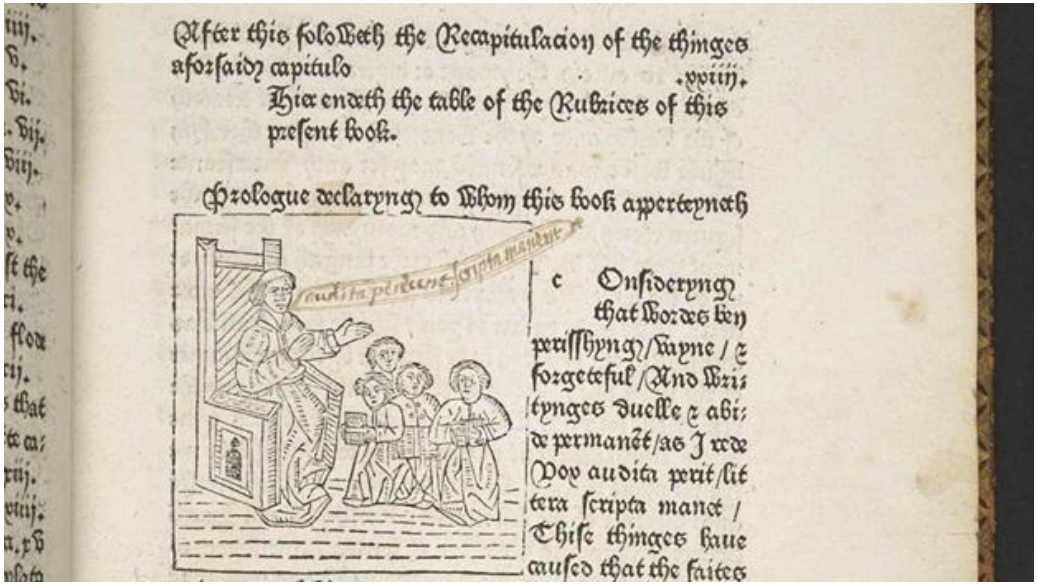
Incunabula catalogue data workshop

DHNB, Reykjavik
29 May 2024


After this foloweth the Recapitulacion of the thinges aforsaidy capitulo .xviij.

Hier endeth the table of the Rubrices of this present book.

Prologue declaryng to whom this book appertayneth

audita pereunt scripta manent

Onsideryng that wordes ben perisshyng/vayne/ & forgeteful/And writynges duelle & abide permanēt/as I rede Voy audita perit/litera scripta manet/ Thise thinges haue causyd that the faites

William Caxton's 1481 edition of the 'Mirror of the World'. BL shelfmark IB.55040.

Agenda

Welcome

08:30-09:15
Intro to Catalogues as Data

Transforming Printed Catalogues into Data:
Use case and data preparation (training data, transcription with Transkribus)
Code development and data outputs

09:15-10:00
Hands-on Exercise:
Catalogue entry extraction from transcribed text using a Jupyter notebook

10:00-10:30
Comfort Break

10:30-10:50
Introduction to the corpus analysis tool AntConc:
Set up, Data import, Tools and Setting

10:50-11:45
Hands-on Exercise: Analysis with AntConc

11:45-12:00
Closing discussion

British Library Digital Research Team

Neil Fitzgerald
Head of Digital Research

Stella Wisdom
Contemporary British

Nora McGregor
Europe & Americas

Dr Mia Ridge
Western Heritage

Dr Adi Keinan-Schoonbaert
Asia & Africa

Dr Rossitza Atanassova
Digitisation

Deirdre Sullivan
Business Support Officer

Harry Lloyd
Research Software Engineer

Erin Burnand
Universal Viewer Product Owner

Saira Akhter
Research Software Engineer

James Misson
Research Software Engineer

Lanie Okorodudu
Senior Testing Engineer

Catalogues as Data: The GLAM Context

- Understand better library and archival collections (publishing trends)
- Digitisation practice and resources (LWM map/newspapers tools)
- Metadata analysis and enhancement at scale (language, subjects, genres, named entities)
- Linked data

Some great examples:

- English Short Title Catalogue (Tolonen, Mikko et al.)
- British and Irish Newspapers Title List and Handbook (Ryan, Yann)
- Automated Language Identification (Morris, Victoria)

Harnessing Machine Learning and AI

- Use with new methods and tools
- Train language models
- Extract information and use crowdsourced information
- Build resources for GLAM professionals
- Facilitates computational analysis of collections

Some great examples:

Introduction — Classifying 19th Century British Library books using Crowdsourcing and Machine Learning (living-with-machines.github.io)

Convert a Card project and web app

Corpus linguistic analysis with catalogue data

- Examine the outputs from automated text recognition (errors, inconsistencies, variants, punctuation, special characters/scripts)
- Understand the language used by cataloguers and curators over time
- Address any problematic or offensive terminology
- Identify patterns in the data and research questions
- Identify curatorial voices and any biases
- Explore the relationships between the descriptions (items, named entities)
- Explore the transmission of descriptions between catalogues and the relation to other resources
- Inform the development of inclusive descriptive curatorial and cataloguing practice

Corpus linguistic method tested by https://cataloguelegacies.github.io/

“Given a set of guidelines for producing curatorial descriptions, corpus techniques could be used to check the extent to which guidelines are being followed at a macro-level, e.g. by identifying what aspects of objects tend to be referred to or not, and by gauging the overall extent of description versus interpretation and evaluation.

Further, such analysis could form a basis for plans to edit and enhance a catalogue by providing areas to focus on and estimates of the person time required. It could also be that a corpus-based characterization of the language used in an exemplary catalogue could be used to develop or refine guidelines by identifying that catalogue’s distinctive linguistic features.”

Salway, Andrew; Baker, James (2020). *Investigating curatorial voice with corpus linguistic techniques: the case of Dorothy George and applications in museological practice.* University of Sussex. Journal contribution. https://hdl.handle.net/10779/uos.23475305.v1

Incunabula catalogue: Project background

AHRC-RLUK funded Professional Practice Fellowship Project (2022-23): Legacies of Curatorial Voice in the Descriptions of Incunabula Collections at the British Library

Project aims to demonstrate new ways of working with digitised catalogues that would also improve the discoverability and usability of the collections they describe.

The focus of the research is the *Catalogue of books printed in the 15th century now at the British Museum* (or *BMC*) published between 1908 and 2007 which describes over 12,700 volumes from the British Library incunabula collection. By using computational approaches and a corpus analysis tool with the catalogue data we gain new insights into this valuable resource and enable its future reuse in contemporary online resources.

Project summary is available at https://app.transkribus.org/sites/BL-Incunabula

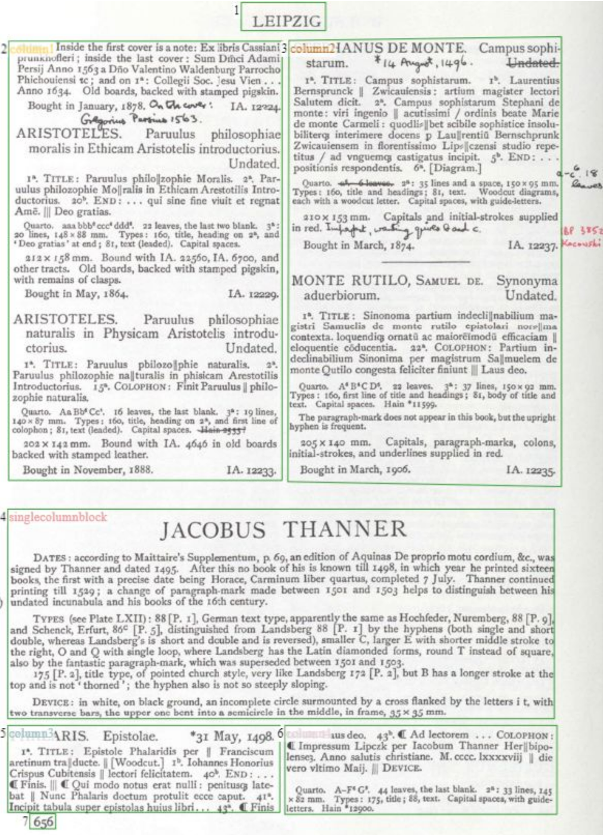
Incunabula catalogue: Data preparation steps

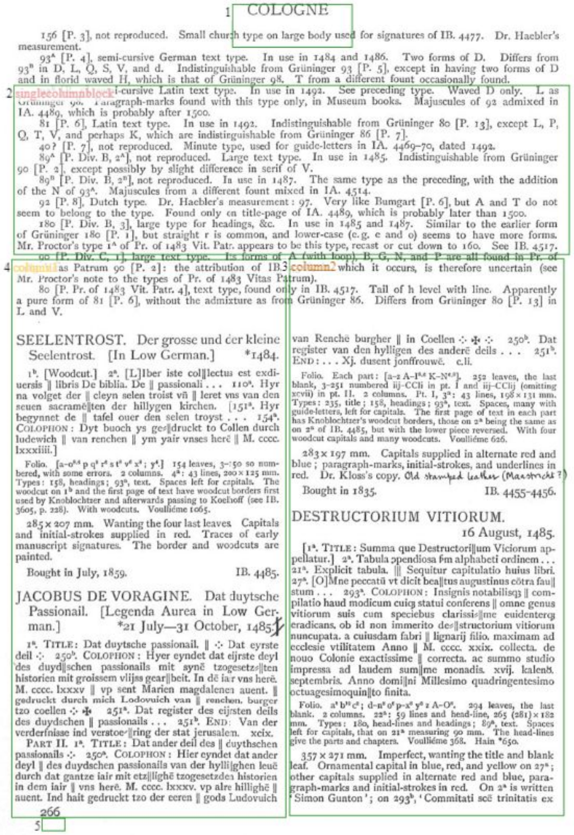
Transkribus:

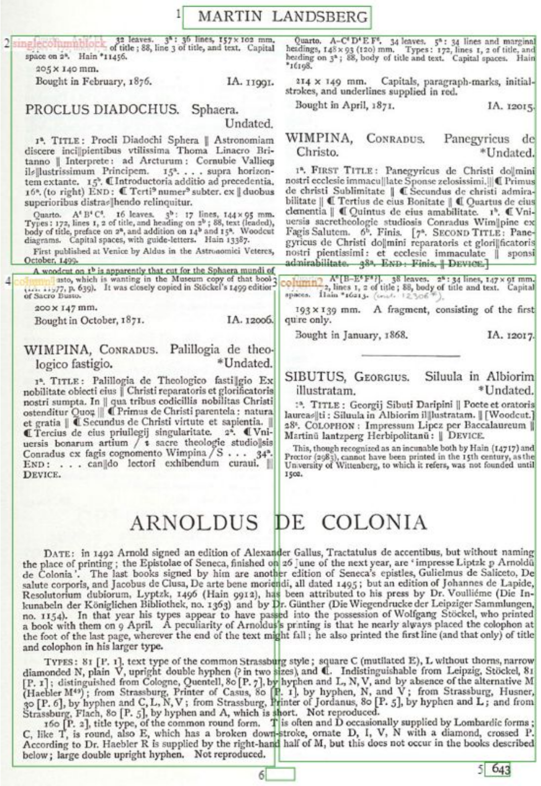
- Training the layout recognition models
- Optical Character Recognition
- Transkribus Outputs

Our text data is noisy. This is due to some errors in the layout recognition and special characters/abbreviations in the transcriptions of the titles.

Early experiments







BMC_10 2 column model output

Select

BMC_7 2 column model output

Select

BMC_7 sixth training set for 4 column model

Select

BMC_3 2 column model output

Select

BMC_2 2 column model output

Select

BMC_3 fifth training set for 4 column model

Select

comparison 2 column on original and new s

Select

BMC 5 2 column model output

Select

Sudio test images

Select

BMC_5 third training set for 4 column model

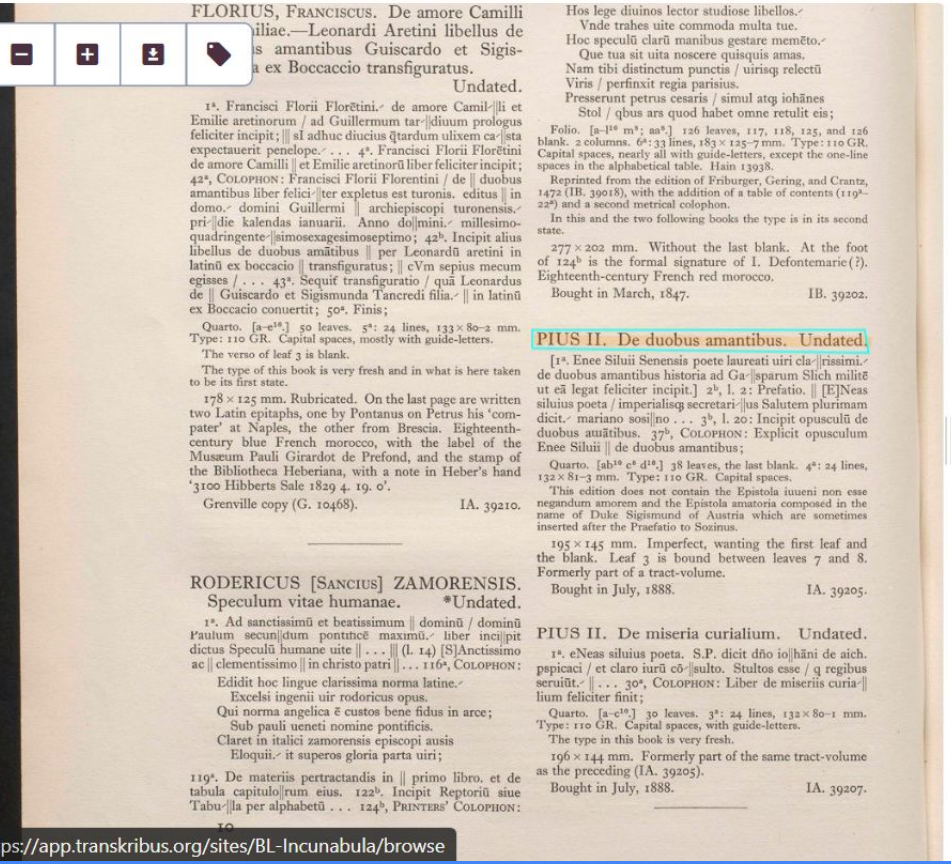
Select

BMC_5 test images for revised 2 column mo

Select

BMC_2 test images for 4 column model

Select



FLORIUS, FRANCISCUS. De amore Camilli
iliae.—Leonardi Aretini libellus de
s amantibus Guiscardo et Sigis-
a ex Boccaccio transfiguratus.
Undated.

1ª. Francisci Florii Florẽtini. de amore Camil-li et Emilie aretinorum / ad Guillermum tar-diuum prologus feliciter incipit; sI adhuc diucius q̃tardum ulixem ca-sta expectauerit penelope. . . . 4ª. Francisci Florii Florẽtini de amore Camilli et Emilie aretinorũ liber feliciter incipit; 42ª, COLOPHON: Francisci Florii Florentini / de duobus amantibus liber felici-ter expletus est turonis. editus in domo. domini Guillermi archiepiscopi turonensis. pri-die kalendas ianuarii. Anno do|mini. millesimo-quadringente-simosexagesimoseptimo; 42ᵇ. Incipit alius libellus de duobus amãtibus per Leonardũ aretini in latinũ ex boccacio transfiguratus; cVm sepius mecum egisses / . . . 43ª. Sequit transfiguratio / quã Leonardus de Guiscardo et Sigismunda Tancredi filia. in latinũ ex Boccacio conuertit; 50ª. Finis;

Quarto. [a-e¹⁰.] 50 leaves. 5ª: 24 lines, 133×80–2 mm. Type: 110 GR. Capital spaces, mostly with guide-letters.

The verso of leaf 3 is blank.

The type of this book is very fresh and in what is here taken to be its first state.

178×125 mm. Rubricated. On the last page are written two Latin epitaphs, one by Pontanus on Petrus his 'compater' at Naples, the other from Brescia. Eighteenth-century blue French morocco, with the label of the Musæum Pauli Girardot de Prefond, and the stamp of the Bibliotheca Heberiana, with a note in Heber's hand '3100 Hibberts Sale 1829 4. 19. 0'.

Grenville copy (G. 10468). IA. 39210.

RODERICUS [SANCIUS] ZAMORENSIS.
Speculum vitae humanae. *Undated.

1ª. Ad sanctissimũ et beatissimum dominũ / dominũ Paulum secun|dum pontificẽ maximũ. liber inci|pit dictus Speculũ humane uite | . . . | (l. 14) [S]Anctissimo ac | clementissimo | in christo patri | . . . 116ª, COLOPHON:

Edidit hoc lingue clarissima norma latine.
Excelsi ingenii uir rodoricus opus.
Qui norma angelica ẽ custos bene fidus in arce;
Sub pauli ueneti nomine pontificis.
Claret in italici zamorensis episcopi ausis
Eloquii. it superos gloria parta uiri;

119ª. De materiis pertractandis in primo libro. et de tabula capitulo|rum eius. 122ᵇ. Incipit Reptoriũ siue Tabu-la per alphabetũ . . . 124ᵇ, PRINTERS' COLOPHON:

Hos lege diuinos lector studiose libellos.
Vnde trahes uite commoda multa tue.
Hoc speculũ clarũ manibus gestare memẽto.
Que tua sit uita noscere quisquis amas.
Nam tibi distinctum punctis / uirisq; relectũ
Viris / perfinxit regia parisius.
Presserunt petrus cesaris / simul atq; iohãnes
Stol / qbus ars quod habet omne retulit eis;

Folio. [a-l¹⁰ m⁸; aa⁸.] 126 leaves, 117, 118, 125, and 126 blank. 2 columns. 6ª: 33 lines, 183×125–7 mm. Type: 110 GR. Capital spaces, nearly all with guide-letters, except the one-line spaces in the alphabetical table. Hain 13938.

Reprinted from the edition of Friburger, Gering, and Crantz, 1472 (IB. 39018), with the addition of a table of contents (119ª–22ª) and a second metrical colophon.

In this and the two following books the type is in its second state.

277×202 mm. Without the last blank. At the foot of 124ᵇ is the formal signature of I. Defontemarie(?). Eighteenth-century French red morocco.

Bought in March, 1847. IB. 39202.

PIUS II. De duobus amantibus. Undated.

[1ª. Enee Siluii Senensis poete laureati uiri cla-rissimi. de duobus amantibus historia ad Ga-sparum Slich militẽ ut eã legat feliciter incipit.] 2ᵇ, l. 2: Prefatio. [E]Neas siluius poeta / imperialisq secretari-us Salutem plurimam dicit. mariano sosi|no . . . 3ᵇ, l. 20: Incipit opusculũ de duobus auãtibus. 37ᵇ, COLOPHON: Explicit opusculum Enee Siluii de duobus amantibus;

Quarto. [ab¹⁰ c⁸ d¹⁸.] 38 leaves, the last blank. 4ª: 24 lines, 132×81–3 mm. Type: 110 GR. Capital spaces.

This edition does not contain the Epistola iuueni non esse negandum amorem and the Epistola amatoria composed in the name of Duke Sigismund of Austria which are sometimes inserted after the Praefatio to Sozinus.

195×145 mm. Imperfect, wanting the first leaf and the blank. Leaf 3 is bound between leaves 7 and 8. Formerly part of a tract-volume.

Bought in July, 1888. IA. 39205.

PIUS II. De miseria curialium. Undated.

1ª. eNeas siluius poeta. S.P. dicit dño io|hãni de aich. pspicaci / et claro iurũ cõ-sulto. Stultos esse / q regibus seruiũt. | . . . 30ª, COLOPHON: Liber de miseriis curia-lium feliciter finit;

Quarto. [a-c¹⁰.] 30 leaves. 3ª: 24 lines, 132×80–1 mm. Type: 110 GR. Capital spaces, with guide-letters.

The type in this book is very fresh.

196×144 mm. Formerly part of the same tract-volume as the preceding (IA. 39205).

Bought in July, 1888. IA. 39207.

Bought in March, 1847.

IB. 39202.

PIUS II. De duobus amantibus. Undated.

1a. Enee Siluii Senensis poete laureati uiri cla rissimi. de duobus amantibus historia ad Ga-sparum Slich militem ut ea legat feliciter incipit.) 2b, l. 2: Prefatio. ENeas siluius poeta/ imperialisque secretari us Salutem plurimam dicit. mariano sosi no ... 3b, l. 20: Incipit opusculu de duobus auiatibus. 376, COLOPHON: Explicit opusculum Enee Siluii de duobus amantibus;

Quarto. Jab'° ce d'°. 38 leaves, the last blank. 42: 24 lines, 132X 81-3 mm. Type: 110 GR. Capital spaces.

This edition does not contain the Epistola iuueni non esse negandum amorem and the Epistola amatoria composed in the name of Duke Sigismund of Austria which are sometimes inserted after the Praefatio to Sozinus.

195 X 145 mm. Imperfect, wanting the first leaf and the blank. Leaf 3 is bound between leaves 7 and 8.

Incunabula catalogue data: Code development

Entry detection algorithm

Language detection

GitHub Repository

Exercise:

Follow the link to the Binder notebook to understand the catalogue entry extraction process.

Reusable datasets

Incunabula catalogue dataset published on the British Library Shared Research Repository include:

- 3500 Images (JPG) and corresponding XMLs (ALTO and PAGE)
- Catalogue entries in full and English-only descriptions (TXT)
- Metadata (CSV)
- Code

DH2023 Poster

COFFEE

BREAK

Computational Analysis with AntConc

Intro to AntConc

Set up v.4.2.4

Our files are in this folder Download files for vols.1-10.

Data import - File-Open Corpus Manager. Raw Files. Add files. Create.

Tools and Settings (also described in the AntConc Help pages)

Global Settings - Restore Settings - Yes - saves your Settings for the next session

Word lists

Word lists are a useful starting point for getting an overview of the linguistic features of a corpus.

- Select the Word Tool and press Start to generate a word list, using the default settings.
- Note the top 30 word types (over 30% of the words in the corpus)? Do all the words/strings make sense?
- Select the case setting and run the query again. Note the difference.
- What named entities appear in the top 200 word types?
- Click on *King* to explore how the word is used in the corpus.
- Select the Invert Order setting to browse the long tail of infrequently used word types. Eg. mostly Latin words some names, strings of letters. Note some of the transcription errors!
- Sort by Type to browse word types alphabetically (capitalised hits at top). Eg. exchange (25) vs. exchanged (1)

For the domain-specific terms you may refer to this Glossary of Incunabula

If saving the outputs of your searches, ensure the filename includes information about your corpus, query, settings, or AntConc version. You can copy and paste the results or save results from the File menu (Ctrl+S)

KWIC (Key-Word-in-Context)

Search for a word or a ‘string’ to narrow your enquiry to a subset of the corpus. KWIC displays the results in a concordance.

Search for *appear* (case selected) as a word and then as a string (deselect word)

Sorted by pattern frequency in which the word appears (by default Sort to Right)

Use the Sort to left option to examine how ‘appear’ has been used by cataloguers, eg. with modal verbs, with negative words, the consistency of use.

Adjust the Sort Options according to which word relative to the search term you want to inspect.

Order by value to browse the word alphabetically (to the left or right).

This tool can help with comparative analysis of catalogue data, eg. language patterns, curatorial voices, across collections and over time.

For example the cataloguer of vol.7 and 8 adopted the phrase ‘there appear to be no (sure) means of determining/deciding …’

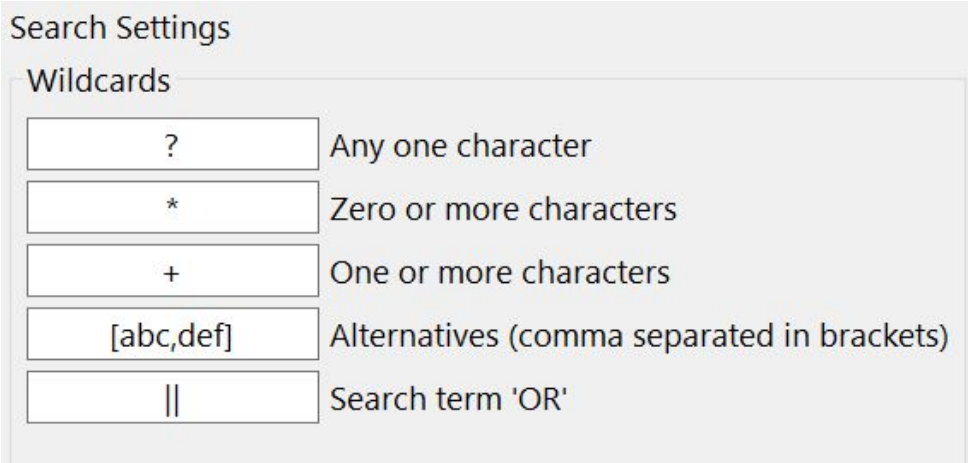
Expand the Context Size to see more of the lines and click on the word to go to the relevant File for close reading of the text.

Use the Case Setting to search with/without capitalisation - eg. ‘*Wanting* the blank leaf’ vs ‘Imperfect, *wanting*…’ These patterns reflect the rules of cataloguing, others - the cataloguers’ linguistic choices - eg. ‘was *presumably* printed **not later than/long after/very long after**….

Wildcards

- *determin+* results show it is only used as a verb/verb form and not as a noun
- *wear** (query as a word or a string)
- **ly* brings up mostly adverbs but also names and nouns, July, Lilly, fly. Note some transcription errors, hyphenated words
- *Be||being||was||were||am||been*
- In the Word Tool you can refine your search by defining the min Frequency or Range (size of the corpus)
- You can also use regex commands
- Choose a word type to explore in context
- Discuss with your neighbour

- Wildcards help to explore known variants



Unusual or historical language

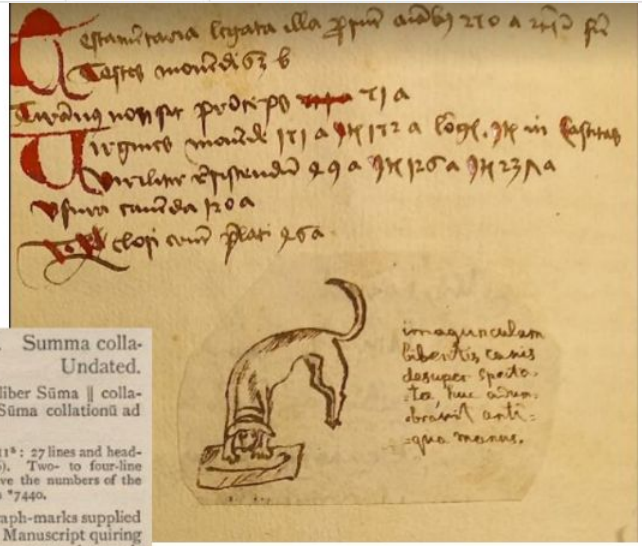
	File	Left Context	Hit	Right Context
1	BMC_1_...	green cap, wearing a black sword, and one in a blue	jerkin	and dull crimson hose and cap. Title. De uidicta libertate
2	BMC_1_...	and cap and green hose, wearing sword ; one in dark blue	jerkin,	hose, and cap ; one in grey jerkin, green hose white
3	BMC_1_...	cap. Before them, with back to spectator, a man in blue	jerkin	and hose and scarlet cap. On the left a man
4	BMC_1_...	trimmed with green. Before him four men: one in dull crimson	jerkin	and cap and green hose, wearing sword ; one in dark
5	BMC_1_...	hose white apron, and black hat, and one in dull crimson	jerkin,	blue hose, and green cap. Title. De seruis fugitiuis et
6	BMC_1_...	green-clad figure of the Prologue miniature. A man in grey	jerkin	and scarlet hose is leaving the room. Title. De edendo.
7	BMC_1_...	one in dark blue jerkin, hose, and cap ; one in grey	jerkin,	green hose white apron, and black hat, and one in
8	BMC_1_...	in blue robe trimmed with grey fur, the other in green	jerkin	and hose, and holding a golden rod. Both wear dull
9	BMC_1_...	in his hand a docu ment with seals; one in red	jerkin,	grey hose, and green cap, wearing a black sword, and
10	BMC_1_...	robe and dull crimson cap ; behind him a man in yellow	jerkin	and hose and dull crimson cap shows a book to

Appropriate use

	File	Left Context	Hit	Right Context
1	BMC_1_s...	capital on 22 supplied in red, green, and yellow, with a	grotesque	head, other capitals (with some omissions), paragraph-marks, initial-
2	BMC_1_s...	page cuts. A five-line woodcut D on 22 and 4°; 15-line	grotesque	A on 2b ; spaces, with guide-letters, left for
3	BMC_1_s...	phia, and the Muse Meretricule' on 12, 4°, and", "1092. Thirteen-line	grotesque	figure capitals. Hain '3359.", '183X 130 mm. On the title-page
4	BMC_8_s...	evet, 1493). From the Old Library. Undated. Type: 114 G. Calligraphic	grotesque	L beginning title, Lom¬ bard Q on 22.', "The grotesque
5	BMC_8_s...	Calligraphic grotesque L beginning title, Lom¬ bard Q on 22.', "The	grotesque	L'on 1e was also used in an unsigned
6	BMC_8_s...	be earlier than this. The calli- graphic L with three	grotesque	faces which starts the title apparently occurs again, with
7	BMC_8_s...	Passion, etc. are those used by Jehannot. The borders, representing	grotesque	beasts and faces and in one case fleur-de-
8	BMC_8_s...	in this style used at Lyons. Calligraphic L, with two	grotesque	faces to the right, the neck of the upper

Total Hits: 8 **Page Size** All hits 1 to 8 of 8 hits

	File	Left Context	Hit	Right Context
1	BMC_1_s...	before a charnel house (1 6 verso), a Gallant followed by his	dog (17	recto), a Deathbed (I 8 recto). 210X 145 mm. Bought in
2	BMC_1_s...	second is pasted a pen-and-ink drawing of a	dog	drinking. Old stamped leather, rebacked. Bought in February, 1875. IA.
3	BMC_2_s...	in the neck; dexter, azure. That on the right, a	dog (?)	On 22 is the stamp of the Stadtbibliothek, Lüneburg. Old
4	BMC_4_s...	x 231 mm. With a small circular book-label (head of	dog	carrying bone). Bound before IB. 19453, a copy of the
5	BMC_5_s...	the last blank. On 16 is painted a shield with arms (	dog).	Bought in April, 1865. IC. 21419. DECISIONES. Decisiones Rotae Romana
6	BMC_5_s...	is", 'pasted a round label with the head of a	dog	holding a bone. Bought in April, 1908. IB. 22473 WITH PAGANINUS
7	BMC_6_s...	very fresh, and three of the watermarks', "found in it (	dog,	ornamental A, cardinal's hat) occur also in books", '
8	BMC_8_s...	a border-strip with the figure of a man-headed	dog	on a black ground, 75 X 23 mm., represents King Carnuant



JOHANNES GALLENSIS. Summa collationum. Undated.

1ª. Ad Omne hoīm genus Incipit liber Sūma ‖ collationū dictus . . . 262ª. COLOPHON: Sūma collationū ad ōne ge‖nus hoīm Explicit feliciter.

Quarto. [A-Z 1³ a-f⁸ g⁶.] 262 leaves. 11ª: 27 lines and headline, 143 (130) × 82 mm. Type: 96 (106). Two- to four-line spaces left for capitals. The head-lines give the numbers of the part and distinction. Voulliéme 657. Hain *7440.

211 × 145 mm. Capitals and paragraph-marks supplied in red or blue, initial-strokes in red. Manuscript quiring and foliation. A manuscript index on two leaves is bound in at the beginning. On the second is pasted a pen-and-ink drawing of a dog drinking. Old stamped leather, rebacked.

Bought in February, 1875. IA. 2897.

Collocates

Discover the words (collocates) that appear frequently near the search term.

Our corpus includes words for animals as part of descriptions of illustrations/woodcuts/watermarks, but mainly used in bookbindings descriptions.

Search for *sheep* (as a string, no case) and explore its collocate types. Results rankings show frequency to the left or right of search term.

Use the Span (proximity to search term) and Min. Freq. to refine the results. Adjust the Range (across how many volumes)

Note *sheep* and *sheepskin* are used interchangeably! This can inform revision of cataloguing rules.

Find the collocates of *before* and explore the Sort by options.

Likelihood and Effect are different ways of showing statistically significant relationships.

You can adjust the measures and thresholds in the Tool Settings for Collocates

Collocates

Shows words that appear frequently within a certain distance of the search term.

Find the collocates of *before* and explore the Sort by options.

Likelihood and Effect are different ways of showing the strength of the word relationship.

You can adjust the measures and thresholds in the Tool Settings for Collocates

Refine the search by using the Min. Frequency and Window Span options.

Cluster and N-Gram Tools

Find word patterns based on the search term.

What 2-word clusters (sequences) are there for *blind* in the corpus?

Change the Min. Frequency to display the top 5 clusters. Click on the results.

Change the Search Term Position to Right and increase the Cluster Size to 3.

What observations can you make about the cluster ‘gold and blind’?

The cataloguers of vols. 9 and 10 are most consistent in their phraseology to describe bookbindings. The detail seemed important to them and they had the expertise (and the time!)

You get similar results with the N-Gram Tool.

Plot Tool

Visualise where the search term appears in the corpus.

Useful for comparison between different search terms and examining large corpora.

Find the mentions of *morocco* (goatskin used in bindings) in the corpus

Sort by Dispersion to see how widely spread it is across the file

You can change the Dispersion statistics measures in the Tool Settings

Normalised Frequency takes into account the total number of tokens in the file.

Plot the use of colour patterns in the corpus. Use the Overlay option to compare the use of different patterns.

Search for *blue and red*, then select the Overlay option, choose another colour and search for *red and gold* Note that the majority of mention of *red and gold* is in volume 8. Click on the outliers to see the result in the file.

To save the graph as png file, select the Graphic View option in the Tool Settings

Keyword Tool

Compare words that appear unusually frequently in a target corpus in comparison with a reference corpus (eg. a specialised vs a generalised corpus) and can help characterise or understand the features of the target corpus.

Uses a number of statistical measures and thresholds that can be customised in the Tool Settings.

Likelihood - measures what is ‘statistically significantly unusual’ in the target corpus compared to the reference corpus
Effect - measures ‘the degree of unusualness’, and the higher the effect, the higher the unusualness

You can import a new corpus into the Corpus Manager and open it using the Reference Tab

You can switch easily between the Target and Reference Tabs in the Manager or from the File menu.

Selecting the negative keywords setting in Tool Settings would list the most unusually infrequent word in the target corpus.

This sheet lists the positive and negative Keyword Types characteristic of our target corpus with reference to a corpus describing historical photographs, and vice versa. Note the similarities in terms of function words, verbs, object information. The photographic descriptions refer to people, places, events, but also provide measurements and archival information.

See the Keyness lesson which compares the historical photographs corpus with a wordlist generated from the British National Corpus

Discussion

Feedback on the approach and tool. Will you apply it to your data?

What are your research questions (printing culture, ownership and provenance, named entities, outdated or problematic language)

Explore further:

NER, punctuation in the descriptions - eg. Original?, quotation marks, referenced sources, annotations for digitised corpora

Ground truth for incunabula titles transcriptions

Combine data with MSS annotations in the curators' copy of the catalogue and Contextual information from the digitised catalogue

Recommendations for inclusive descriptive policy and practice

Thank you!

digitalresearch@bl.uk

https://blogs.bl.uk/digital-scholarship/