Metadata & Discovery Group

Metadata in motion

building datasets from library catalogues

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

Received: 26 August 2026 | Published: 16 September 2026

ABSTRACT

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

KEYWORDS collections as data; metadata datasets

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

Introduction

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

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

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

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

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

Can we do this?

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

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

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

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

What is "Collections as Data"?

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

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

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

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

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

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

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

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

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

Building the LSE print PhD theses dataset[4]

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

Initial decisions

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

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

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

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

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

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

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

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

Weeding the records

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

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

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

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

Which data to include?

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

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

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

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

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

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

Data tidy

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

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

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

Author reconciliation

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

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

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

[10] id.loc.gov

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

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

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

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

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

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

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

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

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

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

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

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

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

Subject reconciliation

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

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

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

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

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

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

Transformation into CSV and JSON

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

Documenting the dataset

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

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

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

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

AI, ethics and responsible stewardship

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

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

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

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

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

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

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

Outlook and conclusion

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

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

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

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

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

AI statement

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

References

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Tools, software and services

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

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

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

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

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

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

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

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