


# Identify, obtain, explore

using NLP to link article and journal records in the NHM library catalogue

**Benjamin Cornish**

Natural History Museum

**Ben Scott**  0000-0002-5590-7174

Natural History Museum

Received: 01 June 2025 | Published: 17 June 2025

## ABSTRACT

This paper addresses the critical need to link related records in library catalogues, particularly for aggregated volumes and serial articles in physical collections, to enhance user access and discoverability. The Natural History Museum (NHM) Library, facing a significant backlog of unlinked article-level records within physical journal holdings, developed a semi-automated solution. We outline a two-stage pipeline utilising natural language processing (NLP) and record linkage techniques to automate the matching of article (child) records to journal (parent) records. This approach, leveraging title similarity and metadata extraction, achieved 60% accuracy against a previously-linked dataset. We discuss the challenge article level records can present and how such computational methods can optimise metadata workflows, allowing human effort to be redirected to more complex tasks and improving access across vast collections with limited resources.

**KEYWORDS** Natural language processing; serials cataloguing; linked data

**CONTACT** Benjamin Cornish  benjamin.cornish1@nhm.ac.uk  Natural History Museum

Relating records in a library catalogue can be achieved in several ways. Traditionally this would be via the use of authority files, access points and controlled vocabularies. Such relationships in modern Library Management Systems (LMS) allow users to navigate between records to find material by the same author or on the same subject. The Library Reference Model (LRM) – the conceptual basis for Resource, Description and Access (RDA) – highlights 5 users tasks which “confirm its outward orientation to the end user’s need” (Riva, Le Boëuf, Žumer, 2024, p. 15). Related records are mostly concerned with meeting the *explore* user task defined as allowing users to “discover resources using the relationships between them and thus place the resources in a context” (Riva, Le Boëuf, Žumer, 2024, p. 15).

In some cases however, relating records in a catalogue is arguably more integral and is required for the user to *obtain* (“To access the content of the resource”) or even

*identify* (“to clearly understand the nature of the resources found and to distinguish between similar resources”) a resource ([Riva, Le Bœuf, Žumer, 2024](#), p. 15). This is the case for records which have a horizontal relationship, where bibliographic records have been created for each work in an aggregated volume, or a vertical relationship, such as where bibliographic records have been created for a serial and the associated articles published in the serial.

In this paper we will outline how we have utilised natural language processing (NLP) methods to automate the process of finding such relationships in our library catalogue and combined with data cleaning tools are in the process of embedding links in our records. We will begin by discussing the types of records we are linking, the challenges of finding such links and the benefits to users of providing them. We will then lay out the process used and results from the project so far. We will conclude by considering how this type of work could be extended and how we see this work as leveraging tools to allow for human time to be channelled more effectively giving us the opportunity to work across many records with limited resources.

### Article level search

The Natural History Museum (NHM) like most research libraries provides access to journal articles in two main ways – via physical holdings of the journals the library has purchased or via electronic holdings either available open access or via library subscriptions.

Until the introduction of so called “Web Scale Discovery Services” (WSDS) article-level searching for electronic article was not possible directly from the library catalogue ([Sonawane, 2017](#), p. 27). Before WSDS or other federated search facilities to access electronic journal articles, a user had to know the title of the journal which could then be searched for. A link in the journal record would take the user to the databases which held the journal, from which a further search would often be needed to find one’s desired article. As Breeding mentions, the problem with this is that “asking library users to use one interface to find books and another to perform article-level searching adds a level of difficulty” ([Breeding, 2007](#)). The introduction of WSDS has changed this and the discovery layer of many library catalogues now allow users to search for material across a wide variety of resource types. This means any given search of a catalogue would likely return a mix of electronically accessible articles and books alongside the physical holdings of the library which require in person access.

Searching for articles in journals the library only holds in hard copy was, and remains, a challenge from a library catalogue. To find physical articles one has always had to use physical indexes or cross reference from online sources. To address this limitation some libraries have maintained policies of cataloguing articles in journals in certain circumstances, for example the BFI has extensive article level cataloguing ([British Film Institute, no date](#)). At the NHM we maintained a policy of cataloguing articles in journals for articles written by NHM colleagues, are about NHM collections

or obituaries of noted natural historians. This was and is beneficial to users as it gave them more resources in one place to search from, preempting the obvious benefits of WSDS.

### Relating Serial and Article records

Since 2017 the NHM has been using Alma as its Library Management System and Primo VE as its Discovery layer. Alma introduces a three-tiered inventory management model for physical titles as follows ([Ex Libris, no date a](#)):

- Bibliographic records – Marc21 records describing the bibliographic content of the material sitting in a 1:many relationship to –
- Holding records – Marc21 records recording the location of the material and its call number sitting in a 1:many relationship to –
- Item records

For the purpose of this project we are interested in two sets of records which we can refer to as parent and child records. The parent records are the serial records. These are bibliographic records providing the metadata for the physical journal holdings, holding records describing where runs of journals are held and requestable item records for each volume or issue of the journal (see [Figure 1](#)). Related to these are child records. These are bibliographic records providing the metadata for articles in the physical journal collections. The issue comes from the fact that these records also have holding and items despite the fact that they are not themselves "held" anywhere as they are component parts of their respective parent records.

There are several issues with the situation as it stands. Firstly, as the records are currently unlinked they are only related via the pressmark as it stood in 2017 and the citations in the article bibliographic records indicating which journal title they belonged to. This has meant that if journal holdings are moved there is very little hope of maintaining access via an article level search. Secondly, our previous system lacked a concept of a holding. In the move to Alma it was decided that these article level records, monographic as they were (i.e. they are records for single articles) should sit at monographic holdings, separating them further from the serial records to which they belong. For example, an article level record for an article in the Journal of Zoology might be sitting on a different holding than the journal in which it resides. Thirdly, there is a real mixed quality of citations provided in the article records. These citations are often abbreviated, could use journal titles which are not the main entries in our catalogue and thus cannot be cross referenced easily, and may be in MARC21 fields not actually displayed in Primo VE. Finally, these records were not necessarily correctly

## Bibliographic

022	—  a 0952-8369
022	—  a 1469-7998 (Internet)
035	—  a (UK-LoNHM)11310-44nhm_inst
035	—  a Catkey 11948
035	—  a (Sirsi) 11310
040	—  a UK-LoNHM  b eng  e rda
210	00  a J. Zool. Lond.
245	00  a Journal of Zoology :  b Proceedings of the Zoological Society of London.
246	10  a Proceedings of the Zoological Society of London
246	13  a Journal of Zoology (London, England: 1987)
264	_1  a London :  b Published for the Zoological Society of London by Academic

Holdings (1 - 5 of 5)

Sort by: **Alma Ranking** ▼

1	South Kensington / <b>Zoology Serials</b> Call Number SERIALS S 1A	Holdings ID 22145389340002081 ● Vol. 146 (1965) - Vol.315 (2021)
2	Tring / <b>Tring Serials</b> Call Number SERIALS S 130 A	Holdings ID 22162403190002081 ● Vol. 146-236 (1965-1995)
3	South Kensington / <b>Entomology Serials Wandsworth</b> Call Number SERIALS S 18	Holdings ID 22162149440002081 ● Vol. 146-180 (1965-1976)
4	South Kensington / <b>Zoology Discard</b> Call Number MAMMALS 17.1	Holdings ID 22160453480002081 ● 1 items out of 1 available
5	South Kensington / <b>Zoology Mollusca</b> Call Number MOLLUSCA S 8	Holdings ID 22226973490002081 ● 1 items out of 1 available

## Holding

## Item

<input type="checkbox"/>	Barcode	Library	Location	Item Call Number	Call Number	Volume	Description	Year	ora ry Loc ati on	Status	Proces s type	Acce ss Num ber	Receivin g date
<input type="checkbox"/>	000399958	South Kensi...	<b>Zoology</b> Serials	-	SERIALS S 1A	Vol.315	Vol.315 No.4 (2021 Dec)	2...	No	Item in place	-	-	05/01/...
<input type="checkbox"/>	000337976	South Kensi...	<b>Zoology</b> Serials	-	SERIALS S 1A	Vol.315	Vol.315 No.3 (2021 Nov)	2...	No	Item in place	-	-	08/12/...
<input type="checkbox"/>	000353253	South Kensi...	<b>Zoology</b> Serials	-	SERIALS S 1A	Vol.315	Vol.315 No.2 (2021 Oct)	2...	No	Item in place	-	-	02/11/...
<input type="checkbox"/>	000354125	South Kensi...	<b>Zoology</b> Serials	-	SERIALS S 1A	Vol.315	Vol.315 No.1 (2021 Sep)	2...	No	Item in place	-	-	28/09/...
		South	<b>Zoology</b>				Vol.314			Item in			

Figure 1: Serial inventory

identified as articles by Primo VE which would confuse the user even further<sup>1</sup>. All this has meant that, although there is a huge potential benefit to the user of having these records in the catalogue, this material is often very hard to find and obtain.

Since the move to Alma we have been addressing this problem on a record-by-record basis – utilising the MARC21 773 field to add title and, where available, barcode and bibliographic ID (MMSID) to relate the article and serial records. Alma takes the information entered in a 773 and will use this to generate holdings information on the part record. Which means that once a 773 is added with item level information it will display on the article record in Primo and hence direct the user to the serial volume they require (see [Figure 2](#)).

**Details**

Author: [Birkhead, Mike.](#) >

Title: Causes of mortality in the mute swan *Cygnus olor* on the River Thames  
M. Birkhead.

Format: p. 15-25 : ill.

Series: [Journal of zoology 0952-8369 ; Vol. 198, 1982](#) >  
[Journal of Zoology Vol. 198, 1982](#) >

Language: English

General note: 000196827

Taxonomic Name: *Cygnus olor*. Mute Swan

**Locate Item**

NHM staff sign-in for request options  
External visitors click [here](#) for requesting items. [NHM Staff Sign in](#)

[← BACK TO LOCATIONS](#)

**LOCATION ITEMS**

South Kensington

→ RELATED TITLE: *Journal of Zoology : Proceedings of the Zoological Society of London.*

May be available, Zoology Serials ; SERIALS S 1A

Note: Binding for Vols 211-215 incorrectly indicates Series A

Holdings: Vol. 146 (1965) - Vol. 315 (2021)

Item in place (0 requests) Vol.198 (1982)

Loanable Item Barcode: 000196827

**Figure 2:** Example of a linked record

This link also means that if this serial is moved or another enumeration is updated, it won't affect the link, meaning a dynamic connection can be maintained. This is an elegant solution and, in this system, has discovery benefits. However, it is quite labour intensive as it requires work across item, holding and bibliographic level records. We

<sup>1</sup> This is down to a quirk of how Alma and PrimoVE recognise articles and a misreading of the MARC21 guidance on the a vs. b indicator in LDR/07. The table detailing how the Primo VE article resource type is mapped from MARC21 LDR/07 bibliographic level indicates that the only value that will generate a resource type of article in Primo VE is "b" – serial component part ([Ex Libris, no date b](#)). However, this refers to a serialised component such as an editorial in a newspaper rather than a component part of serial. The value of "a" – monographic component part – should also be a valid mapping to the article resource type and indeed the vast majority of articles in journals are themselves monographic, so would be properly encoded as LRD/07 = "a" ([Library of Congress, 2016](#)).

have calculated that it takes approximately 10 minutes to complete each link. As there are approximately 100,000 links to make, simply working through them manually would take over 5 years of work which is clearly not ideal!

To try and work through this issue we have established some semi-automated approaches using the functionality of Alma to delete items and holdings en masse. However, the application of linking the child record to the parent's based on the citation has always proved elusive. It was this that lead us to thinking about some more computational approaches and to contact the NHM's AI team.

### Working with the AI Team

We developed a two-stage pipeline to link child records with their corresponding parent records (e.g., journals or series) using natural language processing and record linkage techniques ([de Bruin, 2023](#)). The goal was to match child items with appropriate parent candidates based on title similarity and associated metadata, including information such as year of publication, volume, and issue number.

#### Stage 1: Metadata Extraction and Normalisation

The initial step involved normalising and extracting structured metadata from both parent and child datasets. In the parent dataset, metadata was embedded within a free-text description field. This field was parsed using a rule-based NLP pipeline built on the spaCy library ([Honnibal, Montani, Van Landegem and Boyd, 2024](#)) specifically employing the Matcher component to identify patterns corresponding to metadata elements such as volume numbers, publication years, and issue identifiers.

In the child dataset, titles and the 773 field were similarly parsed. The series titles were processed using the same spaCy-based rule system, while additional metadata embedded in the 773 field was extracted using regular expressions. All titles were normalised (e.g., lowercasing, punctuation stripping), and any metadata extracted was retained as key-value pairs for downstream matching.

#### Stage 2: Candidate Matching and Best-Fit Selection

The second stage focused on identifying the most likely parent record for each child item. This was carried out using the RecordLinkage library, which provides tools for probabilistic and deterministic record linkage ([de Bruin, 2023](#)).

We applied a hybrid fuzzy matching strategy that leveraged both Jaro-Winkler similarity and partial ratio string matching ([Bachmann, 2024](#)). For each child title, candidate parent records were selected if they had either:

- The highest Jaro-Winkler similarity to the child title, or
- The highest partial ratio similarity, provided their Jaro-Winkler score was within 0.2 of the top match.



This ensured tolerance for partial or noisy matches without sacrificing precision.

Among the shortlisted candidates, a deterministic matching step compared each child record's structured metadata against that of the parent candidates. This comparison was conducted using a strict multi-field match: any candidate for which a field (e.g., volume, year) did not match the corresponding child field was excluded. If multiple candidates passed this filter, the candidate with the highest number of matching fields was selected as the best match.

## Evaluation and Performance

To assess the performance of the parent-child linking system, a manually curated gold-standard dataset was assembled by the library team. This dataset comprised approximately 3,000 child records that had been manually linked to their corresponding parent journal entries. The pipeline achieved an overall accuracy of 68% when evaluated against this gold-standard dataset. Accuracy was defined as the proportion of child records for which the system's predicted parent matched the manually assigned parent. This result suggests that the title-based matching combined with deterministic metadata comparison performs reasonably well in structured contexts.

When applied to the full, real-world dataset — for which no gold-standard manual labels were available — the system yielded predictions for approximately 60% of child records. Analysis of the remaining 40% of unmatched cases revealed that the series title in the child record did not sufficiently resemble any parent title, resulting in no candidates being passed forward to the deterministic metadata comparison stage. This points to limitations in the fuzzy matching threshold and strategy, which may fail to accommodate certain types of string variation, abbreviation, or inconsistency.

Despite these limitations and given the amount of records that we need to work with, these matching percentages are excellent and we are currently working on ways to try and improve the matching with some human tweaking and intervention. Once the model has output its predicted links, we can update our records utilising a combination of MarcEdit and OpenRefine to add 773 fields to each child record, linking it to its respective parent item. We can then run jobs in Alma to remove the item and holdings records associated with these child records. What we are left with are article records linked both in the bibliographic record but also through inventory, thus allowing our users to find and request the precise volumes they need to conduct their research (see [Figure 2](#) above).

## Conclusion

This paper has demonstrated a practical and effective approach to addressing a long-standing challenge in the NHM's library catalogue: linking of article-level records to their parent journal holdings. The traditional manual process of creating these vital connections was prohibitively time-consuming, posing a significant barrier to user

access and efficient resource management. By developing a two-stage pipeline leveraging natural language processing and advanced record linkage techniques, we have successfully semi-automated this complex task.

This project not only enhances the discoverability and obtainability of valuable research materials for our users but also provides a dynamic, resilient linking mechanism that adapts to changes in physical holdings. More broadly, this work exemplifies how computational methods can empower metadata professionals to tackle large-scale, repetitive tasks with limited resources, freeing up human expertise for more nuanced and strategic cataloguing efforts. We believe this methodology holds significant potential for extension to other types of complex record relationships, ultimately fostering a more accessible and user-centric library catalogue.

## References

- Bachmann, M. (2024) *rapidfuzz/RapidFuzz* (v3.11.0) [Software]. Zenodo. Available at: <https://doi.org/10.5281/zenodo.14509091> [Accessed: 31 May 2025]
- Breeding, M. (2007) 'Next-Generation Library Catalogs: Chapter 1 Introduction', *Library Technology Reports*, 43(4). Available at: <https://librarytechnology.org/document/18344/> [Accessed: 31 May 2025]
- British Film Institute (no date) *Simple Search*. Available at: <https://collections-search.bfi.org.uk/web/search/simple> [Accessed: 31 May 2025]
- de Bruin, J. (2023) *Python Record Linkage Toolkit: A toolkit for record linkage and duplicate detection in Python* (v0.16) [Software]. Zenodo. Available at: <https://doi.org/10.5281/zenodo.8169000> [Accessed: 31 May 2025]
- Ex Libris (no date a) *Introduction to Alma Inventory*. Available at: [https://knowledge.exlibrisgroup.com/Alma/Product\\_Documentation/010Alma\\_Online\\_Help\\_\(English\)/040Resource\\_Management/050Inventory/010Introduction\\_to\\_Alma\\_Inventory#](https://knowledge.exlibrisgroup.com/Alma/Product_Documentation/010Alma_Online_Help_(English)/040Resource_Management/050Inventory/010Introduction_to_Alma_Inventory#) [Accessed: 31 May 2025]
- Ex Libris (no date b) Mapping to the Display, Facets, and Search Sections in the Primo VE Record. Available at: [https://knowledge.exlibrisgroup.com/Primo/Product\\_Documentation/020Primo\\_VE/Primo\\_VE\\_\(English\)/120Other\\_Configurations/Mapping\\_to\\_the\\_Display%2C\\_Facets%2C\\_and\\_Search\\_Sections\\_in\\_the\\_Primo\\_VE\\_Record#MARC21\\_and\\_KORMARC\\_Resource\\_Type\\_Mapping](https://knowledge.exlibrisgroup.com/Primo/Product_Documentation/020Primo_VE/Primo_VE_(English)/120Other_Configurations/Mapping_to_the_Display%2C_Facets%2C_and_Search_Sections_in_the_Primo_VE_Record#MARC21_and_KORMARC_Resource_Type_Mapping) [Accessed: 31 May 2025]
- Honnibal, M., Montani, I., Van Landeghem, S., and Boyd, A. (2024) *spaCy: Industrial-strength Natural Language Processing in Python* (v3.7.5) [Software]. Zenodo. Available at: <https://doi.org/10.5281/zenodo.1212303> [Accessed: 31 May 2025]
- Library of Congress (2016) *Leader (NR)*. Available at: <https://www.loc.gov/marc/bibliographic/bdleader.html> [Accessed: 31 May 2025]
- Riva, P., Le Bœuf, P., Žumer, M. (2024) *IFLA Library Reference Model: A Conceptual Model for Bibliographic Information*. International Federation of Library Associations and Institutions (IFLA). Available at: <https://repository.ifla.org/handle/20.500.14598/40.2> [Accessed: 31 May 2025]



Sonawane, C.S. (2017) 'Library Discovery System: An Integrated Approach to Resource Discovery', *Informatics Studies*, 4(3), 27-38.