities, such as those in Wikidata. Traditional authority files centre on single author descriptors (such as dates), presented as facts, as a means of disambiguation. In contrast, Wikidata items and properties lend themselves to connecting a multitude of descriptors to a single data entry point to reflect multiple identities. In a chapter entitled “The Oklahoma Native Artists Project : Oral History to Linked Open Data”, Megan Machen et al detail a project to increase the discoverability of work of living Native American artists in Oklahoma through recording interviews and encouraging self-representation, with linked open data as the end product. The emphasis is placed on getting consent to showcase work and artists, working collaboratively with the artists to create accurate metadata. In addition, linked open data becomes a vehicle for moving away from the dominance of metadata describing printed works, to metadata for works created in different media.

Some of the early chapters examine linked data through theories such as pragmatic sociology, reflective/diachronous language theory and Marxism. Whilst these provide an interesting angle to ethical consideration, they seem based on the academic background of the writers and are not always useful paradigms. While this book highlights injustice and asks challenging questions about inclusivity and equity in linked data practices, authors are not always able to suggest clear solutions. It should be noted that this book is primarily based on linked data developments in North America - where the richer institutions have had funding for linked data projects and infrastructure - itself an inequality in the world of linked data (which authors acknowledge).

Overall, *Ethics in Linked Data* provides much food for thought. It should be considered key reading for any institution embarking on a small-scale linked data project, with stand-alone chapters and an appendix containing an extremely extensive ethics toolkit. Whilst this book assumes a basic understanding of linked data, its chapters would be accessible for the uninitiated. Key concepts and theories are clearly presented and provide critical reflection on what could have been approached differently. The focus is on open knowledgebases such as Wikidata and locally created ontological frameworks, with little mention of the library-focused BIBFRAME ontology that tends to be maintained by large library management system vendors. Case studies are detailed, with descriptions of decision-making and practical suggestions that could be adapted to any local linked data project. Wikidata lends itself to small projects to showcase local collections. Ethical considerations presented would also have application in other areas within GLAM institutions.

Ethics in Linked Data is a tool for learning more about ethics, for challenging biases and cultural assumptions, as well as extending knowledge about the history of oppressive colonial power structures and how those continue to have an impact on our knowledge organisation and vocabularies today. What becomes clear throughout this book is that ethical decision-making in linked data often needs to be made at local level

and with care; metadata quality needs to be maintained in a way that is sustainable; and a collaborative approach between linked data practitioners and creators of works implemented where possible. Above all linked data practice needs to be an iterative process as meaning and vocabularies are always historically situated and constantly evolve over time.

Book review: Records and information management

Reviewed by: **Sarah Henning**

Museum Archivist and Information Governance Manager, Imperial War Museum

Received: 26 August 2025 | Published: 22 September 2025

Franks, Patricia C. (2025) *Records and information management*. 3rd ed. Chicago: ALA Neil Schuman; London: Facet Publishing. ISBN 979-8-89255-588-3 (paperback, ALA); ISBN 978-1-78330-818-7 (paperback, Facet)

CONTACT Sarah Henning  SHenning@iwm.org.uk  Imperial War Museum

Records Management is one of those things that most people take for granted. There is a general assumption in organisations that it just happens, and that the digital world has made things easier. After all, why bother with all that organisation when you can keep everything and do a free text search?

This book explains why. It provides a clear picture of the endless amount of systems, requirements and developments that records managers have to be on top of and it leaves you wondering about whether anyone manages to do it all.

As the introduction says, it isn't intended to be a cover to cover read, unless you're a student, and is more of a reference work. Its aim is to 'provide stability in a world that gets overly excited about the next new thing'. And Chapter 1 gives you a potted history of when each 'next new thing' came along. It took 5,867 years to go from using tokens for accounting to the invention of the manual typewriter, but only 50 to go from the invention of the PC to the widespread use of AI. As a records manager I envy my Victorian counterparts for the simplicity of their jobs, but not enough to go back to poor sanitation, no NHS and having to stay at home and do the housework.

Starting with the basics, the layout is good with a clear chapter structure. Each chapter has an introduction, central text and a summary, following the tried and tested 'say what you are going to say, say it and then say what you have said' formula. Text is split up into clear, headed sections, and there are a lot of illustrations and diagrams – really welcome with so dense a subject. What I particularly liked was the 'Paradigms and Perspectives' section at the end of each chapter, where guest authors give their own take on the topic discussed in the chapter. These are often very practical and reassuring after all the theory.

The book is comprehensive, with chapters on all the usual subjects, including creation, retention, classification, emerging technologies, physical record centres and archives. It is particularly strong on definitions and descriptions of software, tools and processes, and doesn't shy away from stating the obvious, something often overlooked when it comes to technology. Records managers have to be able to ask the obvious questions about how technology works to make sure it does the right things by the information it processes. This book provides the background to give records managers the confidence to ask those questions – experience shows you can't assume anything with technology.

Patricia Franks is writing from an American perspective and for a largely American audience and the contrast with records management in the UK is striking. It is a long time since I trained to be a records manager, but the approach here doesn't seem to have changed much – records management is still very much second best to its more exciting archive sister. It is also not taken nearly as seriously by governments and institutions, unless something goes wrong. Compare the government legislation on public records in the UK, mainly covered by two Public Record Acts and a smattering of subsequent legislation such as the Freedom of Information Act, with the significantly larger amount of legislation, directives and standards produced in the US and cited by Franks in this book. There is also a refreshing assumption that resources, time and support will be freely available as everyone realises what you are doing is important. Where they are not, Chapter 14, focusing on leadership and management skills, might help.

The US focus means that if you want to know about key legislation in the UK that affects records management, such as Data Protection legislation or the Freedom of Information Act, these are just briefly referred to. But if Franks were to cover every piece of legislation worldwide, this would be a very long book indeed.

Sometimes, the book can feel like an exhausting list of things to do. In a world where technology advances so quickly, records and information managers find it hard to keep up. Having a grasp of the basics, and what you can and can't compromise on, is crucial for those struggling with resource and the relentless focus on the next shiny thing. The end of chapter essays help with this, and Chapter 13 brings everything neatly together with advice on how to develop a records management programme and information governance strategy.

If I'm being really picky, the section on archives isn't as strong as the other chapters but this doesn't claim to be an archive textbook. There is also more of a focus on the commercial value of information rather than its value as evidence in other ways. However, this may not be a bad thing as emphasising how information can make and save you money is usually the main driver for managing it properly.

It is sobering to realise that despite the greater focus on laws and standards that there seems to be in the US, records remain very vulnerable if the people in charge don't care. Records management is undervalued but crucial for good governance and we need better training and the guidance that books like this provide.



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.

Editors: Karen F. Pierce and Fran Frenzel

ISSN 2399-9667

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

MDG blog: <http://catandindexgroup.wordpress.com/>

