A Corpus Linguistic Analysis of Catalogue Data: Understanding Curatorial Practice Over Time

Rossitza Atanassova (British Library) and James Baker (University of Southampton)

In this chapter we discuss the application of corpus linguistic approaches to printed catalogue data and how they open up new areas of enquiry into cataloguing and curatorial practice over time. While similar methods are being adopted by scholars and practitioners in a range of fields, here we are primarily concerned with research in libraries that aims to develop collections metadata processes and practice for description, discovery, enrichment and computational use.

Our use case is Atanassova's project 'Legacies of Curatorial Voice in the Descriptions of Incunabula Collections at the British Library and Their Future Reuse' undertaken between 2022-23, with Baker as her mentor. This research was funded under the Research Libraries UK and Arts and Humanities Research Council Professional Practice Fellowship scheme that enables library professionals to be active participants in research (RLUK, 2024?) and was one of ten funded projects in the first year of the scheme.[1] The project took a 'catalogues as data' approach by using corpus linguistics methods to analyse legacy catalogue descriptions of books printed before 1500, with the ambition of gaining new insights into both the catalogues and the collections they describe. The motivation for undertaking this project was to use and embed the research methodology within British Library (BL) operations as a means of demonstrating the impact of digital scholarship applied to collection catalogues.[2]

The sections below cover the library context for our approach, the processes and tools used to prepare the data, results from exploratory corpus linguistic analysis and a reflection on our experience of working with legacy printed catalogues as data.

[1] RLUK (2022) AHRC-RLUK Professional Practice Fellows announced, https://www.rluk.ac.uk/ahrc-rluk-professional-practice-fellows-announced/ accessed 26/01/24

[2] The Fellowship project was inspired by Legacies of Catalogue Descriptions (2020-2023) https://catalogueleg acies.github.io/ accessed 26/01/24 Aptly, the Principal Investigator of this project acted the mentor for the fellowship project.

1. The context: Collections Catalogues and the British Library

Efforts like ours to enable the (semi) automated generation of library catalogue data take place in the context of longstanding institutional efforts to produce, maintain, and provide access to catalogue data. The British Library's first collection metadata strategy 'Unlocking the Value' published in 2015 laid out the principles for the effective stewardship of metadata as a key organisational asset in its own right and paved the way for the adoption of automated workflows to upgrade, enhance and share open metadata (Wilson, 2019). The follow-on 'Foundations for the Future' 2019-2023 collection metadata strategy set out clear priorities for maintaining and developing the collection metadata assets, including to enhance the quality of existing information through automated generation of metadata and crowdsourcing and to make visible hidden resources through digitisation and extraction of data from printed collection catalogues. Such activities have improved the discoverability and accessibility of collection metadata and support the Library's mission to 'make our intellectual heritage accessible to everyone, for research, inspiration and enjoyment' (British Library, 2023a).

A very early example of such efforts was the publication in 2011 of the Linked Open British National Bibliography, a resource that contains approximately five million records in machine-readable structured data formats that allow for the machine reuse of bibliographic information.[3] More recently, Victoria Morris - Metadata Analyst at the British Library - used machine learning techniques to enhance over four million catalogue records with information about the language of the resource described (Morris, 2019). Alongside this, there has been ongoing work across the Library to augment the online catalogue records for digitised printed collections with annotations about subjects, genre, named entities, and geographical information, often created initially as a result of research collaborations and crowdsourcing initiatives (British Library, 2021). Additionally, AHRC funded projects such as Locating a National Collection[4] have enabled the

[3] (2023) Share Family: British National Bibliography (Beta) service is live. *The British Library Digital Scholarship Blog,* July 14. https://blogs.bl.uk/digital-scholarship/2023/07/share-family-british-national-bibliography.html See also The Linked Open British National Bibliography https://www.data.gov.uk/dataset/6fa6a421-e515-4ab7-bbd6-1e60c2b60706/the-linked-open-british-national-bibliography accessed 31 March 2024

[4] https://britishlibrary.github.io/locating-a-national-collection/

Library and other cultural heritage institutions to explore the application of new methods for connecting and engaging the public with cultural heritage (Rees et al., 2022).

By working with catalogue records as data and by applying different digital methods and tools to those records, institutions and communities are generating new insights into the collections that these records describe. In 2019, Yann Ryan - then Newspaper Data Curator at the British Library - published a title-level list of British and Irish newspapers held by the Library which he updated with information about changes to titles over time, dates of publication, and the availability or otherwise of digital surrogates.[5] This work stimulated new research into the collection. For example, the flagship Humanities Data Science project Living With Machines (2023) used this title-level list as a starting point for identifying candidate historical sources for ingest and analysis. The list has also facilitated the creation of new visualisation tools to explore collections (Press Tracer (Vane, 2021)) and support selection for digitisation (Press Picker[6]) and prompted the development of new resources for working with cultural heritage data (Ryan, 2021). Such approaches to catalogue records can reveal new patterns and highlight new or missing information in the data which in turn can better inform understanding of the scope of the collection, including how well described or accurate the catalogue is, and how representative it is of what was published at particular historical moments. Through this cycle of catalogues as data work we gain a more secure grasp of how the catalogue may introduce potential bias when used in research, in particular when catalogue data is used as a proxy for the collection it documents and describes.

Importantly, the above approach to catalogues as data facilitates the delivery of the British Library's "Enacting Change", Race Equality Action Plan (REAP) (British Library, 2022). Embedded in *REAP* are best practice recommendations for cataloguing and metadata, the implementation of which seeks to enhance the safety, inclusivity and accessibility of the Library's collections for all potential users. Following this work funding was obtained for the Metadata and Provenance Pilot Project (2023-25) to co-develop sustainable workflows for inclusive descriptive

[5] https://bl.iro.bl.uk/concern/datasets/7da47fac-a759-49e2-a95a-26d49004eba8 For the full title list of catalogued newspapers held by the British Library see https://bl.iro.bl.uk/concern/datasets/943bd083-6355-44a1-97eb-b8ff898f87d5

[6] https://github.com/Living-with-machines/PressPicker_public

practice and collections discovery (Danskin, 2023). This initiative brings together metadata analysts, cataloguers, and curators to explore how the Library catalogues can best reflect the historical, cultural, and material contexts of the collections the Library holds. The project will also make use of and further develop tools such as Carissa Chew's community maintained 'Inclusive terminology: guide & glossary for the cultural heritage sector',[7] thus contributing findings from the Library's internal-focused work to a wider network.

Doing this work requires Library staff to develop new skills and to use new tools for working with catalogues as data. The Library's Digital Scholarship Training Programme, managed and run by the Digital Research Team, has for the last decade offered an average of 50-60 events a year attended by over a thousand individual staff members, with more targeted training for those who work regularly with digital content (British Library, 2023b). In the video[8] celebrating the 10th anniversary of the Programme, Digital Curator Nora McGregor explains how the training has given staff the confidence to put into practice the skills and techniques they have learnt about and be part of new internal and external digital research collaborations.

The Library's Digital Curators have done ground-breaking work to trial and implement OCR/HTR technology for the automated transcription of data in non-European languages, enrich the Library's collections metadata and support digital humanities initiatives (Keinan-Schoonbaert, 2024). They have also led developments to automate the workflows around crowdsourced metadata using technologies such as IIIF (Ridge, 2023). As part of innovative research collaborations such as Living with Machines, the Library hired project staff who contributed to the development and delivery of training on using collections data with Python, NLP, Machine Learning, AI, Computer Vision, for the benefit of the wider cultural heritage sector (Living With Machines, 2023). The RLUK Statement of Support for the Technician's Commitment (RLUK, 2023), which celebrates the importance of technical expertise to the research process, has strengthened the case at the Library and in the wider sector for more Research Software Engineers roles, that bring the much

[7] Chew Inclusive Terminology Glossary (2023) http://itg.nls.uk/wiki/Introduction
[8] YouTube (2023) Introducing the British Library Digital Scholarship Training Programme https://www.youtube.com/watch?v=UwKFb6Uzui8

needed programming and computational skills to collections curation and research, as demonstrated in the next section (Lloyd, 2023).

2. The *BMC*: from Catalogues to Data

The primary objective of the Fellowship project was to explore how corpus linguistic analysis of catalogue data could be used to better understand the language of cataloguing and curatorial practice over time. The 13 volume *Catalogue of books printed in the 15th century now at the British Museum* (hereafter *BMC*) was identified as an ideal candidate for this research. This printed catalogue contains detailed descriptions of the Library's internationally important incunabula collection and is an essential finding aid and scholarly resource for users. Initially published between 1908 and 2007, most volumes of the *BMC* have subsequently been made available as searchable pdfs, but the relevant catalogue entries have not been integrated into the Library catalogue.[9] Thus, the project would bring the additional benefit of preparing data from the *BMC* to enrich the Library's catalogue.

The project began by testing and implementing a semi-automated workflow for, first, identifying and extracting individual catalogue entries from OCR/HTR files derived from digitised images using the AI-powered *Transkribus* platform, and, second, identifying and extracting the relevant information for analysis. The output of this process were volume level text files assembled as a corpus for the first 10 volumes of the printed *BMC*.[10] Several intermediate steps were required to prepare the corpus.[11] Following digitisation, page images were divided according to page layout. An annotated subset was then created and used to train two different structure recognition models.

[9] The lithographic reprint of vols.1-8 published in 1963 reproduced the Museum's working copy, with 'numerous manuscript additions and corrections by various hands, the majority being by Dr. Victor Scholderer.' See note to the edition by R.A. Wilson. We therefore had to scan a clean copy of the catalogue to remove the noise from the annotations in the curatorial copy from which the pdf was created. The newly digitised images were published to the Library's IIIF-compliant Universal Viewer, however access has been disrupted since October 2023 due to a ransomware attack.

[10] Vol. 11 has substantially longer entries with new subheadings, vol.12 requires further correction of the layout analysis and vol. 13 requires Hebrew text recognition.

[11] Due to the word limit we have not included information about the initial tests with the commercial text recognition software Abbyy.

Whilst this was a laborious and lengthy part of the process, which took an estimated 80 hours, it was a vital step which ensured the catalogue records were extracted in the correct reading order across multiple columns. Due to inconsistency in the printed layout and the quality of some digitised images, we did not achieve 100% text recognition accuracy,[12] but the output was of sufficient quality for corpus linguistic analysis, the strength of which is to support analysis that illuminates the structures and distinguishing features of a corpus (as opposed to fine grained content analysis, where high levels of text accuracy is vital).

Next, individual catalogue records were identified and extracted in an automated way.[13] The record structure had several consistent features, and after some testing and iteration (Dunford, 2023),[14] a programmatic rule was developed that split individual catalogue entries using the block letter headings at the start of each entry and a shelfmark at the end of the entry. These split entries were then saved as separate text files, giving a series of catalogue entry level text files assembled as a corpus.

To filter these entry level files for content that represented the cataloguing and curatorial practice, language detection was then used to encode passages of text within catalogue entries. This step was necessary as catalogue entries in the *BMC* typically include both information about incunabula and transcriptions of text from incunabula. Much of the latter is written in Latin, Greek, old English, or other non-Anglophone languages. As we were interested in analysing text produced as part of cataloguing and curatorial practice – including *inter alia* contextual information and copy-specific notes on an acquisition, provenance, or binding – language detection enabled the creation of a clean corpus. Further, by replacing non-interpretive text with a placeholder [Non-English

[12] The pages were scanned as double openings which introduced some curvature/warping or loss of text in the gutter and although the images were flattened during post-processing, there was still some inconsistency which affected the layout recognition and therefore the text recognition.
[13] This was undertaken by Isaac Dunford, postgraduate student with programming skills, as part of his Digital Humanities internship at Southampton University. See https://github.com/Southampton-Digital-Humanities/2023_Catalogue-Entry-Detection
[14] The code is maintained on GitHub at https://github.com/britishlibrary/Incunabula-Catalogue-Entry-Detection/tree/hny

section lasting x lines], the flow of catalogue entries between cataloguer/curator created text and exposition through transcription was maintained.

The final stage of data preparation involved recombining the entry level data at volume level.[15] This stage balanced requirements for text analysis with plans for enrichment of the online catalogue records. In order to facilitate both reproducibility and code review, the final output was published as a dataset accompanied by a csv file listing the shelfmarks of each volume and raw data used for the information extraction (i.e. the images and XML files referenced in the csv file).[16]

The use of AI-based tools and approaches were central to transforming the *BMC* into structured and actionable data. Specifically, Transkribus was used to adapt and semi-automate the process of text recognition.[17] There was manual labour involved in the process of selection and annotation of images to create structure models suitable for our content: we annotated 400 images in order to process 3,000 images. We had to make decisions about how many annotated images to provide for Transkribus to train the models, and finding a balance between the effort of providing training data for machine learning and the benefits of automation would vary from project to project and will depend on the type of content and available resources. The more variation there is in the data, the higher the effort will be, and the results will need careful assessment. For the double-columned layout model we did not see any significant improvement after the third training set (50 images per set), whereas in the case of the mixed layout model (single/horizontal and double/vertical text blocks) we noticed some deterioration when we increased the volume of the training dataset from 250.[18] The results are also dependent on how the Transkribus tool is used, which settings are

[15] This work was completed by the Research Software Engineer with the Digital Research Team at the Library working in close collaboration with the project lead and curator for the incunabula collection.
[16] See the incunabula catalogue dataset with raw and processed files at https://bl.iro.bl.uk/collections/a0a057dd-bd00-414a-ba2c-9fe61ee6fba0?locale=en For information see the readme.txt.
[17] https://www.transkribus.org/ We used the Transkribus Print 0.3 language model for the text recognition.
[18] This was probably due to the inaccuracies in the print layout and to some extent the imaging quality. We only inspected the images visually and did not use any scientific metrics

selected in the training model option, the base model (we used the P2PaLa model[19]) and any improvements. It was important to separate the training from the test data for the evaluation of the models. During the project the Transkribus web interface was developed incrementally, and the annotation and evaluation tasks became easier.[20]

The more challenging part of the process was the development of the algorithm to detect the separate catalogue records (entries) which involved implementing rules coded with python.[21] The approach selected worked with the PAGE XML files output by Transkribus, making use of the XML mark-up tags for the heading in upper case letters, and the record number (in a regular format eg. IA/IB/IC followed by digits). Some adjustments to the code had to be subsequently made to deal with where there were other copies following the main catalogue record for which the heading was not repeated. The algorithm was also able to reassemble entries which were split across two different pages (XML files) and reject results which seemed to be false positives (where the record appears to be too long).

Closer examination identified some inaccuracies in the outputs and required iterative code development. There were instances of an incorrect reference number having been picked by the algorithm (other references in the text), or some anomalies in the format of the reference number (Proctor, 1898)[22] where it includes a small letter after the digits, or where the reference number is different from the shelfmark for the item as used in the online catalogue. This work required human labour and curatorial knowledge of the history of the collection, how it has been organised and

[19] https://readcoop.eu/transkribus/docu/p2pala/

[20] During the project the Transkribus app was upgraded and since then new features, such as field and table recognition, were introduced which could have improved the layout recognition. We processed the images between Nov 2022 and February 2023 and used different versions of the platform, from v.0.8.6 to 1.2.0.

[21] See this notebook created by Lloyd, H. for a demonstration of how the code works: https://github.com/britishlibrary/Incunabula-Catalogue-Entry-Detection/tree/v1.0.0

[22] Proctor devised the index to the collection assigning reference numbers that typically begin with IA, IB or IC (the second letter referring to the size of the volume) followed by a digit. The collection was systematically arranged under the name of each country, town, and printer, in chronological order.

developed, catalogued and searched.[23] This part of the work was particularly important for the preparation of the catalogue data as metadata for ingest into the online catalogue. Some of the improvements were achieved by shared tasks using spreadsheets.

For the language detection of English-only text in the record we used langdetect python package[24], the outputs were not verified at this stage and could be managed as part of the analysis with AntConc. The text of greatest interest for the analysis is contained in the section introduced by the item dimensions and has copy-specific information. This is also the information that the curators are interested in updating the online catalogue with.

3. Towards data analysis for catalogue enrichment

To analyse the incunabula descriptions corpora and to test the viability of corpus linguistics methods in curatorial/documentation settings (e.g. to develop new insights into both the catalogues and the collections they describe), we used AntConc[25], a GUI software platform widely used in corpus linguistics and adjacent research communities. 45,595 unique tokens were identified in the *BMC* corpus (case not selected), 35% of which were accounted for by the most frequent 30 words. Significantly, rather than comprising function words, many of these 30 most frequent words (Figure 1) were those used to identify incunabula, suggesting the use of specialist language - for example, information relating to the number of leaves and precise measurements ('mm') of the full page and/or text block; the number of lines contained on a page; notes on first, blank (or missing) leaves; capitals, rubrication, or illumination; and provenance information. Further, we find formulations relating to referencing conventions within incunabula cataloguing: 'IA' and 'IB' ('Proctor numbers'); 'Hain' introduces a reference number in a canonical bibliographical source for incunabula (Hain's Repertorium); and 248 of the 'G's introduce the shelfmark used for the Thomas Grenville collection of incunabula.

[23] The collections of King George III (King's Library), of Thomas Grenville and Revd. C.M. Cracherode and few other volumes retained their original shelfmarks.
[24] https://pypi.org/project/langdetect/
[25] Laurence Anthony's Website. AntConc https://www.laurenceanthony.net/software/antconc/ For our analysis we used AntConc v.4.2.

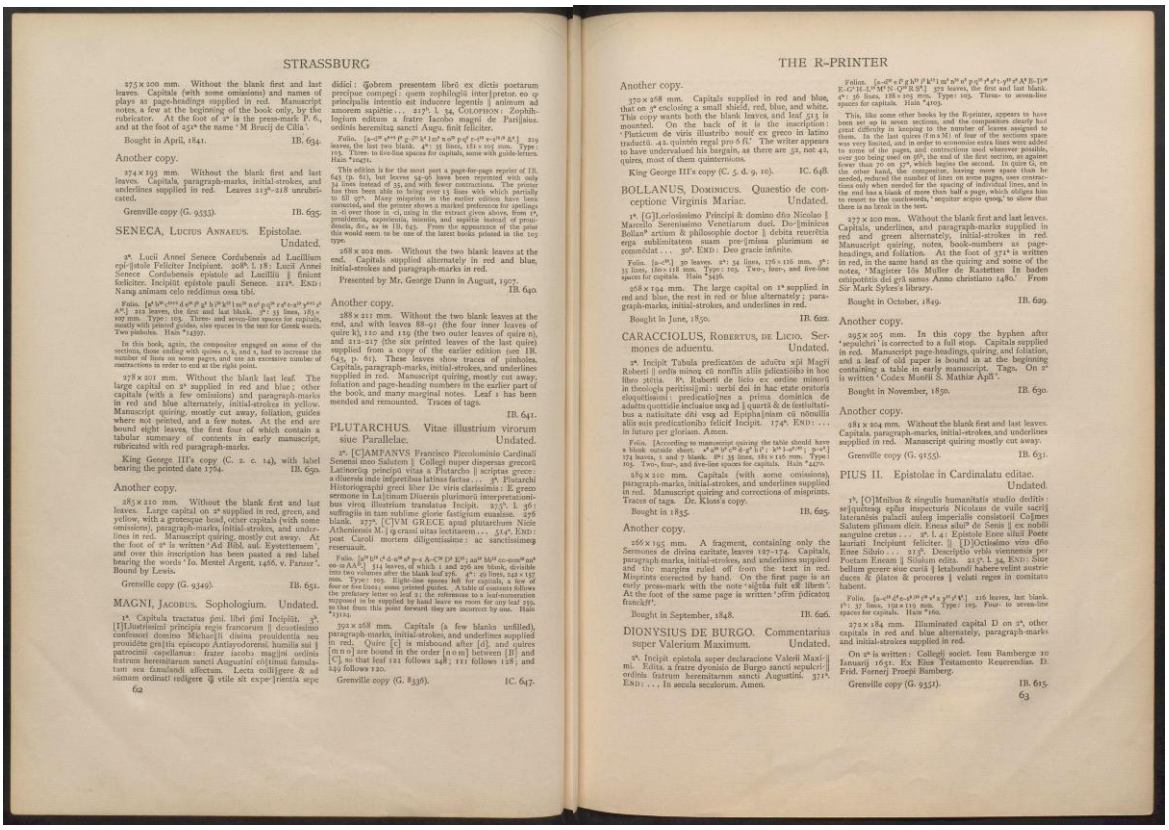
Of the common function words in the *BMC* corpus, prepositions were - compared with everyday speech - used with atypical frequency, for example in helping the reader locate specific features of a text (e.g. 'on folio') or in providing explanatory information (e.g. 'edited by' or 'owned by'). The function words at the start of a sentence introduce in a concise way important information about the object or its contents. The majority of uses of "With" (1588 occurrences) relate to ownership references - 'With the bookplate/bookstamp/note of ownership', while others note the presence of woodcuts or marginalia. Missing information about the object is equally important and begins with 'Without' (2392 occurrences) when used to indicate the absence of a blank leaf or blank leaves present in other copies.[26] In the phrase 'Without the blank leaf/leaves' was so well-established that from volume 5 onwards it is abbreviated to 'Without the blank(s)', suggesting attention to both economy of language and print space. This cataloguing practice of recording 'absences' is important to bibliographic research with early printed collections and the 21st century cataloguers of *BMC* elaborate on this practice.

Rank	Type	Frequency	Rank	Type	Frequency	Rank	Type	Frequency
1	the	58485	11	lines	10485	21	spaces	5847
2	of	32514	12	Leaves	8902	22	capital	5796
3	in	28271	13	Blank	8241	23	capitals	5555

[26] In volumes 1-3 'Wanting' is used to note missing leaves, and elsewhere 'wanting' (1169) as part of the string 'Imperfect, wanting …' regarding various contents missing in a volume.

4	and	25311	14	B	7937	24	by	5415
5	a	18015	15	To	7524	25	title	5360
6	mm	17350	16	G	6903	26	red	5324
7	x	16740	17	text	6507	27	first	5288
8	on	13955	18	hain	6246	28	type	5267
9	with	12705	19	Ia	5913	29	bought	5170
10	is	11544	20	Ib	5855	30	head	4846

Figure 1: The 30 most frequent tokens or word types (setting no case selected) in the *BMC* corpus. In this list the single letter 'b' refers to the pagination (or signatures) and 'x' refers to both measurements and pagination. 'Type' is just outside the 30 most frequent words when the 'case' setting is selected



STRASSBURG

275 x 200 mm. Without the blank first and last leaves. Capitals (with some omissions) and names of plays as page-headings supplied in red. Manuscript notes, a few at the beginning of the book only, by the rubricator. At the foot of 3ª is the press-mark P. 6, and at the foot of 251ª the name 'M Brucij de Cilia'.

Bought in April, 1841. IB. 634.

Another copy.

274 x 193 mm. Without the blank first and last leaves. Capitals, paragraph-marks, initial-strokes, and underlines supplied in red. Leaves 213ª–218 ururubricated.

Grenville copy (G. 9353). IB. 635.

SENECA, Lucius Annaeus. Epistolae. Undated.

3ª. Lucii Annei Senece Cordubensis ad Lucilium epi|stole Felicter Incipiunt. 2078. I. 18: Lucii Annei Senece Cordubensis epistole ad Lucilliū ‖ finiunt feliciter. Incipiūt epistole pauli Senece. 211ª. END: Nang animam celo reddimus ossa tibi.

Folio. [a⁸ b¹⁰ c¹⁰⁺¹ d–e⁸ f⁸ g⁶ h⁸ i¹⁰ k¹⁰ l–m⁸ n–o⁸ p–q¹⁰ r–s¹⁰ t¹⁰⁺¹⁰ A–C⁸] . 212 leaves, the first and last blank. 3ª: 35 lines, 183 x 107 mm. Type: 103. Three- and seven-line spaces for capitals, mostly with printed guides, also spaces in the text for Greek words. Two pinholes. Hain *14397.

In this book, again, the compositor engaged on some of the sections, those ending with quires e, k, and t, had to increase the number of lines on some pages, and use an excessive number of contractions in order to end at the right point.

278 x 201 mm. Without the blank last leaf. The large capital on 2ª supplied in red and blue; other capitals (with a few omissions) and paragraph-marks in red and blue alternately, initial-strokes in yellow. Manuscript quiring, mostly cut away, foliation, guides where not printed, and a few notes. At the end are bound eight leaves, the first four of which contain a tabular summary of contents in early manuscript, rubricated with red paragraph-marks.

King George III's copy (C. 2. c. 14), with label bearing the printed date 1754. IB. 650.

Another copy.

285 x 210 mm. Without the blank first and last leaves. Large capital on 2ª supplied in red, green, and yellow, with a grotesque head, other capitals (with some omissions), paragraph-marks, initial-strokes, and underlines in red. Manuscript quiring, mostly cut away. At the foot of 2ª is written 'Ad Bibl. aul. Eystettensem', and over this inscription has been pasted a red label bearing the words 'Io. Mentel Argent. 1466. v. Panzer'. Bound by Lewis.

Grenville copy (G. 9349). IB. 651.

MAGNI, Jacobus. Sophologium. Undated.

1ª. Capitula tractatus ſcñi. libet ſcñi Incipiūt. 3ª. [I]Llustrissimi principis regis francorum ‖ devotissimo confessori domino Michae|li divina providentia seu providēcie gra|tia episcopo Antisyodorensi. humilis sui ‖ patrocinii capellanus: frater Iacob⁹ magistri ordinis fratrum heremitarum sancti Augustini cō|stituti famulatum seu famulandi affectum. Lecta coll|igere & ad sūmam ordinatī redigere q̄ vtile sit expe|rientia sepe didici: ꝯgrobrem presentem librū ex dictis poetarum precipue compegi: quem zophilogiū inter|pretor. eo q̄ principalis intentio est inducere legentis ‖ animum ad amorem sapiēcie . . . 217ª. I. 34. Colophon: Zophilogium editum a fratre Iacobo magni de Pari|siius. ordinis heremitaj sancti Augu. finit feliciter.

Folio. [a–d⁸ e¹⁰⁺¹ f–g⁸ h⁶ i–m⁸ n–o⁸ p–q⁸ r–t⁸ u–v¹⁰ A⁶] . 219 leaves, the last two blank. 4ª: 35 lines, 181 x 105 mm. Type: 103. Three- to five-line spaces for capitals, some with guide-letters. Hain *14471.

This edition is for the most part a page-for-page reprint of IB. 643 (p. 61), but leaves 54–56 have been reprinted with only 34 lines instead of 35, and with fewer contractions. The printer has then been able to bring over 13 lines with which partially to fill 57ª. Many misprints in the earlier edition have been corrected, and the printer shows a marked preference for spellings in -ci over those in -ti, using in the extract given above, from 1ª, providencia, experiencia, intencio, and sapiēcie instead of prouidencia, &c., as in IB. 643. From the appearance of the print this would seem to be one of the latest books printed in the 103 type.

268 x 202 mm. Without the two blank leaves at the end. Capitals supplied alternately in red and blue, initial-strokes and paragraph-marks in red.

Presented by Mr. George Dunn in August, 1907. IB. 640.

Another copy.

288 x 211 mm. Without the two blank leaves at the end, and with leaves 88–91 (the four inner leaves of quire k), 210 and 219 (the two outer leaves of quire n), and 212–217 (the six printed leaves of the last quire) supplied from a copy of the earlier edition (see IB. 643, p. 61). These leaves show traces of pinholes. Capitals, paragraph-marks, initial-strokes, and underlines supplied in red. Manuscript quiring, mostly cut away, foliation and page-heading numbers in the earlier part of the book, and many marginal notes. Leaf 1 has been mended and remounted. Traces of tags. IB. 641.

PLUTARCHUS. Vitae illustrium virorum siue Parallelae. Undated.

2ª. [C]AMPANVS Francisco Picolomineo Cardinali Senensi meo Salutem ‖ Collegi nuper dispersas grecorū Latinorūq principū vitas a Plutarcho ‖ scriptas grece: a diuersis inde infportibus latinas factas . . . 3ª. Plutarchi Historiographi grecī liber De viris clarissimis: E greco sermone in La|tinum Diuersis plurimorū interpretationibus viroq illustrium translatus Incipit. 237ª. I. 36: suffragiis in tam sublime gloriæ fastigium euasisse. 276 blank. 277ª. [C]VM GRECE apud plutarchum Nicie Atheniensis M.|| grauis uitas lectitarem . . . 314ª. END: post Caroli mortem diligentissime: ac sanctissimeq reseruauit.

Folio. [a¹⁰ b⁸] c⁶ d–e¹⁰ f⁸ g–s A–C⁸ D⁶ E¹⁰ ; aa¹⁰ bb¹⁰ cc–mm¹⁰ nn⁸ oo–aaA–C⁸] . 314 leaves, of which 1 and 276 are blank, divisible into two volumes after the blank leaf 276. 4ª: 49 lines, 242 x 157 mm. Type: 103. Eight-line spaces left for capitals, a few of four or five lines; some printed guides. A table of contents follows the prefatory letter on leaf 2: the references to a leaf-numeration supposed to be supplied by hand leave no room for any leaf 259, so that from this point forward they are incorrect by one. Hain *13124.

392 x 268 mm. Capitals (a few blanks unfilled), paragraph-marks, initial-strokes, and underlines supplied in red. Quire [c] is misbound after [d], and quires [m n o] are bound in the order [n o m] between [B] and [C], so that leaf 121 follows 248; 111 follows 118; and 249 follows 120.

Grenville copy (G. 8336). IC. 647.

62

THE R-PRINTER

Another copy.

370 x 268 mm. Capitals supplied in red and blue, that on 3ª enclosing a small shield, red, blue, and white. This copy wants both the blank leaves, and leaf 513 is mounted. On the back of it is the inscription: 'Plinicum de viris illustribꝰ nouiſ ex greco in latino tradicti. 43. quintin regal priv 6 fl.' The writer appears to have undervalued his bargain, as there are 52, not 42, quires, most of them quinternions.

King George III's copy (C. 3. d. 9, 10). IC. 648.

BOLLANUS, Dominicus. Quaestio de conceptione Virginis Mariae. Undated.

1ª. [G]Loriosissimo Principi & domino dño Nicolao ‖ Marcello Serenissimo Venetiarum duci. Do|minicus Bollan⁹ artium & philosophie doctor ‖ debita reuerētia ega subhumitatem suam pre|missa plurimum se commēdat . . . 30ª. END: Deo gracie infinite.

Folio. [a–c⁸] 30 leaves. 2ª: 34 lines, 176 x 116 mm. 3ª: 35 lines, 180 x 118 mm. Type: 103. Two-, four-, and five-line spaces for capitals. Hain *3456.

268 x 194 mm. The large capital on 1ª supplied in red and blue, the rest in red or blue alternately; paragraph-marks, initial-strokes, and underlines in red.

Bought in June, 1850. IB. 622.

CARACCIOLUS, Robertus, de Licio. Sermones de aduentu. Undated.

2ª. Incipit Tabula predicatōm de aduēcu xpi Magri Roberti ‖ ordīs minoꝝ eī nonſlis aliis ḋdicatiōbꝰ in hoc libro ꝑfcta. 8ª. Roberti de licio ex ordine minoꝝ in theologia peritissi|mi: verbi dei in hac etate oratoris eloquētissimi: predicatiōnes a prima dominica de aduēcu quotidie inchoate usq ad ‖ quartā & de festiuitatibus a natiuitate dñi vsq ad Epipha|niam cū nōnullis aliis suis predicationibꝰ feliciꝰ Incipit. 174ª. END: . . . in futuro per gloriam. Amen.

Folio. [According to manuscript quiring the table should have a blank outside sheet. a⁸ b⁸ c⁸ d–g⁸ h⁴: i⁸ k–l⁸ m⁶: p–x⁸] 174 leaves, 3 and 7 blank. 8ª: 35 lines, 181 x 116 mm. Type: 103. Two-, four-, and five-line spaces for capitals. Hain *4270.

285 x 210 mm. Capitals (with some omissions), paragraph-marks, initial-strokes, and underlines supplied in red. Manuscript quiring and corrections of misprints. Traces of tags. Dr. Klose's copy.

Bought in 1835. IB. 625.

Another copy.

256 x 195 mm. A fragment, containing only the Sermones de divina caritate, leaves 127–174. Capitals, paragraph-marks, initial-strokes, and underlines supplied and the margins ruled off from the text in red. Misprints corrected by hand. On the first page is an early press-mark with the note 'sigētia fuit eē liberi'. At the foot of the same page is written 'pfem ḋdicatoꝝ franckiff'.

Bought in September, 1848. IB. 626.

DIONYSIUS DE BURGO. Commentarius super Valerium Maximum. Undated.

2ª. Incipit epistola super declaracione Valerii Maxi|mi. Edita. a fratre dyonisio de Burgo sancti sepulcri ‖ ordinis fratrum heremitarum sancti Augustini. 371ª. END: . . . in secula secalorum. Amen.

Folios. [a–d⁸ e–f⁸ g⁶ h¹⁰ i⁸ k¹¹ l–m⁸ n¹⁰ o⁸ p–q¹⁰ r–s⁸ t–v⁸ A–B–D–E–G–H–L–M⁸ N–Q⁸ R⁴] . 373 leaves, the first and last blank. 4ª: 36 lines, 188 x 105 mm. Type: 103. Three- to seven-line spaces for capitals. Hain *4103.

This, like some other books by the R-printer, appears to have been set up in seven sections, and the compositors clearly had great difficulty in keeping to the number of leaves assigned to them. In the last quires (f to M) of four of the sections space was very limited, and in order to accommodate extra lines were added to some of the pages, and contractions used wherever possible, over 300 being used on 358, the end of the first section, as against fewer than 70 on 57ª, which begins the second. In quire G, on the other hand, the compositor, having more space than he needed, reduced the number of lines on some pages, uses contractions only when needed for the spacing of individual lines, and in the end has a blank of more than half a page, which obliges him to resort to the catchwords, 'sequitur scipio quasq.' to show that there is no break in the text.

277 x 200 mm. Without the blank first and last leaves. Capitals, underlines, and paragraph-marks supplied in red and green alternately, initial-strokes in red. Manuscript quiring, notes, book-numbers as page-headings, and foliation. At the foot of 371ª is written in red, in the same hand as the quiring and some of the notes, 'Magister Ioꝝ Muller de Rastetten In badem ostiptentia del grā supas Anno christiano 1480.' From Sir Mark Sykes's library.

Bought in October, 1849. IB. 629.

Another copy.

295 x 205 mm. In this copy the hyphen after 'sepulchri' is corrected to a full stop. Capitals supplied in red. Manuscript page-headings, quiring, and foliation, and a leaf of old paper is bound in at the beginning containing a table in early manuscript. Tags. On 2ª is written 'Codex Monfil S. Mathiæ Apd'.

Bought in November, 1850. IB. 630.

Another copy.

283 x 204 mm. Without the blank first and last leaves. Capitals, paragraph-marks, initial-strokes, and underlines supplied in red. Manuscript quiring mostly cut away.

Grenville copy (G. 9155). IB. 631.

PIUS II. Epistolae in Cardinalatu editae. Undated.

1ª. [O]Mnibus & singulis humanitatis studio deditis: se|quēcesq; ecclie inspectura Nicolaus de valle sacrij lateranēsis palacij ap|lici imperialis consistorij Co|mes Salutem plīmam dicit. Encas silui⁹ de Senis ‖ ex nobili sanguine cretus . . . 3ª. I. 4: Epistole Enee siluij Poete lauriati Incipiunt feliciter. ‖ [D]Octissimo viro dño Enee Siluio . . . 213ª. Descriptio urbis viennensis per Poetam Encam ‖ Siluium edita. 213ª. I. 34. END: Siue bellum gerere siue curiā ‖ letabundi habere velint austrie duces & platos & procures ‖ veluti reges in comitatu habent.

Folio. [a–c⁸ d–e⁸ f⁸ g–k⁸ l⁶ m–x⁸ y⁴ z⁸ A⁴] . 216 leaves, last blank. 1ª: 37 lines, 192 x 110 mm. Type: 103. Four- to seven-line spaces for capitals. Hain *160.

272 x 184 mm. Illuminated capital D on 2ª, other capitals in red and blue alternately, paragraph-marks and initial-strokes supplied in red.

On 2ª is written: Collegij societ. Iesu Bambergæ 20 Ianuarij 1632. Ex Elias Testamento Reuerendias. D. Frid. Fornerj Froejt Bamberg.

Grenville copy (G. 9351). IB. 615.

63

Figure 2: *Catalogue of books printed in the 15th century now at the British Museum* Part 1 (1908) (BL shelfmark J/2704.aa.30/1.) pp.62-63 with entries for copies from the Grenville Library and King George III's Libraries. Note the structure of the entries and the brevity and formulaic nature of some of the 'Another copy' descriptions (eg. IB 635). Image: © British Library Board, CC-BY 4.0

If we expand our frame to the 300 most frequent words (case selected) in the *BMC* corpus, we find frequent use of specialist vocabulary relating to bibliographic description, including collation and materiality: *quire(s)* (2886 occurrences),[27] *colophon* (1458 occurrences), *strokes* (1864 occurrences), *Type(s)* (6985 occurrences) and *type* (2535 occurrences), *stamped* (1248 occurrences), *Rubricated* (693 occurrences) and *rubricated* (388 occurrences), *foliation* (847 occurrences), *morocco* (584 occurrences), and *vellum* (489 occurrences). There is also present

[27] There are 3619 instances altogether (regardless of case) including verb forms 'the sheets are *quired* by a consecutive number' or 'manuscript quiring'.

some specialist use of common vocabulary: eg. ‘capitals and underlines *supplied* (3123 occurrences) in red’, *signed* (736 occurrences) often referring to the ‘signature’ letters at the bottom of pages, *head* (4703 occurrences), *strokes* (1864 occurrences) and *foot* (407 occurrences) with reference to types and letter shapes.

the	49974	printed	2822	but	1335	D	801	set	583	than	467
of	32410	Quarto	2785	have	1274	three	798	contents	582	TITLE	464
in	26272	page	2754	commentary	1264	October	795	g	572	Voulliéme	456
and	25280	columns	2753	note	1263	Manuscript	776	II	569	L	452
mm	17329	that	2741	stamped	1248	before	772	volume	567	brown	452
a	14339	Folio	2705	July	1211	words	769	library	566	illuminated	442
is	11526	Type	2704	Another	1171	Undated	760	heading	563	DEVICE	441
with	11096	de	2650	wanting	1169	dated	750	mostly	562	leaded	438
on	10779	which	2646	or	1164	gold	744	November	559	signatures	438
X	10662	e	2610	not	1160	border	739	F	558	label	437
lines	10475	type	2535	his	1120	cuts	736	except	557	painted	435
leaves	8687	Without	2392	some	1115	signed	736	m	555	pigskin	435
The	8507	Capitals	2098	woodcut	1115	q	727	fly	554	V	434
blank	8238	two	2079	only	1104	four	715	out	554	Two	433
to	7486	written	2050	after	1102	June	712	vol	549	December	427
b	6570	quire	1938	d	1101	f	702	I	548	pages	427
text	6413	strokes	1864	found	1088	z	700	Gk	543	paper	426
G	6331	In	1833	beginning	1084	hand	698	Spaces	542	use	426
Hain	6246	c	1820	Imperfect	1079	Part	696	February	538	name	424
x	6078	been	1797	end	1077	leather	695	following	538	Dr	423
IA	5907	p	1748	notes	1054	Rubricated	693	woodcuts	538	between	412
IB	5810	initial	1746	capital	1045	v	693	copies	534	later	412
by	5360	Bound	1727	each	1043	Woodcut	689	present	528	Greek	410
spaces	5305	also	1727	one	1041	IIIs	673	register	519	aa	410
first	5148	marks	1671	This	1040	k	669	col	515	black	410
red	5116	C	1669	S	1033	give	668	Lombards	511	foot	407
Bought	5098	same	1661	underlines	998	marginalia	665	H	510	cover	405
title	4850	headings	1658	part	995	colours	660	probably	509	At	404
Capital	4750	paragraph	1638	From	972	here	657	half	508		

head	4703	has	1632	s	962	number	656	early	507	very	404
last	4343	no	1606	quires	948	ad	650	see	507	signature	402
Types	4281	cut	1605	books	939	arms	646	reprint	505	its	399
line	4219	P	1603	table	938	et	646	all	503	column	392
are	4180	With	1601	used	933	March	645	ii	503	rubricated	388
copy	3979	blue	1597	sheet	928	few	645	containing	499	smaller	388
book	3862	I	1571	manuscript	923	January	641	lower	498	tracts	385
leaf	3729	i	1547	Old	918	up	639	stamp	492	t	382
A	3676	letter	1469	being	907	Gesamtkatalog	637	elsewhere	489	Octavo	381
at	3487	left	1466	tract	893	May	632	vellum	489	made	381
capitals	3445	other	1463	numbers	875	Proctor	618	Library	485	boards	379
as	3437	colophon	1458	h	873	n	613	plate	485	It	378
letters	3354	an	1454	April	866	appears	606	about	484	top	378
for	3272	it	1452	bound	859	small	604	etc	484	y	375
guide	3208	De	1428	George	856	century	598	sheets	483	taken	374
On	3175	numbered	1407	King	850	above	594	over	482	O	370
supplied	3123	IC	1405	foliation	847	those	594	E	479	borders	370
edition	3035	be	1375	second	840	work	594	quiring	478	verses	370
from	2957	was	1371	space	840	so	590	large	476	Woodcuts	369
this	2911	B	1367	date	830	Ja	586	device	475	editions	368
R	2864	M	1364	o	824	morocco	584	below	474	press	366

Figure 3: The 300 most frequent tokens or word types (case setting selected) in the *BMC* corpus

From this initial appraisal, a number of analytical pathways opened up. One related to OCR/HTR quality. In our case, corpus analysis enabled us to note the presence of hyphenated words and the transliteration of some characters (e.g. 'dwarf' misread as 'dwart'). Whilst these errors may be corrected if prepared for reuse in online resources - such as the Incunabula Short Title Catalogue (ISTC), the Material Evidence in Incunabula (MEI) and other databases recording information about 15th-century printed books[28] - they were deemed insufficiently frequent to impede analysis into the character of the corpus.

However, word lists can also support user journeys. For example, users interested in exploring the collection by format could start with the word frequency results for 'folio', 'quarto', and 'octavo'.[29] Searching these words in the *BMC* corpus returns 5872 occurrences, divided between a comparable number of 'quarto' (2785 occurrences) and 'folio' (2706 occurrences) and a small number of 'octavo' (381 occurrences). Similarly, the word lists alone start to illuminate the provenance and acquisition history of the collection through words relating to language describing watermarks,[30] woodcuts, and bookbindings, or the presence of names such as 'Sussex' (262 occurrences) that indicate the prominence in the *BMC* of the books from the Library of the Duke of Sussex.[31] Indeed, provenance information increases in the later volumes as the cataloguers adopted new standard phrases such as: "the note of ownership".

[28] See https://15cbooktrade.ox.ac.uk/project/, https://data.cerl.org/istc/, https://data.cerl.org/mei/
[29] Equally 'x' and 'mm.' (or numbers if selected in the AntConc setting) can be used for analysis of the volumes dimensions, eg. 402 x 260 mm.
[30] Eg. vol. 6 Scholderer remarks: "The watermark (a bird with a very long body) is that found in the latter part of the preceding. (IB. 28454)" or in vol.7, "The rosette is the commonest watermark in Milanese incunabula, while the currycomb and the grapes also occur in Milanese books of the seventies." (IB. 36835.) The cataloguers of vols. 6-8 show preference for the past participle 'paper is *watermarked/unwatermarked*' often defined with adverbs (*mostly*, *partly*) which is a feature of their cataloguing style as discussed below.
[31] In a note to the introduction to vol. 1 p. xxvi we are told that due to the large number of books printed by Ulrich Zel from the Duke of Sussex's library, which was dispersed in 1844, the mention of the Duke's 'press-marked book-plate' was omitted from some of the entries.

For catalogue research, word frequency lists are a useful tool for investigating named entities that appear in catalogue descriptions, which comprise roughly half of the *BMC* corpus.[32] In particular, these names deepen understanding of how bibliographic references were used in incunabula descriptions. Specialist catalogues published by the British Museum feature prominently. In volumes 1-3, which describe German printed incunabula, there are 14 citations to Campbell Dodgson's *Catalogue of Early German and Flemish Woodcuts in the British Museum*.[33] In volume 6, the second of four volumes describing Italian printed incunabula, A. M. Hind's *Catalogue of Early Italian Engravings ... in the British Museum (1910)* is cited twice. In volume 10 there are further two references to A. M. Hind's *An Introduction to a History of Woodcut, vol. 2 (1935)*. Seven references to the *Catalogue of the Fifty Manuscripts and Printed Books bequeathed to the British Museum by Alfred H. Huth (1912)* appear across three volumes, varying in appearance from full references to a simple 'see Catalogue of the Huth' or 'Huth Catalogue'. In addition, in volume 9, which describes the incunabula printed in the Netherlands, there are 34 references to W. M. Conway's *The Woodcutters of the Netherlands*. More broadly, and going beyond corpus level word frequency lists, our analysis suggests a trend for including more scholarly references as part of the descriptions in the later volumes, and in particular in the last published, volume 11 (2007), that describes incunabula printed in England and was intended as a scholarly publication. That does not mean earlier volumes were less scholarly but rather is explained in part by the evolving structures of the *BMC* and the prominence of bibliographic citations with other sections of the printed catalogue, such as introductory materials.

Comparable approaches support analysis of historical language relating to descriptions of bookbindings. In *BMC* many bindings are described using specialist terms. For example, the word 'morocco' (584 occurrences) refers to goatskin used in bindings originally from Morocco. But compound words referring to the animal skin used in the binding dominate: 'pigskin' (450 occurrences), 'sheepskin' (64 occurrences), deerskin (9 occurrences), doeskin (2 occurrences),

[32] Due to the OCR inaccuracies and named entity variants more work is needed to standardise the names before they can be processed with NER tools. This work would bring many benefits as part of metadata enrichment for online catalogues in terms of matching the entities to authority files and creating linked datasets.

[33] Eg. "The attribution of the cuts to Dürer is probably incorrect. See Dodgson, Catalogue of Early German and Flemish Woodcuts in the British Museum, vol. I, p. 263." (IB. 7558)

'goatskin' (1 occurrence), and 'fish skin' (1 occurrence). These are complimented by a smaller number that give simply an animal name: 'calf' (251 occurrences), 'sheep' (27 occurrences), 'goat' (3 occurrences). These high-level observations suggest possible disjunctures between historical cataloguing practice and contemporary audiences less familiar with specialist cataloguing language, and this line of enquiry is deepened by more fine-grained linguistic analysis. For example, Figure 4 shows catalogue descriptions containing the string 'morocco' plotted by entries within *BMC* volumes (with the first catalogue entries in the volume at the leftmost extent of the plot and the last entries at the rightmost extent of the plot) and by individual *BMC* volume (with one volume comprising a single row of the plot). This shows the dispersion of this specialist vocabulary across the corpus, with the highest number of 'morocco' occurrences found in volume 8 (150 occurrences), volume 9 (119 occurrences) and volume 5 (102 occurrences), and very few occurrences of in volumes 1-4, in which the binding type 'pigskin' is dominant. Collocates of 'morocco' indicate further layers of specialist language relating to colour attribution. While the most frequent descriptors for 'morocco' are 'crushed red' (72 occurrences), 'crushed blue' (33 occurrences) and 'crushed brown' (25 occurrences), we also we find less usual and more impressionistic chromatic vocabularies: 'chocolate', 'claret', 'crimson', 'maroon', and 'slate'.[34] The word 'morocco' appears in nearly twice as many trigrams as 'pigskin' (148 occurrences versus 78 occurrences), which is explained by the fact that pigskin was typically used in bookbinding in undyed form. And within these trigrams, we observe subtle variants that indicate choices and preferences of individual cataloguers: 'deep red' and 'dark red', 'blue crushed' and 'crushed blue', 'crushed green' and 'crushed olive'.

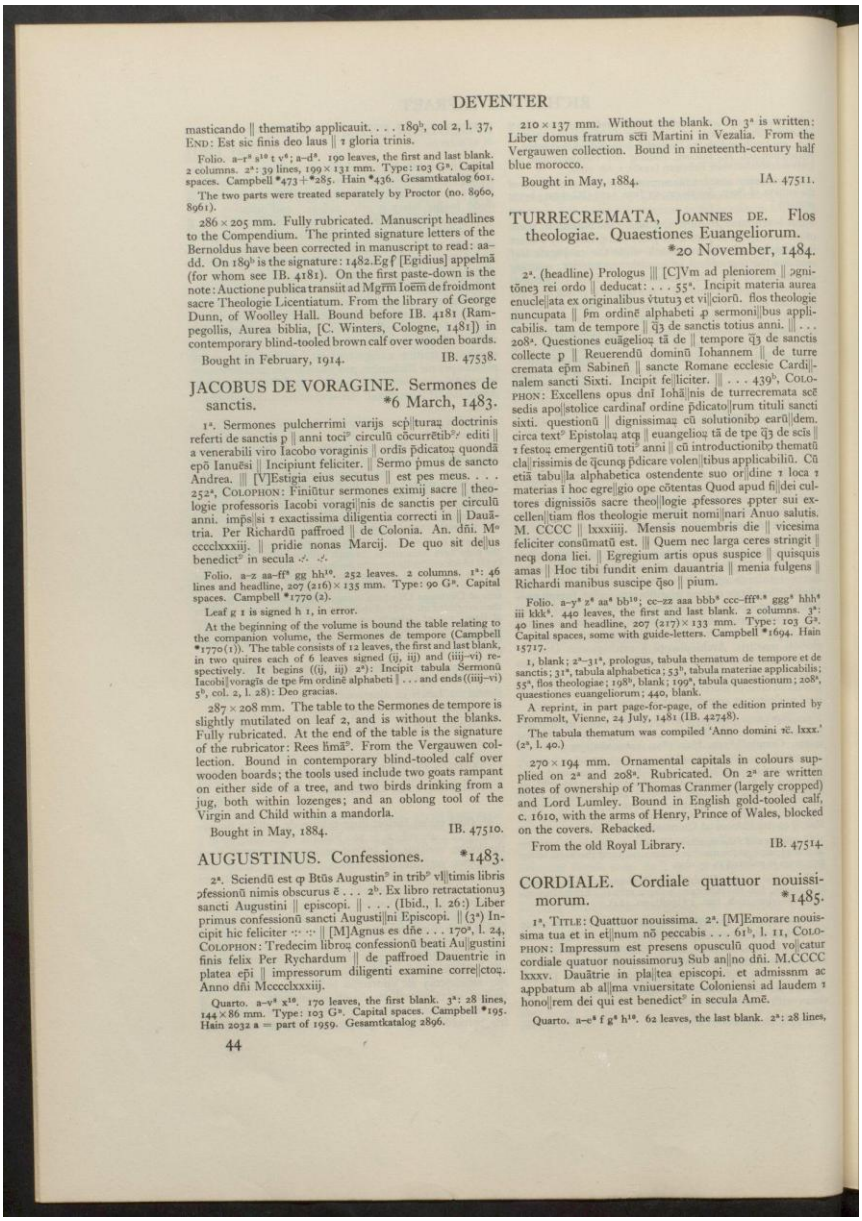
Indeed, iterative use of corpus linguistic approaches is particularly amenable to detecting cataloguers' personal preferences, and the broad continuity of cataloguing practice, across the *BMC* corpus. For example, when describing the same binding technique some cataloguers, especially those who worked on volumes 9 and 10 of the *BMC*, chose 'blind tooled' (59 occurrences) as opposed to 'blind stamped' (18 occurrences). And in the same two volumes we

[34] The following colour descriptors are used by the cataloguer of vol. 8: claret (2), slate (1), chocolate (1).

find 16 occurrences of the string 'blind impression of type… is distinguishable'.[35] Later cataloguers also introduce new descriptive vocabularies into the corpus: 'marbled (36 occurrences) leather', 'panelled (1 occurrences) calf', and the frequent use of nouns such as 'bindery' (121 occurrences, used once before volume 9) almost always in reference to 'the British Museum bindery'. Binding descriptions are also more-detailed and consistent in these later volumes. For example, the lines 'Bound, in contemporary Netherlandish blind-tooled brown calf over wooden boards. The tools used include a square with an oriental linear design, a rectangle with a lion passant in front of a tree, a fleur-de-lis within a lozenge, and an Agnus Dei and a rose within circles' both begin with a regular ten-word n-gram and contain a new regularised vocabulary - 'lozenge' - for a diamond shape.[36] Finally, this form of analysis can suggest the different uses of punctuation marks between volumes. For example, in the context of bindings the string 'original(?)' occurs over 150 times, the largest number of which occur in volume 8. This flags the need for possible revisions in future uses of catalogue data, as well as the complexity of some collection items described in the *BMC*. More significantly it opens up lines of enquiry around changing expertise between volumes of the *BMC* and the circumstances of cataloguing production.

[35] Altogether there are 118 occurrences of the technical term 'blind impression(s)' in the *BMC* corpus.

[36] 7 out of the 11 occurrences of 'lozenge' are in vol.9 and in the context of binding descriptions.



DEVENTER

masticando ‖ thematibꝫ applicauit. . . . 189ᵇ, col 2, l. 37, END: Est sic finis deo laus ‖ ꝛ gloria trinis.

Folio. a–rˢ sˡᵒ ꝛ vˢ; a–dˢ. 190 leaves, the first and last blank. 2 columns. 2ᵃ: 39 lines, 199 × 131 mm. Type: 103 G*. Capital spaces. Campbell *473 + *285. Hain *436. Gesamtkatalog 601.

The two parts were treated separately by Proctor (no. 8960, 8961).

286 × 205 mm. Fully rubricated. Manuscript headlines to the Compendium. The printed signature letters of the Bernoldus have been corrected in manuscript to read: aa–dd. On 189ᵇ is the signature: 1482.Eg f [Egidius] appelmā (for whom see IB. 4181). On the first paste-down is the note: Auctione publica transiit ad Mgrm̄ Ioēm de froidmont sacre Theologie Licentiatum. From the library of George Dunn, of Woolley Hall. Bound before IB. 4181 (Rampegollis, Aurea biblia, [C. Winters, Cologne, 1481]) in contemporary blind-tooled brown calf over wooden boards.

Bought in February, 1914. IB. 47538.

JACOBUS DE VORAGINE. Sermones de sanctis. *6 March, 1483.

1ᵃ. Sermones pulcherrimi varijs scp̄turaꝝ doctrinis referti de sanctis p ‖ anni tociꝰ circulū cōcurrētibꝰ editi ‖ a venerabili viro Iacobo voraginis ‖ ordīs p̄dicatoꝝ quondā epō Ianuēsi ‖ Incipiunt feliciter. ‖ Sermo p̄mus de sancto Andrea. ‖ [V]Estigia eius secutus ‖ est pes meus. . . . 252ᵃ, COLOPHON: Finiūtur sermones eximij sacre ‖ theologie professoris Iacobi voragi‖nis de sanctis per circulū anni. imp̄s‖si ꝛ exactissima diligentia correcti in ‖ Dauātria. Per Richardū paffroed ‖ de Colonia. An. dñi. Mᵒ cccclxxxiij. ‖ pridie nonas Marcij. De quo sit de‖us benedictꝰ in secula ·ſ· ·ſ·

Folio. a–z aa–ffˢ gg hhˡᵒ. 252 leaves. 2 columns. 1ᵃ: 46 lines and headline, 207 (216) × 135 mm. Type: 90 G*. Capital spaces. Campbell *1770 (2).

Leaf g 1 is signed h 1, in error.

At the beginning of the volume is bound the table relating to the companion volume, the Sermones de tempore (Campbell *1770(1)). The table consists of 12 leaves, the first and last blank, in two quires each of 6 leaves signed (ij, iij) and (iii–vi) respectively. It begins ((ij, iij) 2ᵃ): Incipit tabula Sermonū Iacobi‖voragis de tpe p̄m ordinē alphabeti ‖ . . . and ends ((iiij–vi) 5ᵇ, col. 2, l. 28): Deo gracias.

287 × 208 mm. The table to the Sermones de tempore is slightly mutilated on leaf 2, and is without the blanks. Fully rubricated. At the end of the table is the signature of the rubricator: Rees h̄mā. From the Vergauwen collection. Bound in contemporary blind-tooled calf over wooden boards; the tools used include two goats rampant on either side of a tree, and two birds drinking from a jug, both within lozenges; and an oblong tool of the Virgin and Child within a mandorla.

Bought in May, 1884. IB. 47510.

AUGUSTINUS. Confessiones. *1483.

2ᵃ. Sciendū est ꝙ Btūs Augustinꝰ in tribꝰ vl‖timis libris p̄fessionū nimis obscurus ē . . . 2ᵇ. Ex libro retractationuꝫ sancti Augustini ‖ episcopi. ‖ . . . (Ibid., l. 26:) Liber primus confessionū sancti Augusti‖ni Episcopi. ‖ (3ᵃ) Incipit hic feliciter ·:· ·:· ‖ [M]Agnus es dñe . . . 170ᵃ, l. 24, COLOPHON: Tredecim libroꝫ confessionū beati Au‖gustini finis felix Per Rychardum ‖ de paffroed Dauentrie in platea ep̄i ‖ impressorum diligenti examine corre‖ctoꝝ. Anno dñi Mcccclxxxiij.

Quarto. a–vˢ xˡᵒ. 170 leaves, the first blank. 3ᵃ: 28 lines, 144 × 86 mm. Type: 103 G*. Capital spaces. Campbell *195. Hain 2032 a = part of 1959. Gesamtkatalog 2896.

210 × 137 mm. Without the blank. On 3ᵃ is written: Liber domus fratrum sc̄ti Martini in Vezalia. From the Vergauwen collection. Bound in nineteenth-century half blue morocco.

Bought in May, 1884. IA. 47511.

TURRECREMATA, JOANNES DE. Flos theologiae. Quaestiones Euangeliorum. *20 November, 1484.

2ᵃ. (headline) Prologus ||| [C]Vm ad pleniorem ‖ ɔgnitōneꝫ rei ordo ‖ deducat: . . . 55ᵃ. Incipit materia aurea enucle‖ata ex originalibus ꝯtutuꝫ et vi‖ciorū. flos theologie nuncupata ‖ p̄m ordinē alphabeti p̄ sermoni‖bus applicabilis. tam de tempore ‖ q̄ de sanctis totius anni. ‖ . . . 208ᵃ. Questiones euāgelioꝝ tā de ‖ tempore q̄ de sanctis collecte p ‖ Reuerendū dominū Iohannem ‖ de turre cremata ep̄m Sabinēn ‖ sancte Romane ecclesie Cardi‖nalem sancti Sixti. Incipit fe‖liciter. ‖ . . . 439ᵇ, COLOPHON: Excellens opus dnī Iohā‖nis de turrecremata scē sedis apo‖stolice cardinal ordine p̄dicato‖rum tituli sancti sixti. questionū ‖ dignissimaꝝ cū solutionibꝫ earū‖dem. circa textꝰ Epistolaꝝ atq ‖ euangelioꝝ tā de tpe q̄ de scis ‖ ꝛ festoꝝ emergentiū totiꝰ anni ‖ cū introductionibꝫ thematū cla‖rissimis de q̄cunq p̄dicare volen‖tibus applicabiliū. Cū etiā tabu‖la alphabetica ostendente suo or‖dine ꝛ loca ꝛ materias ī hoc egre‖gio ope cōtentas Quod apud fi‖dei cultores dignissiōs sacre theo‖logie p̄fessores p̄pter sui excellen‖tiam flos theologie meruit nomi‖nari Anuo salutis. M. CCCC ‖ lxxxiiij. Mensis nouembris die ‖ vicesima feliciter consūmatū est. ‖ Quem nec larga ceres stringit ‖ neq dona liei. ‖ Egregium artis opus suspice ‖ quisquis amas ‖ Hoc tibi fundit enim dauantria ‖ menia fulgens ‖ Richardi manibus suscipe q̄so ‖ pium.

Folio. a–yˢ zˢ aaˢ bbˡᵒ; cc–zz aaa bbbˢ ccc–fffˢ·ˢ gggˢ hhhˢ iii kkkˢ. 440 leaves, the first and last blank. 2 columns. 3ᵃ: 40 lines and headline, 207 (217) × 133 mm. Type: 103 G*. Capital spaces, some with guide-letters. Campbell *1694. Hain 15717.

1, blank; 2ᵃ–31ᵃ, prologus, tabula thematum de tempore et de sanctis; 31ᵃ, tabula alphabetica; 53ᵇ, tabula materiae applicabilis; 55ᵃ, flos theologiae; 198ᵇ, blank; 199ᵃ, tabula questionum; 208ᵃ, quaestiones euangeliorum; 440, blank.

A reprint, in part page-for-page, of the edition printed by Frommolt, Vienne, 24 July, 1481 (IB. 4274ß).

The tabula thematum was compiled ‘Anno domini m̄cc. lxxx.’ (2ᵃ, l. 40.)

270 × 194 mm. Ornamental capitals in colours supplied on 2ᵃ and 208ᵃ. Rubricated. On 2ᵃ are written notes of ownership of Thomas Cranmer (largely cropped) and Lord Lumley. Bound in English gold-tooled calf, c. 1610, with the arms of Henry, Prince of Wales, blocked on the covers. Rebacked.

From the old Royal Library. IB. 47514.

CORDIALE. Cordiale quattuor nouissimorum. *1485.

1ᵃ, TITLE: Quattuor nouissima. 2ᵃ. [M]Emorare nouissima tua et in et‖num nō peccabis . . . 61ᵇ, l. 11, COLOPHON: Impressum est presens opusculū quod vo‖catur cordiale quatuor nouissimoruꝫ Sub an‖no dñi. M.CCCC lxxxv. Dauātrie in pla‖tea episcopi. et admissnm ac a̅ppbatum ab al‖ma vniuersitate Coloniensi ad laudem ꝛ hono‖rem dei qui est benedictꝰ in secula Amē.

Quarto. a–eˢ f gˢ hˡᵒ. 62 leaves, the last blank. 2ᵃ: 28 lines,

44

Figure 4: *Catalogue of books printed in the 15th century now at the British Museum* Part 9 (1962) (BL shelfmark J/2704.aa.30/9) p.44 showing a lengthy description and detail on binding. Note the reference to ‘lozenge’-shape in IB.47510. Image: © British Library Board, CC-BY 4.0

It is established knowledge that the entire *BMC* was produced over 100 years by several members of staff in the Department of Printed Books at the British Museum (and later British Library)[37], this in spite of major interruptions during the two World Wars (Harris, 1998, 394, 494, 560, 578-9.). The volumes on which our corpus is based were produced between 1906 and 1971 as the result of the collaborative effort of several contributors: A.W. Pollard, A. Esdaile, V. Scholderer, H. Thomas, L.A. Sheppard, G.D. Painter. Of these, Victor Scholderer, the Department's incunabulist until 1945, has been acknowledged as the "one continuous worker", who gradually took over as the catalogue editor,[38] became the principal author of volumes 5 to 8, and contributed to volumes 9, 10 and 12 after his retirement.[39] Especially notable is the endurance of his work during the Second World War, during which time catalogued the French incunabula (volume 8) whilst based at the National Library of Wales to which some of the collection had been evacuated (Harris, 1998, 560, 578). Thus, as well as providing some continuity with the cataloguing standards of the Department, we would expect to detect some of Scholderer's authoritative voice in the descriptions.[40]

[37] Intro to vol.11. p.2: "If the previous twelve volumes in the series, the work of only five editors, [A.W. Pollard, V. Scholderer, G. D. Painter, D. E. Rhodes, A. K. Offenberg] show a steady but often subtle adaptation to the expectations of their time, the present volume, covering the collections of books printed in England, shows more radical departures from past practice by considerably expanding the information provided for each item and by giving more attention to what can be deduced about their production in the printing house".

[38] Pollard writes in the preface to vol. 5 *"*Mr. Scholderer has definitely taken my place as Editor, and it is with a great satisfaction to me that the completion of the catalogue is in such eminently competent hands*"*. In the preface to vols. 7-9 there is a reference to "a Mrs. L. G. Clark who has rendered an invaluable service to the catalogue by undertaking the laborious task of compiling the indexes and helping to see them through the press."

[39] Julius Victor Scholderer (1880-1971), Assistant in Printed Books, 1904-20; Assistant Keeper, 1921-9; Deputy Keeper, 1930-45. The prefaces to volumes 1-8 acknowledge Scholderer's contribution: He was one of two assistants for vol 1 (1908), was responsible for more than half of the descriptions in vol. 2 (1912), collaborated on the text, introduction and map in vol. 3 (1913), collaborated on the text and Intro in vol.4 (1916), is "the only continuous worker" and takes over as editor to vol. 5 (1924), he is the main cataloguer of vol.6 (1930) and "at various times received most valuable assistance from L.A Sheppard", was editor for vol. 7 (1935) with "the descriptions and introductions entirely by him", was the sole cataloguer of vol. 8 except for the sections on Rouen in the text and the introduction which were by L. A. Sheppard.

[40] We are currently processing the data from vol.11 and 13 to enable further analysis of how the incunabula cataloguing practice has evolved.

For example, corpus linguistic analysis of qualifying adverbs in the *BMC* corpus adverbs gives us a proxy for the levels of assertiveness - and uncertainty - in catalogue descriptions. 260 ‘-ly’ type adverbs are present in the corpus with a total frequency of 7823 occurrences. 87 percent of the total represent the 30 most frequent types, and a small number appear in quotations, and therefore represents descriptive decisions made by third-parties rather than by *BMC* cataloguers.

only	1179	doubtfully	18	characteristically	6	singly	3	daily	sightly
probably	606	extremely	18	fairly	6	strictly	3	delicately	slantingly
mostly	575	practically	18	precisely	6	constantly	3	deuoutly	sparsely
apparently	451	shortly	18	carelessly	5	additionally	2	diligently	steeply
presumably	385	directly	17	hardly	5	coarsely	2	disproportionately	strikingly
fully	337	simultaneously	17	simply	5	comparatively	2	distantly	suddenly
alternately	282	consecutively	16	uniformly	5	confusedly	2	evenly	summarily
slightly	268	crookedly	15	commonly	5	conjecturally	2	financially	symmetrically
formerly	256	regularly	15	anonymously	4	conjointly	2	finely	thickly
previously	207	widely	15	appreciably	4	conspicuously	2	foolishly	thoroughly
respectively	207	actually	14	briefly	4	copiously	2	formally	typographically
partly	198	notably	14	conceivably	4	curiously	2	fraudulently	ungrammatically
closely	137	undoubtedly	14	deadly	4	deeply	2	friendly	universally
nearly	131	properly	13	definitely	4	distinctly	2	heavenly	unmetrically
evidently	119	accordingly	13	deliberately	4	doubly	2	hesitatingly	unnecessarily
possibly	114	approximately	12	exceedingly	4	duly	2	ignorantly	unthinkingly
originally	114	elaborately	12	hastily	4	effectively	2	illegibly	unwarrantably
correctly	90	imperfectly	14	identically	4	eventually	2	improbably	urgently
immediately	89	irregularly	12	independently	4	faithfully	2	inadvertently	virulently
generally	80	completely	12	literally	4	faultily	2	incidentally	satisfactorily
exactly	78	faintly	11	naturally	4	inadequately	2	indifferently	correspondingly
occasionally	74	partially	10	profusely	4	infrequently	2	indirectly	
wrongly	68	rarely	10	purely	4	insufficiently	2	indiscriminately	
usually	65	equally	9	recently	4	intentionally	2	intrinsically	
separately	63	greatly	9	strongly	4	lately	2	invariably	
similarly	60	vertically	9	diagonally	3	necessarily	2	ironically	
entirely	57	exclusively	9	explicitly	3	newly	2		
subsequentl y	51	clumsily	8	expressly	3	plainly	2	jointly	
considerably	47	noticeably	8	finally	3	reasonably	2	laterally	
	46	rightly	8	gradually	3	rudely	2	logically	

roughly	43	substantially	8	highly	3	satisfactorily	2	luxuriantly	
especially	42	unusually	8	horizontally	3	skilfully	2	nominally	
accidentally	38	lightly	8	inaccurately	3			notarially	
certainly	35	continuously	8	neatly	3	successively	2	obviously	
heavily	35	crudely	7	quarterly	3	sufficiently	2	officially	
frequently	32	kindly	7	repeatedly	3	surely	2	ordinarily	
differently	32	particularly	7	seemingly	3	tentatively	2	perfectly	
likely	27	readily	7	singly	3	totally	2	pictorially	
normally	26	sparingly	7	strictly	3	transversely	2	plausibly	
clearly	26	wholly	7	constantly	3	unevenly	2	pointlessly	
merely	24	chiefly	6	additionally	2	unmeaningly	2	proportionally	
reply	24	easily	6	coarsely	2	virtually	2	proportionately	
scarcely	23	inadvertently	6	comparatively	2	temporarily	2	publicly	
erroneously	23	loosely	6	confusedly	2	accidentally	2	purposely	
ultimately	22	mistakenly	6	conjecturally	2	abnormally	1	quickly	
carefully	21	sharply	6	conjointly	2			richly	
freely	21	unlikely	6	conspicuously	2	alternatively	1	safely	
largely	21	variously	6	copiously	2	badly	1	schematically	
really	20	alphabetically	6	curiously	2	concurrently	1	secretly	
mainly	20			deeply	2	conscientiously	1	severally	
incorrectly									

Figure 5: Frequencies of adverbs ending '-ly' in the *BMC* corpus (no case setting selected).

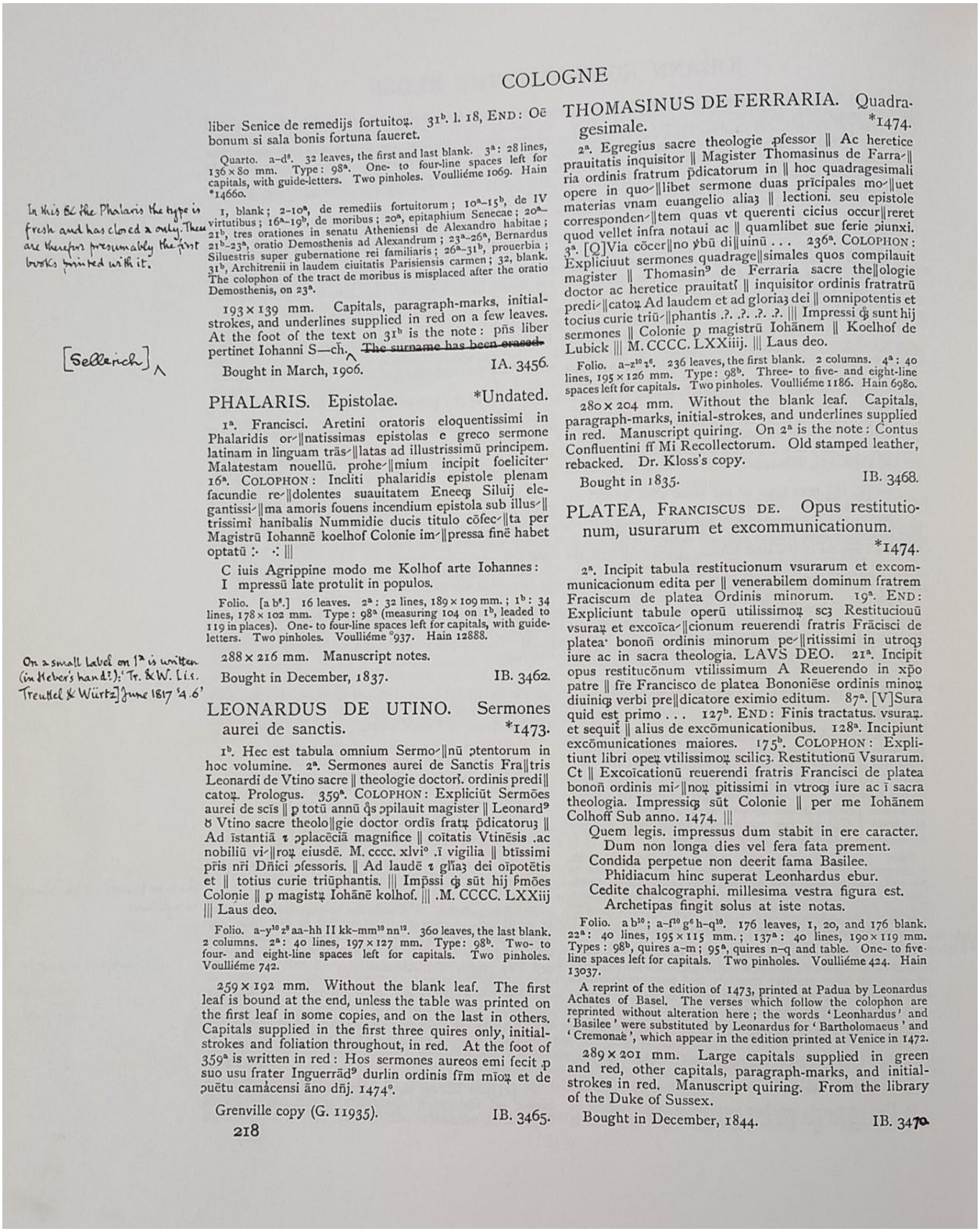
These adverbs occur most frequently in volumes 5 to 8, the volumes which were catalogued primarily by Scholderer. Moreover, signs of his curatorial voice emerge in the use of a wide range of adverb forms. For example, where 'erased'[41] is accompanied by an adverb *BMC* corpus, 'partly' and 'partially' are routinely used. However, in volumes 5 to 8 we observe more nuanced choices of language including 'clumsily'[42], 'thoroughly'[43], 'heavily', and 'carefully'. Indeed, of the 35 occurrences of 'heavily' in the *BMC* corpus, 24 are found in volume 8, the volume principally authored by Scholderer. Other vocabulary characteristic of his voice include 'appreciably' (3 occurrences), 'elaborately' (1 occurrence), 'intrinsically' (1 occurrence), 'luxuriantly' (1 occurrence), 'strikingly' (1 occurrence), 'conscientiously' (1 occurrence)[44], and 'virulently' (1 occurrence). And other parts of speech and syntax in the Scholderer-led volumes display a modal tendency not characteristic of the other volumes, such as '*may* be by another printer' or '*may or may not* indicate..'), as well as formulations nuanced by an accompanying adverb such as 'may have been *deliberately* omitted' (volume 8, IB. 41225), 'may have been *inadvertently* …' (volume 8, IC. 41173), '*apparently* may have been printed', 'very *possibly* may here have been taken over by inadvertence' (IB. 40221.), and 'might equally well; may very well be/cannot well be earlier'.

[41] There are some patterns in the use of synonyms, for example 'erased' (142) is used more frequently than 'obliterated' (44), with 'erased' being used mostly as a verb (something 'is/are/has/have been erased') whilst 'obliterated' used almost always as an adjective (a partly obliterated inscription')

[42] '...In some cases wrong head-lines were supplied, then *clumsily* erased and the correct head-lines substituted. A gothic d is regularly used on leaves 11-24 and is not found elsewhere…"(IC. 19527.)

[43] "... but the significant words of the incipit on a 3b have been too *thoroughly* erased to be decipherable. Doubtfully assigned to the press of Dupré by Proctor (no. 8046)... (IA. 41096.)

[44] "...Thus the repetitions are of the most harmless kind, and ... this must take rank with the most *conscientiously* illustrated of fifteenth century books (Pierpont Morgan Catalogue (1907), no. 612)..." (IB. 41903.)



liber Senice de remedijs fortuitoꝝ. 31b. l. 18, END: Oē bonum si sala bonis fortuna faueret.

Quarto. a–d8. 32 leaves, the first and last blank. 3a: 28 lines, 136×80 mm. Type: 98a. One- to four-line spaces left for capitals, with guide-letters. Two pinholes. Voulliéme 1069. Hain *14660.

In this & the Phalaris the type is fresh and has closed a only. They are therefor presumably the first books printed with it.

1, blank; 2–10a, de remediis fortuitorum; 10a–15b, de IV virtutibus; 16a–19b, de moribus; 20a, epitaphium Senecae; 20a–21b, tres orationes in senatu Atheniensi de Alexandro habitae; 21b–23a, oratio Demosthenis ad Alexandrum; 23a–26a, Bernardus Siluestris super gubernatione rei familiaris; 26a–31b, prouerbia; 31b, Architrenii in laudem ciuitatis Parisiensis carmen; 32, blank. The colophon of the tract de moribus is misplaced after the oratio Demosthenis, on 23a.

193×139 mm. Capitals, paragraph-marks, initial-strokes, and underlines supplied in red on a few leaves. At the foot of the text on 31b is the note: pñs liber pertinet Iohanni S—ch. ~~The surname has been erased.~~

[Selderich]

Bought in March, 1906. IA. 3456.

PHALARIS. Epistolae. *Undated.

1a. Francisci. Aretini oratoris eloquentissimi in Phalaridis or-‖natissimas epistolas e greco sermone latinam in linguam trās-‖latas ad illustrissimū principem. Malatestam nouellū. prohe-‖mium incipit foeliciter· 16a. COLOPHON: Incliti phalaridis epistole plenam facundie re-‖dolentes suauitatem Eneeꝝ Siluij elegantissi-‖ma amoris fouens incendium epistola sub illus-‖trissimi hanibalis Nummidie ducis titulo cōfec-‖ta per Magistrū Iohannē koelhof Colonie im-‖pressa finē habet optatū ∴ ∴ |||

C iuis Agrippine modo me Kolhof arte Iohannes:
I mpressū late protulit in populos.

Folio. [a b8.] 16 leaves. 2a: 32 lines, 189×109 mm.; 1b: 34 lines, 178×102 mm. Type: 98a (measuring 104 on 1b, leaded to 119 in places). One- to four-line spaces left for capitals, with guide-letters. Two pinholes. Voulliéme °937. Hain 12888.

On a small label on 1a is written (in Heber's hand?): 'Tr. & W. Lis. Treuttel & Würtz] June 1817 £4.6'

288×216 mm. Manuscript notes.

Bought in December, 1837. IB. 3462.

LEONARDUS DE UTINO. Sermones aurei de sanctis. *1473.

1b. Hec est tabula omnium Sermo-‖nū ꝓtentorum in hoc volumine. 2a. Sermones aurei de Sanctis Fra‖tris Leonardi de Vtino sacre ‖ theologie doctorī. ordinis predi‖catoꝝ. Prologus. 359a. COLOPHON: Expliciūt Sermōes aurei de scīs ‖ p totū annū q̄s ꝓpilauit magister ‖ Leonard9 ȣ Vtino sacre theolo‖gie doctor ordīs fratꝝ p̄dicatoruꝫ ‖ Ad īstantiā ꝛ ꝓplacēciā magnifice ‖ coītatis Vtinēsis .ac nobiliū vi-‖roꝝ eiusdē. M. cccc. xlvi° .ī vigilia ‖ btissimi pr̄is nr̄i Dn̄ici ꝓfessoris. ‖ Ad laudē ꝛ glr̄iaꝫ dei oīpotētis et ‖ totius curie triūphantis. ||| Impssi q̄ sūt hij p̄mōes Colonie ‖ p magistꝝ Iohānē kolhof. ||| .M. CCCC. LXXiij ||| Laus deo.

Folio. a–y10 z8 aa–hh II kk–mm10 nn12. 360 leaves, the last blank. 2 columns. 2a: 40 lines, 197×127 mm. Type: 98b. Two- to four- and eight-line spaces left for capitals. Two pinholes. Voulliéme 742.

259×192 mm. Without the blank leaf. The first leaf is bound at the end, unless the table was printed on the first leaf in some copies, and on the last in others. Capitals supplied in the first three quires only, initial-strokes and foliation throughout, in red. At the foot of 359a is written in red: Hos sermones aureos emi fecit ꝓ suo usu frater Inguerrād9 durlin ordinis fr̄m mīoꝝ et de ꝓuētu camȧcensi āno dn̄j. 1474°.

Grenville copy (G. 11935). IB. 3465.

THOMASINUS DE FERRARIA. Quadragesimale. *1474.

2a. Egregius sacre theologie ꝓfessor ‖ Ac heretice prauitatis inquisitor ‖ Magister Thomasinus de Farra-‖ria ordinis fratrum p̄dicatorum in ‖ hoc quadragesimali opere in quo-‖libet sermone duas prīcipales mo-‖uet materias vnam euangelio aliaꝫ ‖ lectioni. seu epistole corresponden-‖tem quas vt querenti cicius occur‖reret quod vellet infra notaui ac ‖ quamlibet sue ferie ꝓiunxi. 3a. [Q]Via cōcer‖no ȳbū di‖uinū . . . 236a. COLOPHON: Expliciuut sermones quadrage‖simales quos compilauit magister ‖ Thomasin9 de Ferraria sacre the‖ologie doctor ac heretice prauitatī ‖ inquisitor ordinis fratratrū predi-‖catoꝝ Ad laudem et ad gloriaꝫ dei ‖ omnipotentis et tocius curie triū-‖phantis .?. .?. .?. .?. ||| Impressi q̄ sunt hij sermones ‖ Colonie p magistrū Iohānem ‖ Koelhof de Lubick ||| M. CCCC. LXXiiij. ||| Laus deo.

Folio. a–z10 2a6. 236 leaves, the first blank. 2 columns. 4a: 40 lines, 195×126 mm. Type: 98b. Three- to five- and eight-line spaces left for capitals. Two pinholes. Voulliéme 1186. Hain 6980.

280×204 mm. Without the blank leaf. Capitals, paragraph-marks, initial-strokes, and underlines supplied in red. Manuscript quiring. On 2a is the note: Contus Confluentini ff Mi Recollectorum. Old stamped leather, rebacked. Dr. Kloss's copy.

Bought in 1835. IB. 3468.

PLATEA, FRANCISCUS DE. Opus restitutionum, usurarum et excommunicationum. *1474.

2a. Incipit tabula restitucionum vsurarum et excommunicacionum edita per ‖ venerabilem dominum fratrem Fraciscum de platea Ordinis minorum. 19a. END: Expliciunt tabule operū utilissimoꝝ scꝫ Restitucioūū vsuraꝝ· et excoīca-‖cionum reuerendi fratris Frācisci de platea· bonoñ ordinis minorum pe-‖ritissimi in utroqꝫ iure ac in sacra theologia. LAVS DEO. 21a. Incipit opus restitucōnum vtilissimum A Reuerendo in xp̄o patre ‖ fre Francisco de platea Bononiēse ordinis minoꝝ diuiniqꝫ verbi pre‖dicatore eximio editum. 87a. [V]Sura quid est primo . . . 127b. END: Finis tractatus. vsuraꝝ. et sequiť ‖ alius de excōmunicationibus. 128a. Incipiunt excōmunicationes maiores. 175b. COLOPHON: Explitiunt libri opeꝝ vtilissimoꝝ scilicꝫ. Restitutionū Vsurarum. Ct ‖ Excoīcationū reuerendi fratris Francisci de platea bonoñ ordinis mi-‖noꝝ pitissimi in vtroꝫ iure ac ī sacra theologia. Impressiq̄ sūt Colonie ‖ per me Iohānem Colhoff Sub anno. 1474. |||

Quem legis. impressus dum stabit in ere caracter.
Dum non longa dies vel fera fata prement.
Condida perpetue non deerit fama Basilee.
Phidiacum hinc superat Leonhardus ebur.
Cedite chalcographi. millesima vestra figura est.
Archetipas fingit solus at iste notas.

Folio. a b10; a–f10 g6 h–q10. 176 leaves, 1, 20, and 176 blank. 22a: 40 lines, 195×115 mm.; 137a: 40 lines, 190×119 mm. Types: 98b, quires a–m; 95a, quires n–q and table. One- to five-line spaces left for capitals. Two pinholes. Voulliéme 424. Hain 13037.

A reprint of the edition of 1473, printed at Padua by Leonardus Achates of Basel. The verses which follow the colophon are reprinted without alteration here; the words 'Leonhardus' and 'Basilee' were substituted by Leonardus for 'Bartholomaeus' and 'Cremonae', which appear in the edition printed at Venice in 1472.

289×201 mm. Large capitals supplied in green and red, other capitals, paragraph-marks, and initial-strokes in red. Manuscript quiring. From the library of the Duke of Sussex.

Bought in December, 1844. IB. 3470

Figure 5: *BMC* 1 (1963) (BL shelfmark RAR093.016 ENG) p. 218 showing neat handwritten annotations by Scholderer and his distinctive use of fresh as a descriptor for 'type'. Image: © British Library Board, CC-BY 4.0

Those few examples indicate some playful deliberation in the language of the curatorial observation. As an incunabulist who was confident in his specialism, Scholderer's style of cataloguing appears to be both confident and more informal, with nuances of language that suggests his enjoyment in creating the descriptions. This impression of his skill and linguistic flare gained from corpus linguistic analysis is supported by the testimonies of his colleagues and friends who spoke of his "felicitous style of writing, not without its flashes of wit and humour" (Rhodes 1970, 12).[45]

4. The value of this approach

The automation of processes to prepare catalogues as data is not an easy and straightforward task and involves much experimentation and iteration. In our case this involved the training of structure models, the development of algorithms for entries detection, and the deployment of language identification, all of which required different skills. The benefit of this work is that it has enabled the use of AntConc to query language data for 9,000 records, to navigate through the data via word frequency lists, and to examine that data in context. Our computational analysis of the incunabula catalogue data has enriched our understanding of the curator/cataloguer's observations, i.e. where their interest or expertise lies, the use of terminology and detection of cataloguing styles. It has also revealed some information patterns worthy of further examination to improve discovery and support research with the collection. Finally, managing the descriptions in a structured way has

[45] *Ibid.* M. E. Kronenberg, We Became Friends, p.36 - who left her correspondence with Victor Scholderer to the Royal Library at the Hague. "After reading them he [a librarian] will realise that the old incunabulist, till then as remote to him as Panzer and Hain are to us, was not at all a dry and dull bibliomaniac but a lively, warm-hearted man, many-sided, interested in literature, a poet even, and with artistic gifts …"

brought us closer to integrating them with the online catalogue, as well as plans for further enrichment.[46]

Bringing together Library (and non-Library) staff - the collection expert, digital curator, data scientist, metadata expert, digital humanist - was an essential and valuable part of the process for working with catalogues as data. Such collaboration in the design and evaluation of workflows and tools to work with catalogues as data is transformative and integrates digital scholarship within the daily curatorial and collection management practice. It also enables staff to learn new skills, to stay research active (a key aim of the Fellowship scheme), and to share our learning with others.

The impact of the project goes beyond the immediate objectives of enrichment of the Library's online catalogue. The open data we have shared will enable its reuse for training models (LLMs), enrichment of other metadata, and new digital research. It has created new outreach opportunities to showcase our approach and data with the international community of heritage professionals and digital humanists and support research and practice with catalogues as data.[47]

Our work with catalogues as data supports and amplifies much of what practitioners know about cataloguing practice: it is complex, it is shaped by institutional history, it is entangled with how catalogues are conceived, organised, and published, and it is a product of people and the circumstances they work in. In the case of the *BMC*, the production of its volumes were influenced by major historical events and by interfaces with other cataloguing projects such as the General Catalogue (GK) and various Short-title Catalogues. Whilst decisions about the methods of cataloguing are documented in the introduction to each volume of the *BMC*, as well as in related scholarly works, everyday decisions made in the process of cataloguing are rarely captured in print and, where retained, form part of institutional knowledge.[48] Each volume of *BMC* also can only

[46] Comparative analysis with other catalogue data describing books printed before 1500 and indeed other catalogue data would develop further this research but this was not feasible within the timeframe of this project.

[47] https://blogs.bl.uk/digital-scholarship/2024/07/dhbn-2024-digital-humanities-in-the-nordic-and-baltic-countries-conference-report.html

[48] The Introductions to *BMC* vols. 1-3 mention descriptions of duplicates have been left out of the catalogue due to time pressure. In a conversation with the former curator of incunabula John Goldfinch, he acknowledged that the difference in the level of detail in the descriptions could be

describe the collection particular point in time. Additions made to the collection after the publication of a given volume would change what was known about the collection, as attested by the large number of manuscript annotations made to amend the catalogue entries as knowledge about incunabula and their provenance changed. Thus, as a proxy for the collection, the *BMC* corpus can only reveal what cataloguers and curators knew about the collection at the time of cataloguing and was dependent on the cataloguing aids that were available to them. As Pollard wrote, parts of the incunabula collection he was describing were "merely fragmentary" making in turn his "inferences [..] necessarily tentative".[49]

5. Conclusion

It follows then that pipelines for enriching catalogues using the approaches described in this chapter should acknowledge practitioner and institutional knowledge. In practice, however, scaling up such approaches is dependent on institutional priorities, allocation of resources to experimental work with uncertain outcomes, and collaboration between custodians, experts and users with different skills, perspectives, and needs. We argue that the catalogue research of the kind conducted in this chapter is a useful catalyst for scaling up, foregrounding as it does the complexities of working with historical catalogues and encouraging user communities not only to explore the representativeness of a catalogue in relation to a given collection, but also to consider the complex practice of cataloguing as expressed through linguistic choices rendered at scale. These complexities and choices are now read by machines as well as humans. As libraries, and other cultural heritage institutions, become more actively involved in conversations around artificial intelligence and its application in both collection based and research practice, this work

explained with time pressure, or lack of expertise. For example, the early cataloguers did not have the knowledge to identify provenance marks in books from the Old Royal Library or Hans Sloane collection. As John Goldfinch and colleagues developed expertise in the provenance for these collections, they made annotations about them in the *BMC* curatorial working copy.

[49] See *BMC* vol. 3 Introduction, p. xii. "The Museum collection, of course, aims only at being representative, and in the case of vernacular books and of every kind of fugitive literature it is merely fragmentary. An inferences drawn from it are thus necessarily tentative and certain to have to be modified when the Kommission fuer den Gesamtkatalog der Wiegendrucke has complete its work. But as this side of the interest of Incunabula has been neglected, it is worth while to illustrate its interest even by an avowedly tentative study."

has highlighted the potential and limitations of automation when used with historical catalogues, the importance of sharing institutional knowledge and building staff expertise around data rich approaches, and how collaborations with researchers can promote the value of catalogues as data.[50] As ever it was, the choices we make now will shape the use of our catalogues, the understanding of our collections, and the potential of catalogues as data.

Acknowledgements

Our thanks to the British Library Digital Research Team for supporting Atanassova's fellowship, to the Library's Metadata Services Team for providing access to catalogue data, to Harry Lloyd and Isaac Dunford for writing code and processing data, and to Karen Limper-Herz for sharing her specialist expertise. All errors remain our own.

This research was funded by the Arts and Humanities Research Council (UK) and Research Libraries UK Professional Practice Fellowship Scheme for academic and research libraries.

References

British Library (2021), 19th Century Books - metadata with additional crowdsourced annotations doi.org/10.23636/BKHQ-0312

British Library (2022) Race Equality Action Plan. https://blogs.bl.uk/files/bl-race-equality-action-plan-jan-2022.pdf

British Library (2023a) Knowledge Matters. The British Library 2023-2030 https://www.bl.uk/about-us/Knowledge-Matters-British-Library-Strategy-2023-30.pdf

British Library (2023b) Digital Scholarship Training Programme https://web.archive.org/web/20230927085856/https:/www.bl.uk/projects/digital-scholarship-training-programme

Danskin, A. (2023) Challenging legacies at the British Library. *Art Libraries Journal* **48** Special Issue 2: Cataloguing Ethics, 38-42

[50] https://github.com/LibraryOfCongress/labs-ai-framework

Dunford, I. (2023) Detecting Catalogue Entries in Printed Catalogue Data. *The British Library Digital Scholarship Blog,* May 2. https://blogs.bl.uk/digital-scholarship/2023/05/detecting-catalogue-entries-in-printed-catalogue-data.html

Harris, P. R. (1998), *History of the British Museum Library 1753-1973*, British Library.

Hellinga, L. (2007) *British Museum Catalogue of Books Printed in the 15th century. England. Vol. 11*.

Keinan-Schoonbaert, A. (2024) Join the British Library as a Digital Curator, OCR/HTR. *The British Library Digital Scholarship Blog,* June 26 https://blogs.bl.uk/digital-scholarship/2023/10/join-the-british-library-as-a-digital-curator-ocr-htr.html

Living With Machines (2023) Achievements https://livingwithmachines.ac.uk/achievements/

Lloyd, H. (2023) Convert-a-Card: Helping Cataloguers Derive Records with OCLC and Python. *The British Library Digital Scholarship Blog,* September 21 https://blogs.bl.uk/digital-scholarship/2023/09/convert-a-card-helping-cataloguers-derive-records-with-oclc-and-python.html

Morris, V. (2019) Automated Language Identification of Bibliographic Resources, *Cataloguing & Classification Quarterly,* **58**, 1-27 https://doi.org/10.1080/01639374.2019.1700201

Proctor, R. (1898, 1902) *An index to the early printed books in the British Museum from the invention of printing to the year MD, with notes of those in the Bodleian Library*. Kegan, Paul, Trench, Trübner & Co.

Rees, G., Gadd, S., Horgan, J, Hunt, A., Isaksen, L., Morris, V., Musson, A., Simon, R., Strachan, P., Vitale, V. (2022) *Locating a National Collection Final Report Foundation Projects 2022* https://doi.org/10.5281/zenodo.7071654

Research Libraries UK (2023) Research Libraries make it happen: RLUK Statement of Support for the Technician Commitment https://www.rluk.ac.uk/rluk-technician-commitment/

Rhodes, D. E. (ed.) (1970) *Essays in honour of Victor Scholderer*, Karl Pressler.

Ridge, M. (2022) Importing images into Zooniverse with a IIIF manifest: introducing an experimental feature. *The British Library Digital Scholarship Blog,* April 20 https://blogs.bl.uk/digital-scholarship/2022/04/importing-images-into-zooniverse-with-a-iiif-manifest-introducing-an-experimental-feature.html

Ryan, Y. (2021) A Short Guide to Historical Newspaper Data, Using R. https://bookdown.org/yann_ryan/r-for-newspaper-data/

Trustees of the British Museum (1908-1971). *Catalogue of books printed in the XVth century now in the British Museum* Parts I-X Published by Trustees of the British Museum (1908-1971).

Trustees of the British Museum (1963). *Catalogue of books printed in the XVth century now in the British Museum* Parts I-VIII Lithographic Reprint Published by Trustees of the British Museum (1963).

Trustees of the British Museum (2007). *Catalogue of books printed in the 15th century now in the British Library*, BMC Part XI England, Hes & De Graaf Publishers BV (2007).

Vane, O. (2021) Press Tracer: Visualise Newspaper Lineage. *Living With Machines Blog*, November 17. https://livingwithmachines.ac.uk/press-tracer-visualise-newspaper-lineage/

Wilson, N. (2019) The British Library's new Collection Management Strategy. *The British Library Digital Scholarship Blog*, April 12. https://blogs.bl.uk/digital-scholarship/2019/04/the-british-librarys-new-collection-metadata-strategy.html