Unseen Species Modelling Applied to Medieval Literature

Digital Approaches to Pre-Modern Texts and Manuscripts

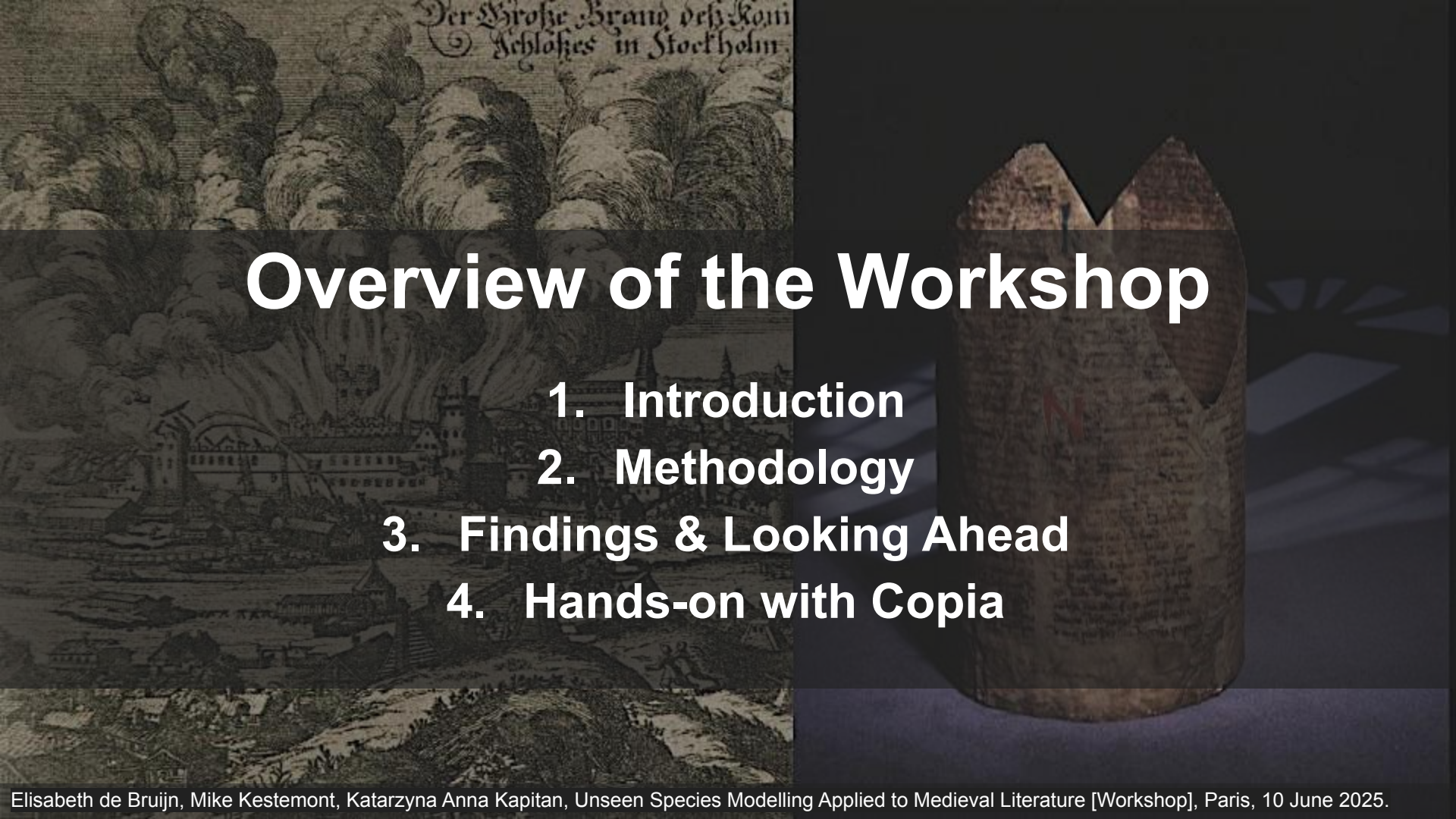
10-12 June 2025, ENC–PSL, Paris

Elisabeth de Bruijn, Mike Kestemont, and Katarzyna Anna Kapitan

Elisabeth de Bruijn, Mike Kestemont, Katarzyna Anna Kapitan, Unseen Species Modelling Applied to Medieval Literature [Workshop], Paris, 10 June 2025.

Aims For Today

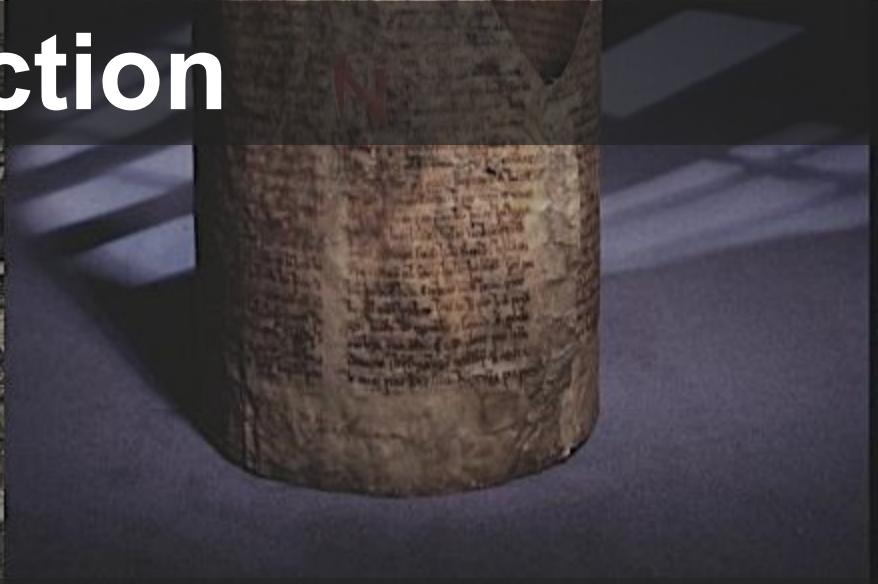
- **Practical introduction (methodology and application)**
  - state of the scholarship
  - statistical methods
- **Presentation and critique of the results**
- **Sneak and peek into current work & call for feedback**



Overview of the Workshop

1. Introduction
2. Methodology
3. Findings & Looking Ahead
4. Hands-on with Copia

Introduction



THE LOSS

What we know is that:

- the state of the cultural heritage available to us for research is only a fraction of what one existed
- diverse factors had influence on preservation of the cultural heritage

What we have is:

- pretty good estimates for the extent of loss for specific types of cultural artefacts, for example liturgical books,
or specific locations.

THE LOSS

What we do not know is:

- what is the extent of loss of other types of cultural heritage, for example medieval literary works

What we do not have is:

- a method which allows us to compare the loss rates across different domains

Loss of Manuscripts

≠

Loss of Witnesses

≠

Loss of Texts

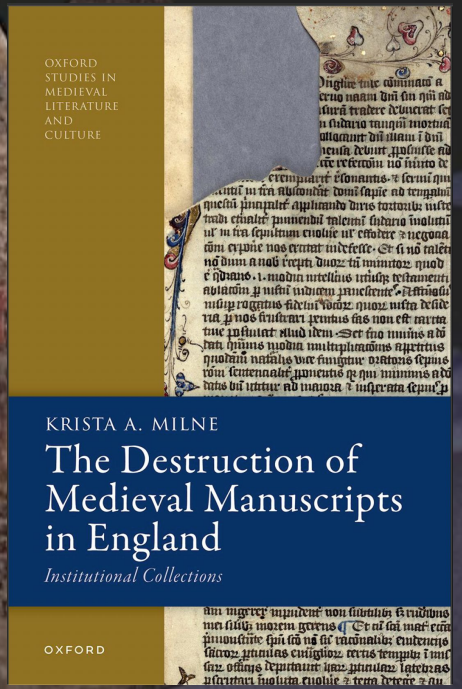
≠

Loss of Works

Factors of Loss

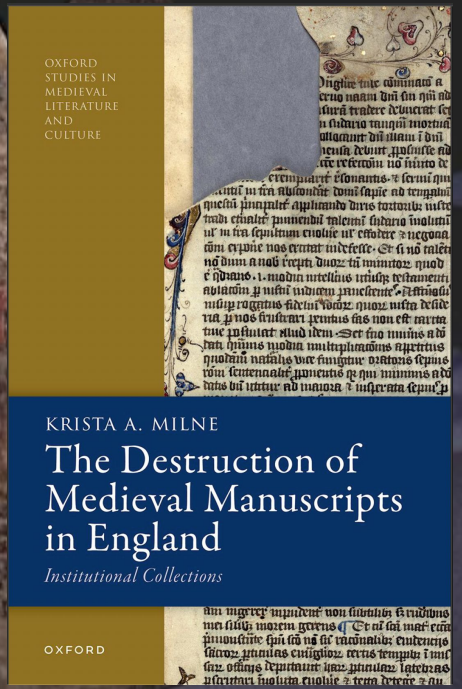
There seems to be consensus among scholars that **certainly** different factors must had influenced different survival rates, but there is little consensus **how**.

For an excellent recent overview of literature concerning English manuscripts, see: Krista Milne, *The Destruction of Medieval Manuscripts in England* (Oxford 2025).



Factors of Loss

- SIZE (small vs large)
- PRICE (decorated vs. plain)
- FORM (bound vs unbound)
- USE (liturgic, schoolbooks)
- CONTENTS (taste)
- PRESERVATION (collections)



Manuscript Loss: Individual Collections and Their Catalogues

Countless studies & broad diversity of **estimates**

1-85% survival rates, frequently around 20%.

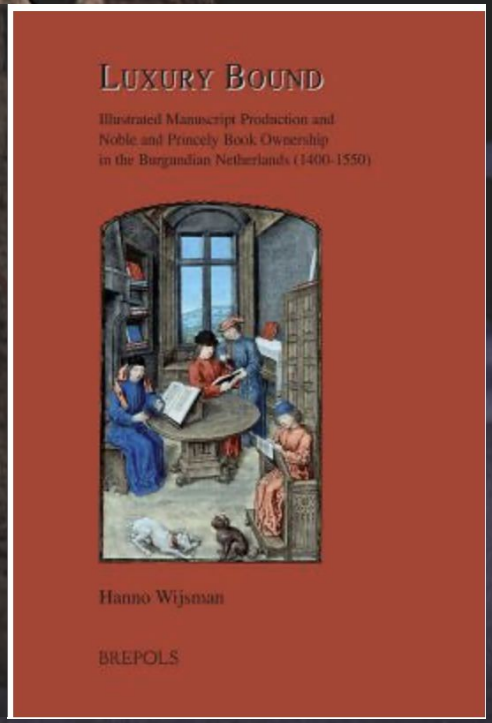
An example:

‘Ultimately, it is about a third of the collection catalogued by Dom Baudry that is currently preserved’
(Lucy Tondreau 1970, my translation)

Manuscript Loss: Type

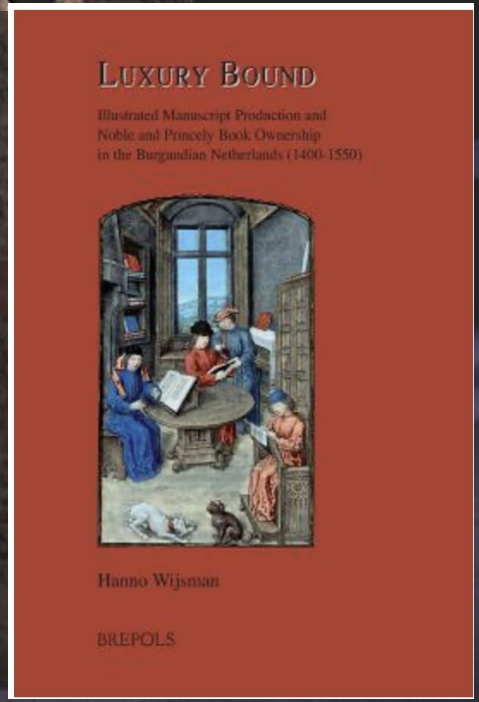
LUXURY MANUSCRIPTS
The survival rate of the luxury, decorated manuscripts: around 20% (Wijsman 2010).

‘I would cautiously suggest a survival rate: approximately one fifth of the illustrated manuscripts originally produced in the Netherlands between 1400 and 1550 has withstood the caprices of time.’ (Wijsman 2010)



Factors of Loss

- **PRESERVATION**
  Survival rates vary and depend on manuscripts' preservation history:
  - 50% - manuscripts that entered institutionalized collections before 1550.
  - 5-25% - manuscripts from private collections that have been dispersed
  - 7-15% - books of hours and prayer books

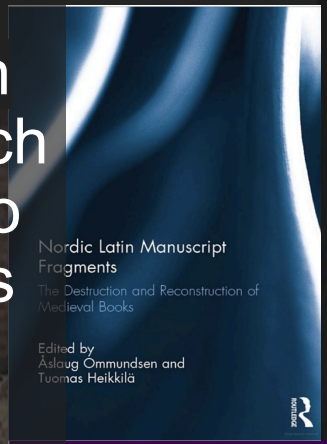


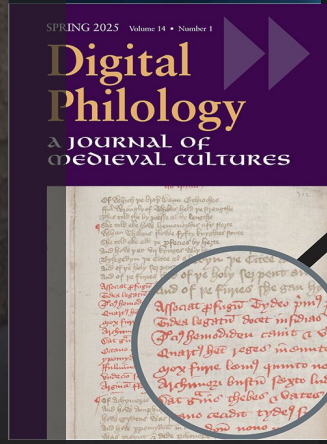
LITURGICAL MANUSCRIPTS

‘’ [In Iceland] the number of churches was about 330 in the late Middle Ages, with up to 1200 chapels [...]. [E]ach church had to have at least four books just to be able to hold a basic service [...]. There are combined remnants of roughly 340 Latin manuscripts, liturgical and others’ (Guðvarður Már 2017)

=> 340 / (330*4) max. 25% survival, but more likely 340 / (1320+1200) => max. 13%

=> Same common sense methodology for French liturgical books => max 0,4 % survival (Jaakko Tahkokallio 2025)

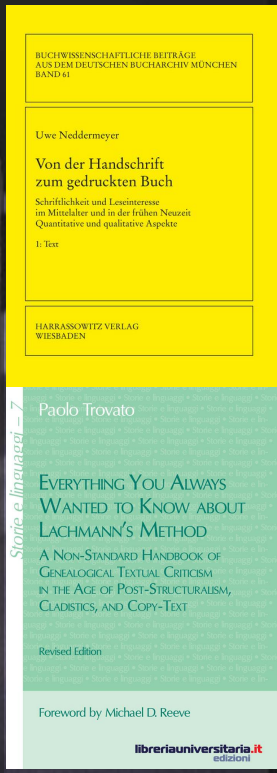




Manuscript Loss: Analogy to Incunabula & Early Prints

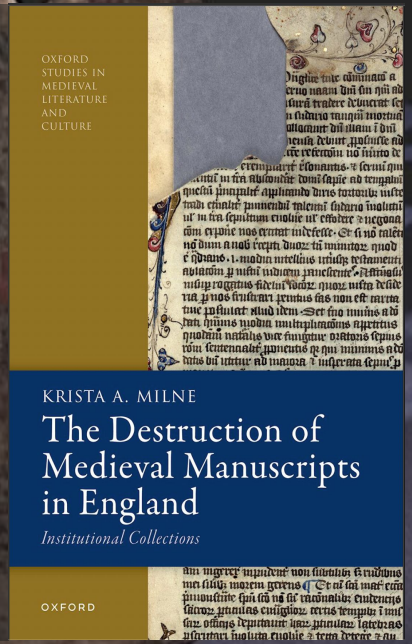
‘It seems appropriate to estimate the percentage of manuscripts contained in the collection somewhat higher than that of the printed works […] I cautiously assume that for every manuscript from the 14th or 15th century available today, there are 14 lost ones, [i.e.] 7%.’ (Neddermeyer 1998, my translation)

‘I do not see valid reasons to imagine that the manuscripts […] stood higher chances of survival.’ (Trovato 2014)



Witness Loss

The stemma-based estimations of lost witnesses can vary greatly, from 0,5% to 50% (Milne 2025)

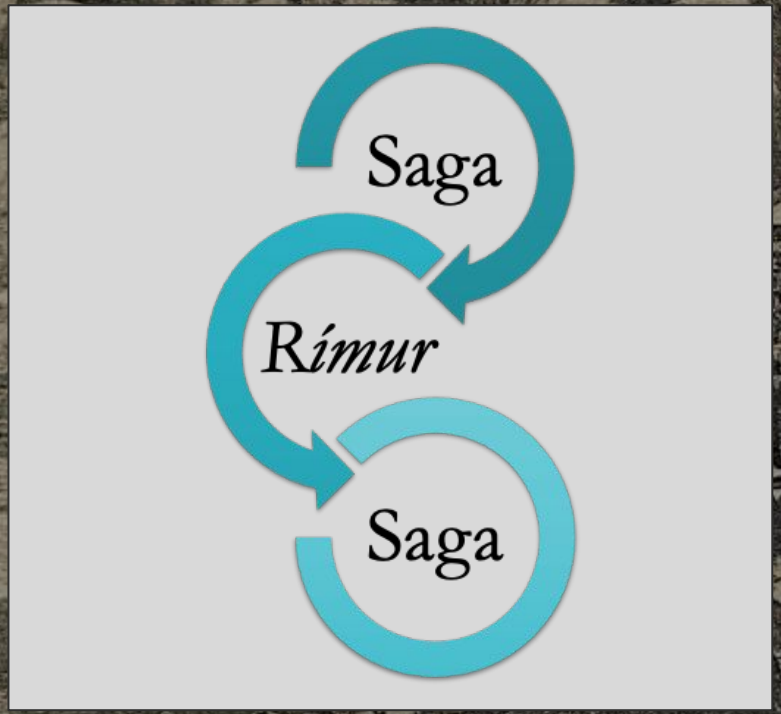


Works Loss

A ChatGPT-generated image of a skald telling a saga (oral tradition).

‘much of the lost literature never had a written existence’ (**Wilson 1952**)

Works Loss



'The *rímur* poets rarely invented their own subject matter. They preferred to take an already existing prose narrative and versify it, remaining faithful to the storyline' (Kuhn 2000).

Works Loss

‘written, medieval Icelandic literature which has not survived in manuscript form into our own times.’

‘a large part of evidence for lost […] [stories] […] is to be found in the surviving literature’ (Jesch, 1984)

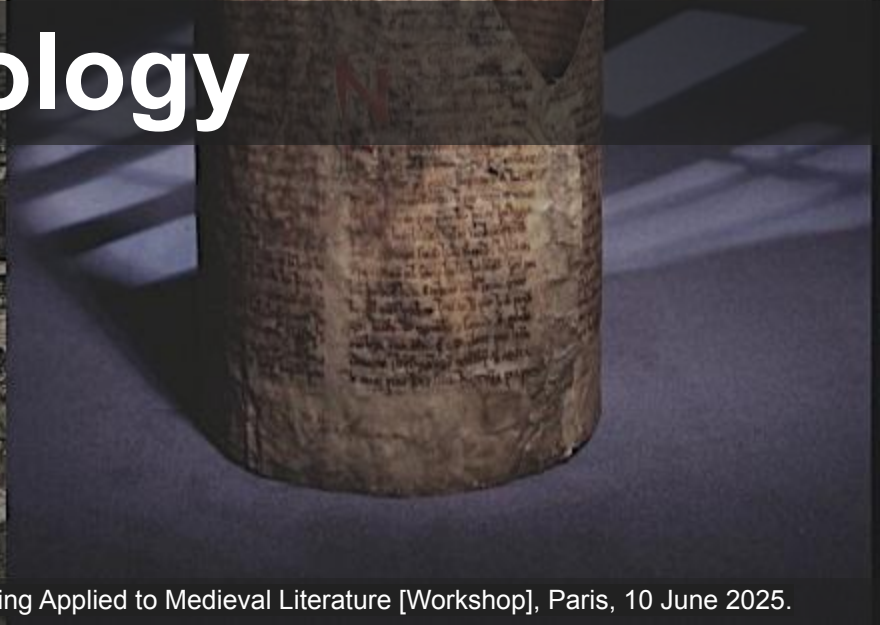
But what about romances?

Shall we accept that we cannot say anything about their survival rates?

Or is there hope?

There might be, if we turn to statistics and use unseen species models.

Methodology



Witness Loss: Statistics

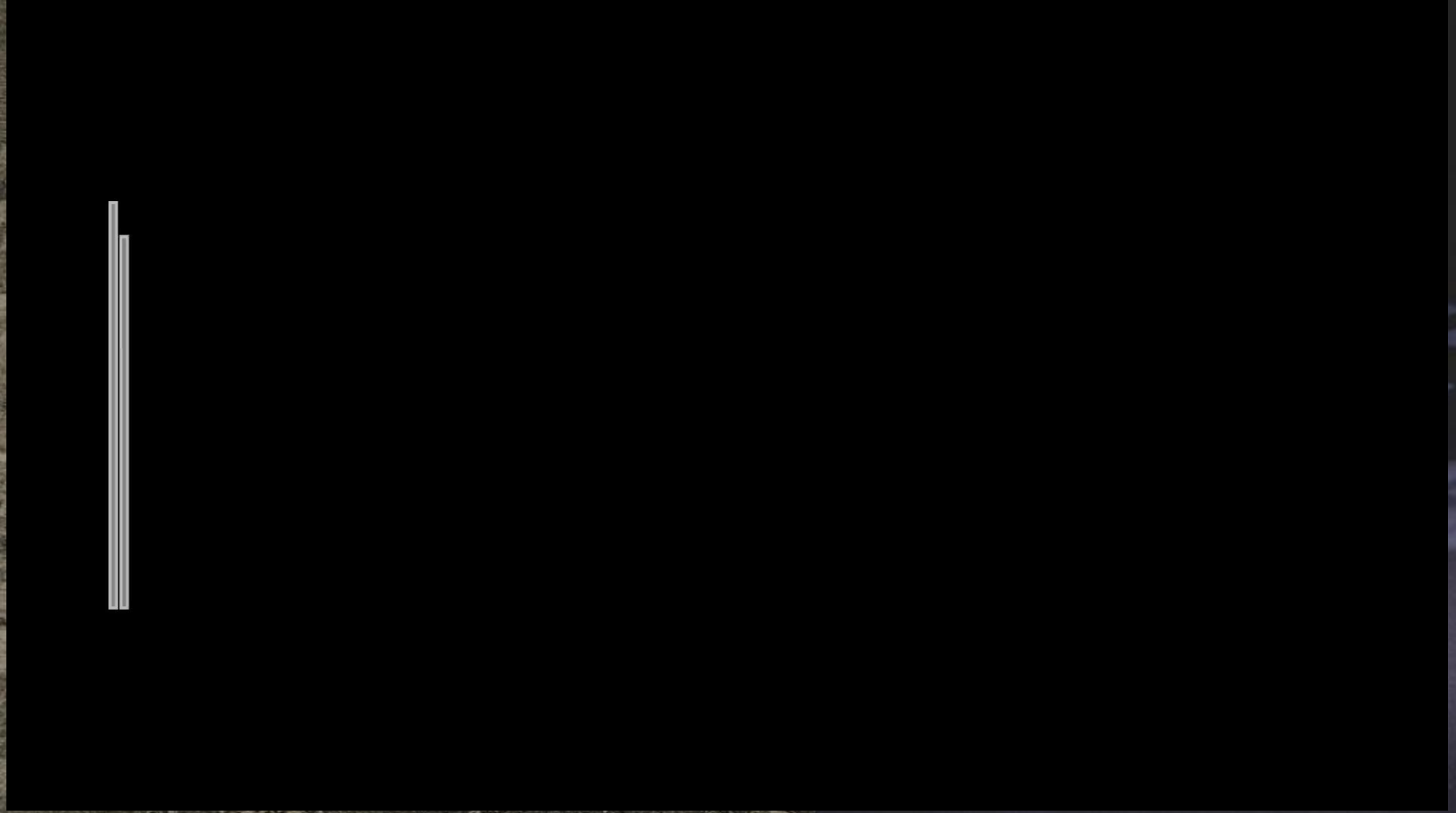
Cisne (2005)
Egghe & Proot 2007
Buringh (2011)
Camps & Randon-Furling (2022)

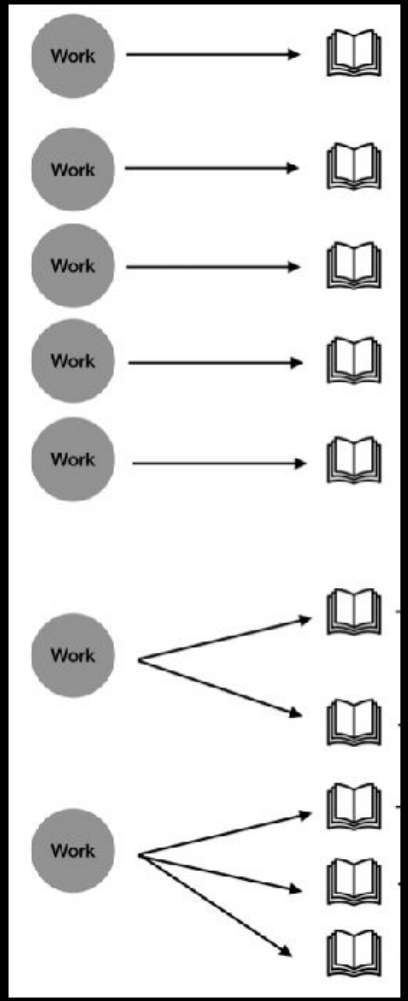
Species.	Sightings
45	1
13	2
6	3
2	4
4	5
1	6
1	7
2	10
1	17
75	167

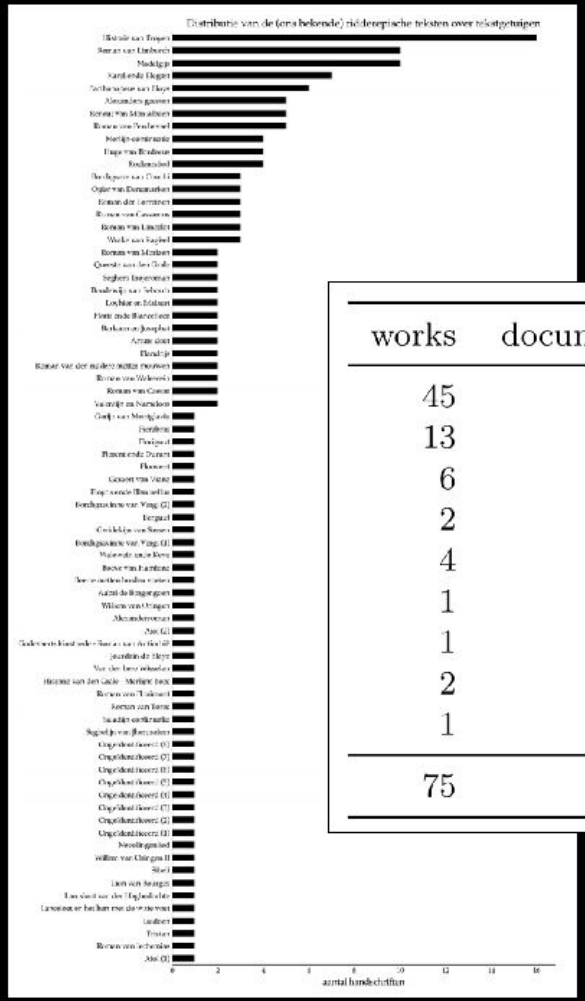
$$\hat{f}_0 = \begin{cases} \frac{(n-1)}{n}\frac{f_1^2}{(2f_2)} & if\, f_2 > 0; \\ \frac{(n-1)}{n}\frac{f_1(f_1-1)}{2} & if\, f_2 = 0 \end{cases}$$

Unseen species models

- *Chao1*: Statistical, non-parametric method by Anne Chao and colleagues
- Core assumption: # low-frequency items predict # unobserved items
- *f1* (singletons) and f2 (doubletons) to predict (lower bound on) *f0*







works	documents
45	1
13	2
6	3
2	4
4	5
1	6
1	7
2	10
1	17
75	167

Key analogy

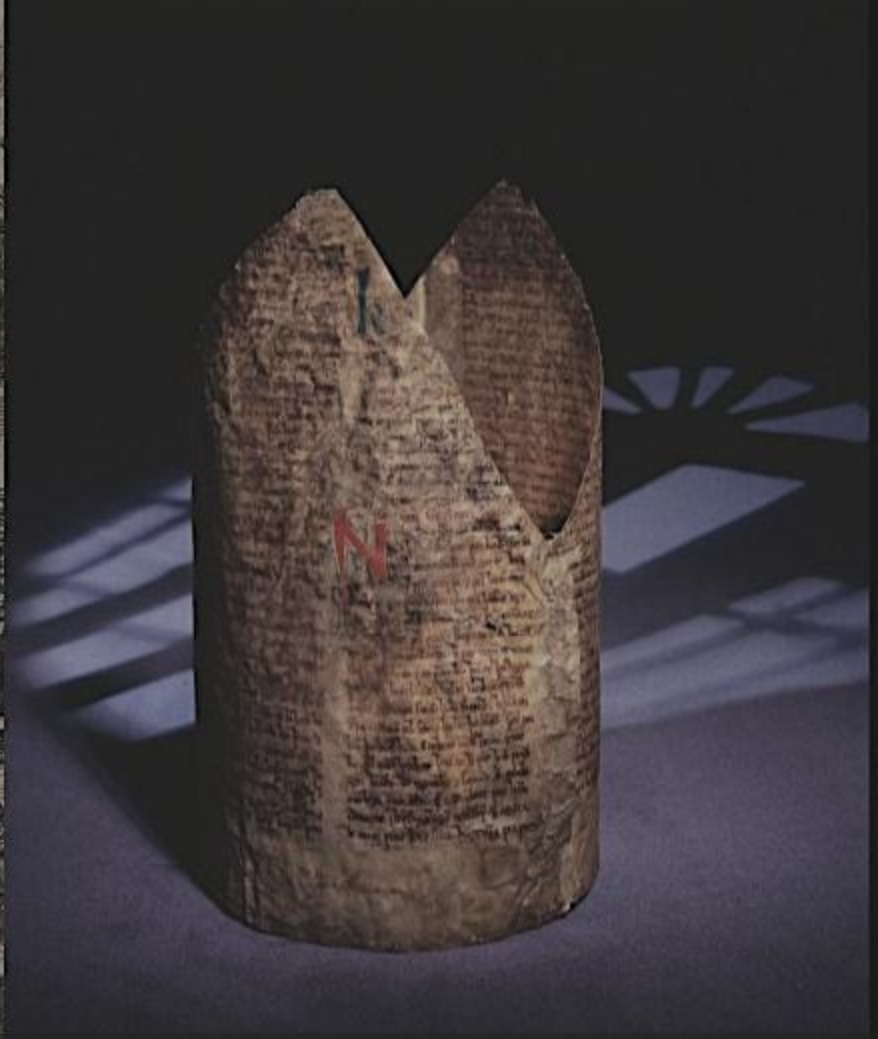
(what you must accept For this trick to “work”)

number of works ~ number of species

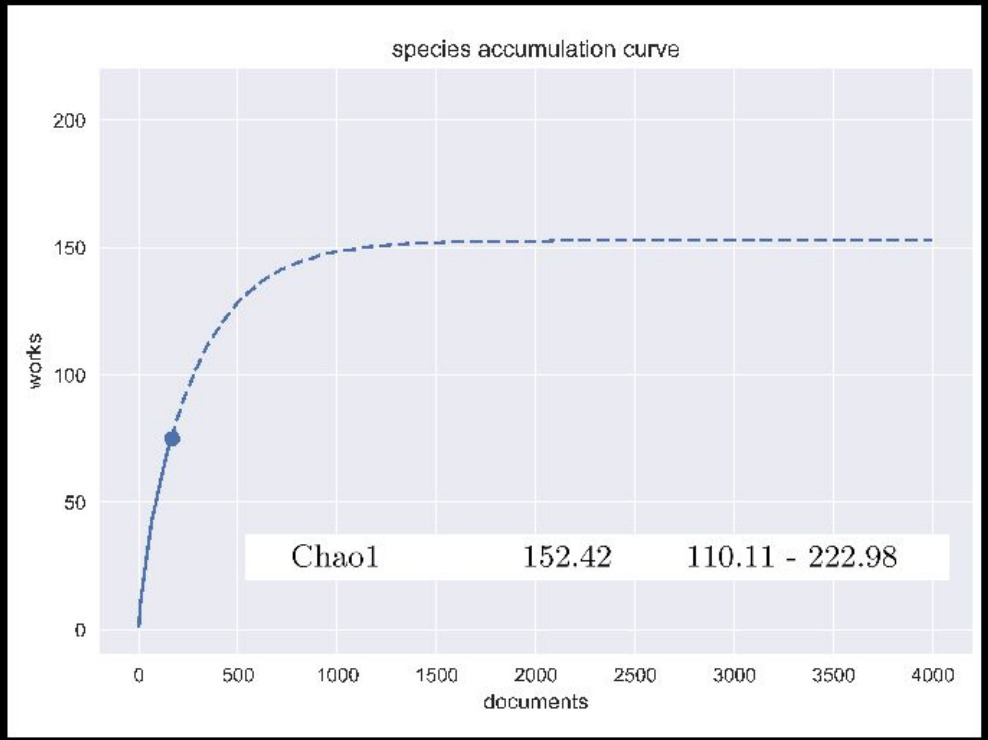
number of documents ~ number of sightings

number of traps ~ number of libraries

[cf. Egghe & Proot 2007]



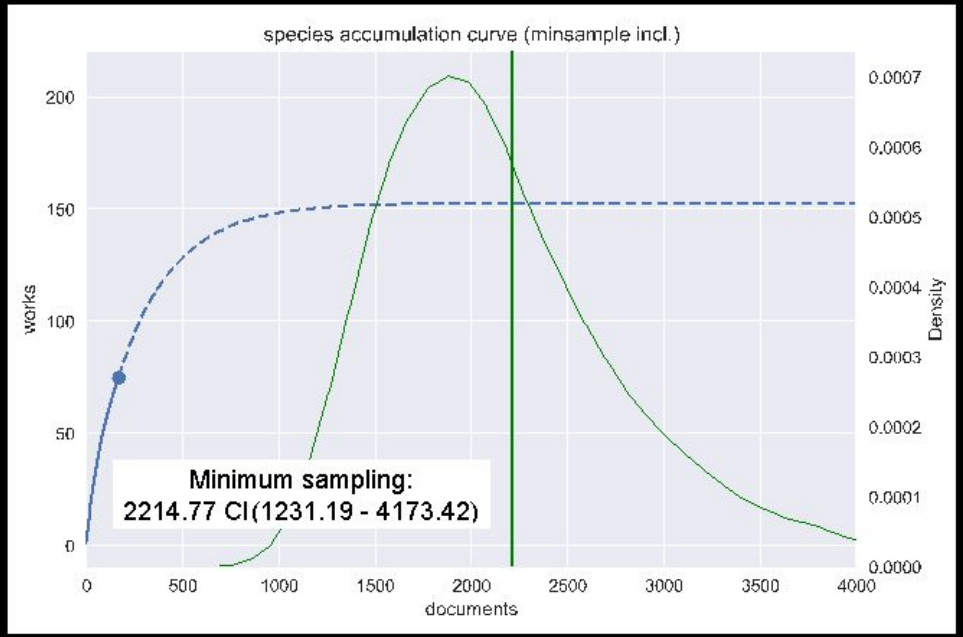
Elisabeth de Bruijn, Mike Kestemont, Katarzyna Anna Kapitan, Unseen Species Modelling Applied to Medieval Literature [Workshop], Paris, 10 June 2025.



Species accumulation curve

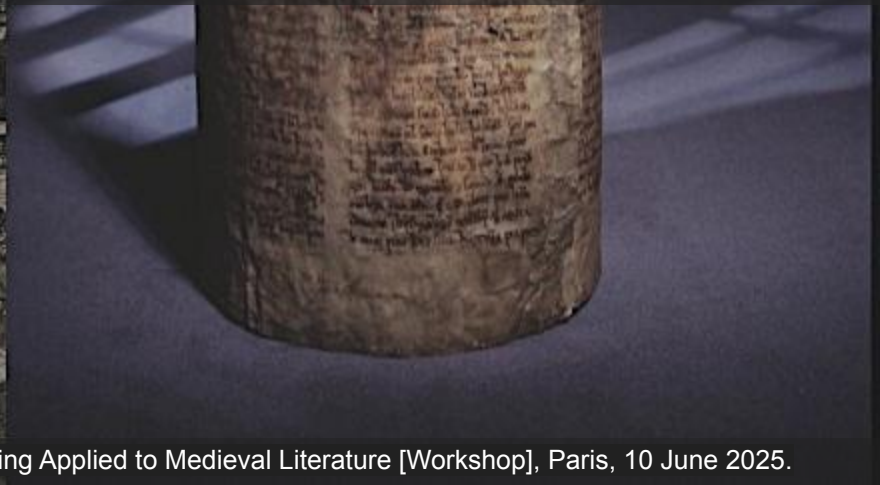
- (At least) half of texts “lost”
- Wide confidence interval (CI)…
- But >>> than (conservative) 100
- Correction of optimism

Document Survival



- Estimate saturation point blue curve
- Wide CI with left-skewed distribution
- Estimate: 7.54% survival rate (~167/2215 MSS)
- Amazingly close to 7% (Wijsman et al.)

Results and Looking Ahead



Differences Between Vernaculars

Table 1. Point estimates of survival ratios in six traditions. For works using Chao1 (i.e., sample completeness at $q = 0$) and documents (ms) using the minimum sampling extension, including the number of works (S_{obs}), documents (n), singletons (f_1), and doubletons (f_2).

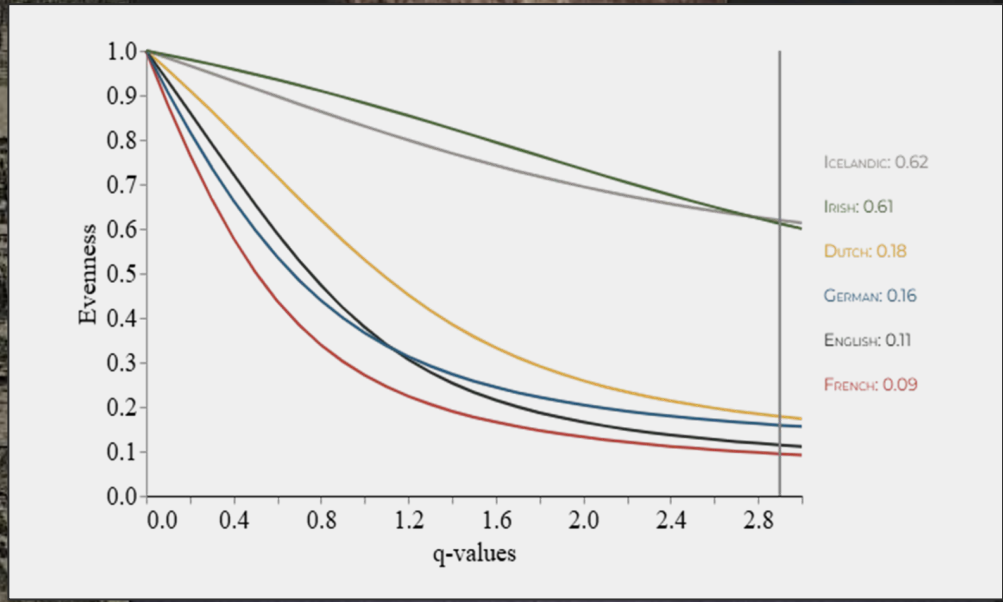
Language	f_1	f_2	S_{obs}	n	Chao1	ms
Dutch	45	13	75	167	0.492	0.075
English	42	8	69	176	0.386	0.049
French	90	21	222	1473	0.535	0.054
German	36	19	128	1088	0.790	0.145
Icelandic	44	28	117	295	0.773	0.169
Irish	69	54	188	449	0.810	0.192
Total	326	143	799	3648	0.683	0.090

Differences with confidence intervals

How can these differences be explained?

- In the limited space of the paper, we suggested one explanation, borrowing again a concept from ecology:

  Evenness



How can these differences be explained?

Indeed, species richness is higher in ‘somewhat remote’ islands (Iceland and Ireland, but not England)

BUT

the concept of evenness in relation to ‘remoteness’ does NOT explain the high survival rate of German

My (and Katarzyna's) not so secret agenda: Physicality and Co-transmission

- Aim: Enrich the *Forgotten Books* dataset with information about physicality & co-transmission
- Purpose: **reception research across vernaculars**
- Question: How to quantify potential factors of influence

How do we quantify potential factors of influence

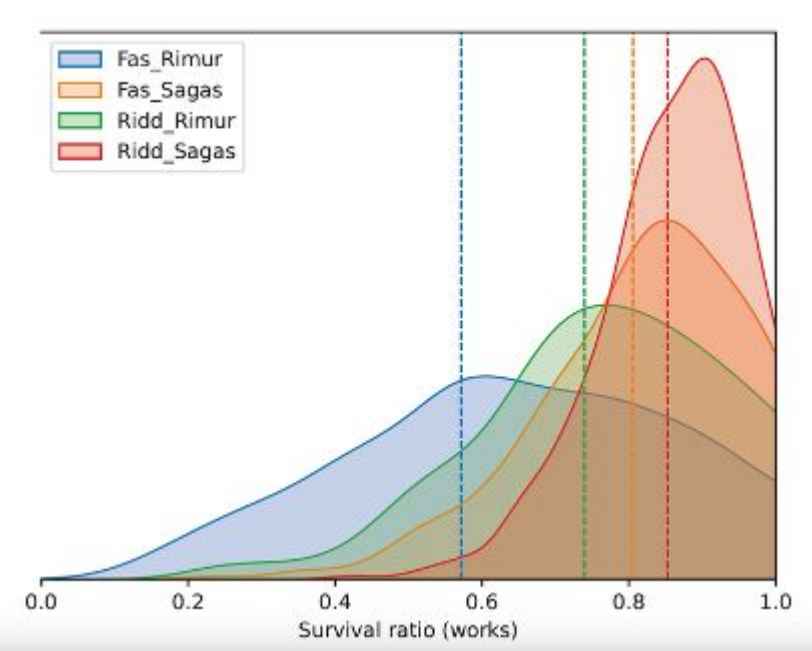
Reflections on:

1. (Sub)genre (work-related)
2. Dating (both work and object-related)
3. Material characteristics, e.g. illuminations, support (object related)
4. Multi-text codices (object related)

1. (Sub)genre

- Research bias in data collection: **we** decide what to include
- No consensus between vernaculars on ‘romance’
- Differences between **subgenres** affecting data?
  - *Romane – Heldenepen*
  - *roman – chansons de geste*
  - *Arthurromans – Karelromans – ‘other’*
  - *riddarasögur – fornaldarsögur – rímur*

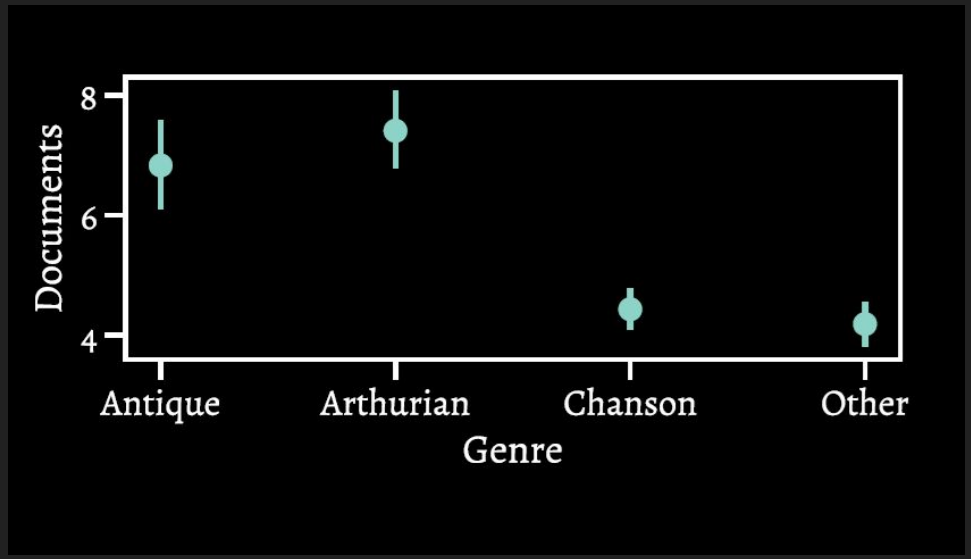
Icelandic Works by Subgenre



- Fas_Rimur (Fornaldarrímur)
  - 57% survived (43% lost) !!!
- Fas_Sagas (Fornaldarsögur)
  - 80 % survived (20% lost)
- Ridd_Rimur (Riddararímur)
  - 74 % survived (26% lost)
- Ridd_Sagas (Riddarasögur)
  - 85% survived (15% lost)

Source: Katarzyna Anna Kapitan, 'Digital Approaches to Loss of Medieval Literature', a lecture delivered in the lecture series *Insights into the Digital Humanities,* at the University of Bern, 28 October 2024. DOI: 10.5281/zenodo.14182601

Popularity of (sub)genre?



Source: Mike Kestemont, 'Modelling the popularity of medieval literature. The case of the chivalric narrative. Lecture: Paris 2024.

2. Dating

- The ‘Middle Ages’ didn’t start and end everywhere at the same time
  - Production of works and documents starts earlier on the Continent
  - Late circulation of medieval literature on islands:
    - Cut-off date Icelandic literature: ca. 1550
    - Cut-off date Irish literature: 1600

→ consequences?

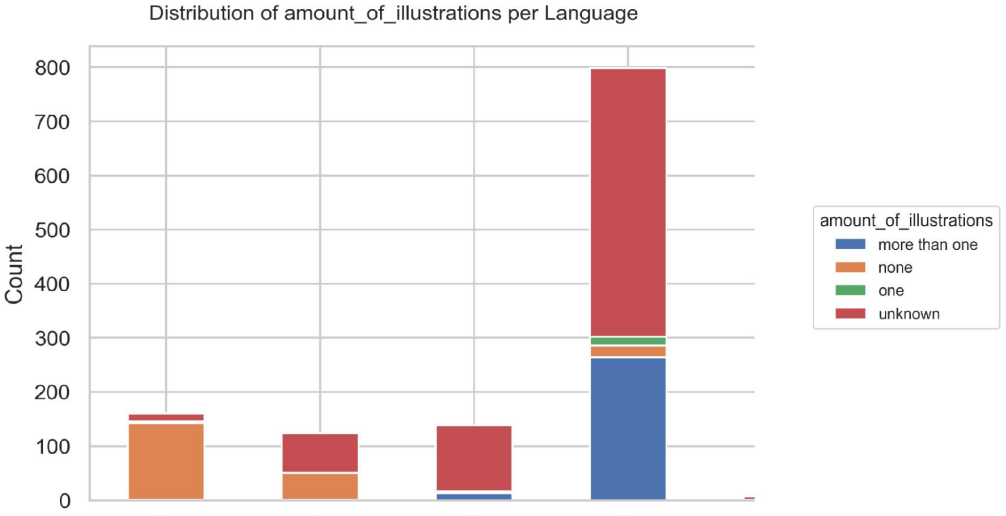
2. Dating

- The count changes significantly depending on the cut-off date.
- In our Icelandic dataset, the story of Konráður is registered in prose and verse as preserved in 7 witnesses (4 of the saga, and 3 of the rímur), but if we narrowed down a cut-off date for the witnesses to *before 1500* we'd have only 3 witnesses (2 of the saga, and 1 of the rímur)
- How to quantify to stay faithful to disciplinary traditions, and enable transdisciplinary comparisons?

3. Material characteristics

- **Illustrations**
  - Hypothesis: Illustrated manuscripts have higher survival chances (e.g. H. Wijsman, *Luxury Bound. Illustrated Manuscript Production and Noble and Princely Book Ownership in the Burgundian Netherlands (1400-1550)* (Brepols, 2010).
  - Is this true for all linguistic traditions?

3. Material characteristics



Contents of the LostMA database documents by 'amount of illustrations' for French, Dutch and English (5/06/2025); see: https://lostma-erc.github.io/corpus

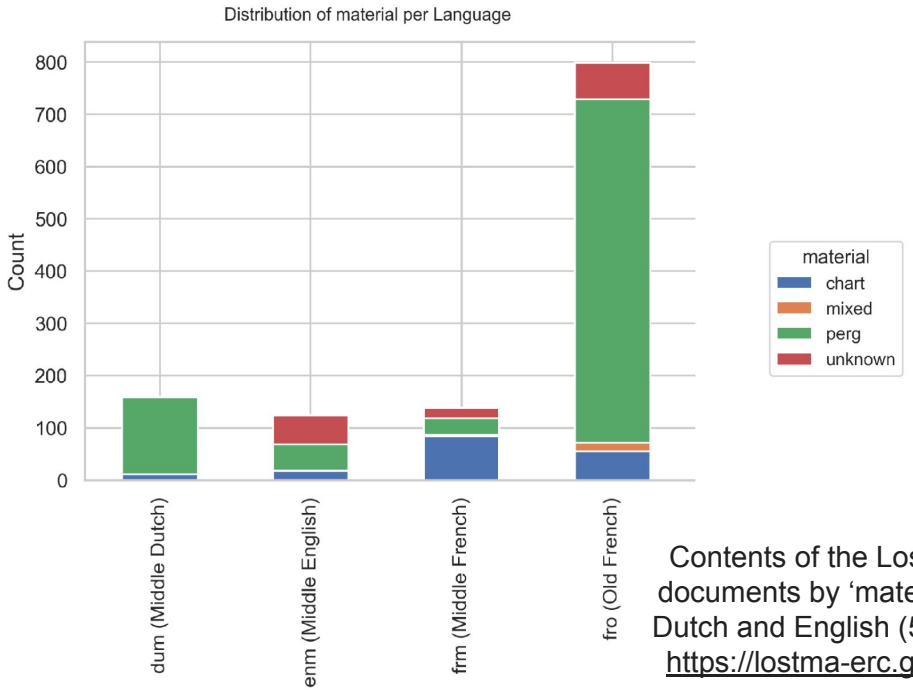
3. Material characteristics

- **Parchment** is reusable (yet parchment manuscripts are generally older) > how does this affect survival rate?

Figure 6. BodL, MS Ashmole 33 outer covers.

Image source: Hannah Ryley, *Re-using Manuscripts in Late Medieval England* (2022), p. 87.

3. Material characteristics



Contents of the LostMA database documents by 'material' for French, Dutch and English (5/06/2025); see: https://lostma-erc.github.io/corpus

4. Multi-text codices

We counted texts and documents, but:

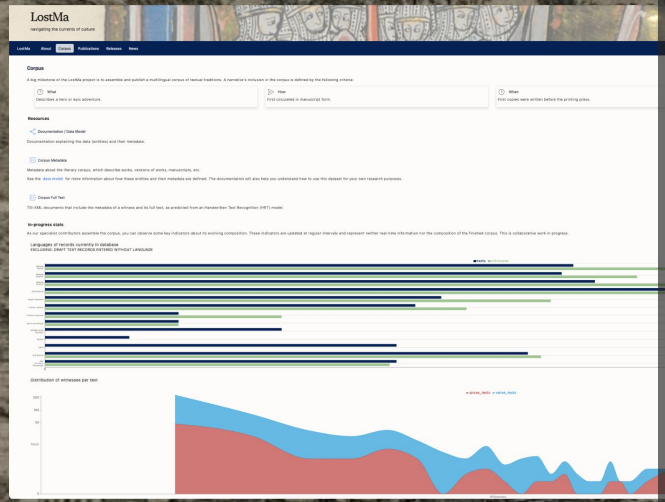
- multiple shelfmarks > 1 witness (manuscript scattered across depositories) - taken into account in FB
- 1 shelfmark > multiple texts (multi-text codices) - not in FB
- Impact of multi-text codices on loss/survival?
  a. If you lose **one** book, you lose **multiple** texts
  b. Texts are being ‘protected’ by the collection they are part of

→ b is believed to be the case, however…

4. Multi-text codices

- My pilot study revealed that 20% of the German romances were transmitted in a collection with other romances, against 40% of the French and 50% of the English romances.
- Yet, German is best preserved, English the worst…
- Note that we only counted codices with 2 or more ROMANCES, not multi-text codices in general, because we do not have those data yet

Hope for a more nuanced view?
Ongoing collaboration with LostMA



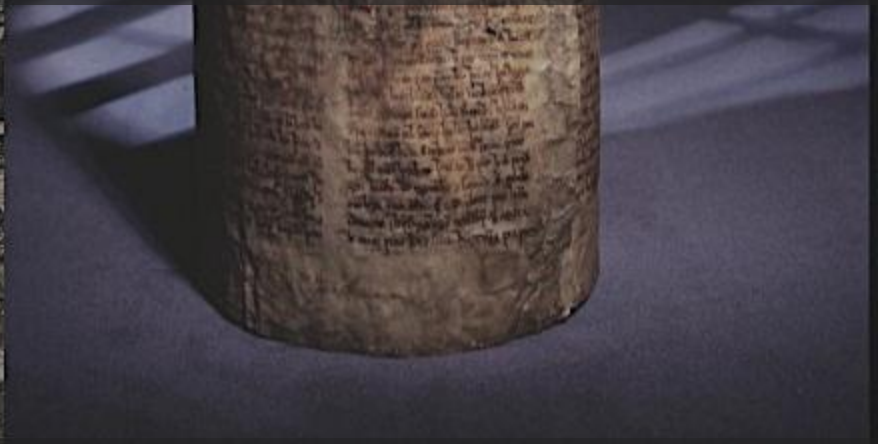
Screenshot of the LostMA website (5/06/2025), https://lostma-erc.github.io/corpus

- Creation of the revised datasets for Dutch & Old Norse.
- Suggestions for the data model, emphasising:
  - **physical features of manuscripts** (Implemented: Christensen & Camps, Greening your database of literary works https://enc.hal.science/hal-05059049v1)
  - **co-transmission of texts** (ongoing discussions within the team of collaborators)

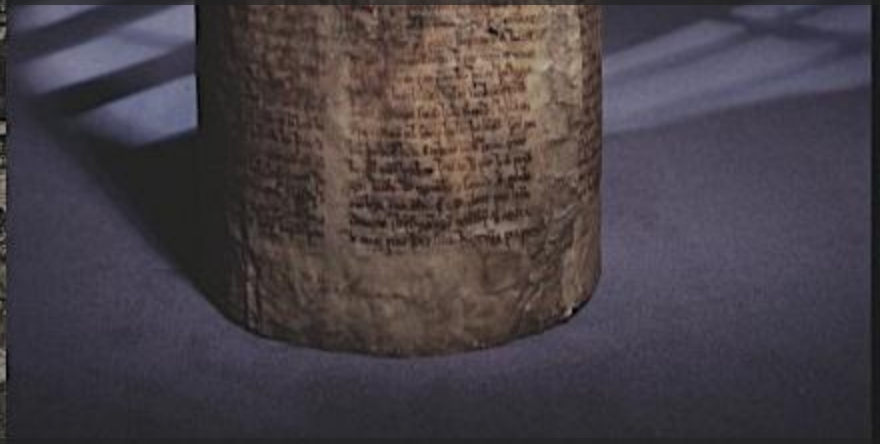
How to include multi-text codices in a database?

1. Co-transmission	Unknown	No (single-text codex)	Yes (multi-text codec)
If 1 yes:			
2. With other romances	Unknown	No	Yes
2.1 Extent of romances in the codex (page count?, or number of texts?)	100%	>75%	< 30%
If 2 no, which genres (multiple choice)	historiography	philosophy	theology, etc.
If 2 yes: which other genres (multiple choice)	historiography	philosophy	theology, etc.

Ideas? Questions?



Hands-on Session



Reproducible Research

Hands-on Session with Copia

- Colab notebook: https://tinyurl.com/CopiaParis2025
- Workshop repo: https://github.com/KAKDH/UnseenSpeciesCopia-

Thank You

With huge thanks to the Forgotten Books team and the LostMA team so all inspiring discussions.

The background image of the slides: Left: Fire of Stockholm, 1697, Source: https://www.wikiwand.com/en/Historical_fires_of_Stockholm#Media/File:Slottsbranden_1697.jpg; Right: Parchment leaf reused as bishop's mitre., Arnamagnæan Institute, Copenhagen. Source: https://twitter.com/arnamagnaean/status/1070357850013294592/photo/1