Forgotten books
The application of unseen species models to the survival of culture
Driscoll, Matthew James; Kestemont, Mike; Karsdorp, Folgert; de Bruijn, Elisabeth; Kapitan, Katarzyna Anna; Ó Macháin, Pádraig; Sawyer, Daniel; Sleiderink, Remco; Chao, Anne

Published in:
Science

DOI:
10.1126/science.abl7655

Publication date:
2022

Document version
Peer reviewed version

Citation for published version (APA):
Title: Forgotten Books: The Application of Unseen Species Models to the Survival of Culture

Authors: Mike Kestemont,† Folgert Karsdorp,† Elisabeth de Bruijn, Matthew Driscoll, Katarzyna A. Kapitan, Pádraig Ó Macháín, Daniel Sawyer, Remco Sleiderink, Anne Chao

Affiliations:

University of Antwerp; Antwerp, Belgium
KNAW Meertens Institute; Amsterdam, the Netherlands
Ruhr-Universität Bochum; Buchum, Germany
Arnamagnæan Institute; University of Copenhagen, Copenhagen, Denmark
Linacre College University of Oxford; Oxford, United Kingdom
Department of Nordic Studies and Linguistics; University of Copenhagen, Copenhagen, Denmark
Vigdís Finnbogadóttir Institute of Foreign Languages, University of Iceland; Iceland
The National Museum of Iceland; Iceland
The Museum of National History, Frederiksborg Castle; Denmark
University College Cork; Cork, Ireland
Merton College, University of Oxford; Oxford, United Kingdom
Institute of Statistics, National Tsing Hua University; Hsin-Chu, Taiwan

Corresponding author. E-mail: mike.kestemont@uantwerp.be

†These authors contributed equally to this work.

Abstract: The study of ancient cultures is hindered by the incomplete survival of material artefacts, so that we commonly under-estimate the diversity of the cultural production in
historic societies. To correct for this survivorship bias, we apply unseen species models
from ecology and gauge the loss of narratives from medieval Europe, such as the romances
about King Arthur. The obtained estimates are compatible with the scant historic evidence.
Besides events like library fires, we identify the original evenness of populations as an
overlooked factor in their stability in the face of immaterial loss. We link the elevated
evenness in island literatures to parallel accounts of ecological and cultural diversity in
insular communities. Our analyses call for a wider application of these methods across the
heritage sciences.

One-sentence summary: Unseen species models from ecology can estimate artifact survival
rates from ancient cultures.

Main Text: Historical studies of human culture are hindered by the fact that they must work
with incomplete samples of material artefacts (books, paintings, statues, etc.) that still survive (1,
2), and that do not necessarily represent the original population faithfully. Because of this
survivorship bias, we risk to under-estimate the diversity of the cultural production of past
societies. In response, we turn to bias correction methods from ecology. For monitoring species
richness reliably, ecologists use statistical models that account for the unseen species in samples
(3). This is necessitated by the common under-detection of species that are hard to observe
during bioregistration campaigns, creating a detection bias that must be quantitatively accounted
for. Following recent studies (4, 5) pointing to parallels between cultural and ecological
diversity, we show that unseen species models can be applied to manuscripts containing
medieval literature. This enables us to estimate the size of the original population of works and
documents and, in turn, the losses that these cultural domains sustained. We offer a large-scale
estimate of the (im)material loss of narrative fiction from medieval Europe. This endeavour
resonates with a broader interest in the persistence of cultural information in human societies, particularly in the domain of cultural evolution (5–9).

**Fig 1.** Top left (A): Fragment of *Strengleikar* (COPENHAGEN, DEN ARNAMAGNÆANSKE SAMLING, AM 666 B 4TO), repurposed to stiffen a bishop’s miter. Used with permission. Top right (B): Intact, lavishly illustrated codex (*Wigalois*; LEIDEN, UNIVERSITY LIBRARY, LTK. 537, F. 72v). CC-BY. Bottom (C): Fragment (binding waste) of an unidentified Dutch romance (KU LEUVEN LIBRARIES, SPECIAL COLLECTIONS, Ms. 1488). Public domain.

Narrative fiction was a mainstay of medieval culture (ca. 600–1450 CE). The courtly chivalric romance, for instance about the King Arthur and the Holy Grail, has had a long-lasting impact. Before movable-type printing in Europe (ca. 1450 CE), handwritten manuscripts were used for the sustainable storage of text (10). In some places – e.g., Ireland and Iceland – manuscript circulation continued in this role into the modern era. Works of narrative fiction circulated through manually produced copies that survive as unique material artefacts, typically in the form of parchment (or paper) codices (11). Thus, multiple parallel witnesses of the same medieval work could circulate. Today, manuscripts constitute the main evidence regarding medieval narrative fiction. Textual witnesses have been subject to various processes of decay and destruction (e.g., library fires) (1, 2, 11, 12). Texts may survive in intact codices (Fig. 1B), but many of those works which survive at all now only exist in manuscripts that are fragmentary, lacking leaves or bearing damage from tearing. Because of parchment’s durability, books were often recycled for more everyday practical uses (Fig. 1A), e.g. into small boxes, tailors’ measures or even packing material for meat. Additionally, strips of parchment were used as binding waste by printers to strengthen book spines (Fig. 1C).
The (material) loss of documents can entail the (immaterial) loss of works: a work becomes “lost” when none of the copies that once represented it are known anymore (13). A theoretical distinction must separate out documents that have been destroyed and those which have not been recovered yet, e.g. because of improper cataloguing: sources in the latter category might still reemerge. Different survival scenarios are represented in Fig. 2. We adopt a distinction between the (non-material) WORK, as listed in pre-existing scholarly repertories, and the (material) DOCUMENTS in which these WORKS are attested (14). While medieval narratives also circulated orally, the present analysis is necessarily limited to the written production.

Fig 2. Schematic representation of example survival scenarios (A–G) for medieval literature. Individual WORKS were copied into one (A–E) or more (F–G) DOCUMENTS, whose survival STATUS varies from intact codices (A) to fragments (C, E), residing in REPOSITORIES, such as libraries, archives or private collections. Lost DOCUMENTS can be fully (D) or partly (G) destroyed, or may not have been recovered yet (B). For lost WORKS (B, D), none of the original documents have been recovered.

The survival rates for medieval DOCUMENTS are traditionally estimated based on medieval library catalogues: if the listed specimens can still be identified, the calculation of the survival rates of these books is straightforward (1). Authoritative studies have suggested (for the Holy Roman Empire) an overall survival rate of 7% for general-purpose manuscripts, which must be adjusted upwards to 20% for higher-end codices (1, 11, 15). Such estimates are nevertheless problematic because they depend on a small sample of catalogues, from protective collection environments, with cataloguers frequently omitting lower-end documents (15). A prior attempt (16) to apply methods from survival studies to this problem met with criticism, because the figures obtained did not fit with other historical evidence (17, 18). Regarding the loss of WORKS, there has been
little quantitative work (19). Conventional approaches rely on allusions to lost works, e.g., in library catalogues (13), but many lost works will not have been mentioned. Egghe and Proot published a pioneering estimator for the loss of multi-copy, printed works (20), which was later identified as an unseen species model. Their approach, however, requires an estimate of the print-runs of hand-pressed books, which does not suit manuscripts.

We build on the information-theoretic analogy that medieval works can be treated as distinct species in ecology, and that the number of extant documents for each work can be regarded as analogous to the number of sightings for an individual species in a sample. Thus, if we treat the available count information for medieval literature as “abundance data” (3), one can apply unseen species models to estimate the number of lost works in a corpus or “assemblage”. We collected count data for surviving medieval heroic and chivalric fiction in six European vernaculars (21): three insular (Irish, Icelandic, English) and three continental (Dutch, French, German). For all works, we listed the number of handwritten medieval documents in which they survive (Table 1). Next, we applied non-parametric methods to estimate the original richness of these traditions. For a given assemblage, let \( N \) represent the abundance-based frequencies for \( n \) unique works which were observed in documents.

Chao1 is a method to estimate a lower bound on \( N \), or the number of undetected species in an assemblage, based on the number of singletons \((n_1, \text{species sighted only once})\) and doubletons \((n_2, \text{species sighted exactly twice})\) in a sample of \( n \) individuals. The original number of works \( N \) can then be estimated as \( n_1 + 2n_2 \). Chao1 is not specific to ecology and has been derived under a very general model: it can be applied as a universally valid lower-bound richness estimator to any hyper-diverse, under-sampled collection of types, such as stone tools, coins or even words (23). Therefore, this estimator is even more widely applicable in the heritage sciences than shown here.
In this framework, the survival ratio for the works can be quantified as the *sample completeness* or: the ratio of the number of unique observed works (\( N \)) over the estimated true species abundance (\( S \)). Species richness is an intuitive measure to quantify species diversity, but alternative measures exist, e.g. the Shannon or Simpson diversity (both put less weight on rare species). The Hill number profile (26) allows us to compare a sample’s diversity across various values of \( z \), a scalar corresponding to different diversity measures at specific points (e.g., for richness, for Shannon, for Simpson). Hill numbers are nowadays the diversity measure of choice in ecology for quantifying species diversity and decomposition (25).

We also use an extension of Chao1 (27) that estimates the minimum number of additional observations that are required to observe each of the species at least once. This number will approximate the number of lost documents in an assemblage, so that we can estimate the original population size as \( N \). Chao1 and the minimum sampling extension were derived as a lower bound, which implies that the estimates of the survival ratios below, strictly speaking, offer an upper bound on the loss of works and documents – it is possible that even more literature was lost. Nevertheless, Chao1 works satisfactorily as a nearly unbiased point estimator when the abundances of rare species are nearly homogeneous or singletons and undetected species have approximately the same mean abundances (23). Because Chao1 is non-parametric, the lower bound is valid for any distribution of entities among types: it should be robust to differences in survival across document types (15).

Finally, we analyzed the evenness in these assemblages or the extent of equity among species abundances (28). A community’s evenness will affect its stability in the face of external forcing, in particular its ability to withstand the impact of diversity-threatening events, such as wildfires (29). Given two equal-sized assemblages, the more even assemblage will be more resistant to the
loss of WORKS through DOCUMENT losses. Below, we chart evenness profiles for one class () of evenness measures (Fig. 5). These curves can be connected to the slope of a Hill number profile: their steepness enables the intuitive comparison of the (un)evenness in the WORKS’ abundances for the reconstructed assemblages. The profiles (Fig. S1) for additional evenness classes () yield consistent findings (21).

**Fig 3.** Estimates for the union of the 6 assemblages. A: Hill number curves (for ), empirical and estimated, showing the absolute underestimation of the original diversity of WORKS. B: Species accumulation curve, plotting the number of WORKS as a function of the number of DOCUMENTS. The dot shows the observable data, the solid line the rarefaction for sample sizes , the dashed line the extrapolation to sample sizes . C: Kernel-density plot for the estimated number of DOCUMENTS.

The results for the union of the corpora (Table 1 and S2) suggest an overall survival ratio of 68.3% CI[63.2%–73.5%] for WORKS and 9.0% CI[7.5%–10.7%] for DOCUMENTS. The species accumulation curve (Fig. 3B) indicates at which rate we might still be discovering new WORKS in the future, by sighting more DOCUMENTS (3). Fig. 3A shows the empirical and estimated Hill number profiles: at the curves indicate the absolute size of our current under-estimation of the original diversity in the combined assemblage of chivalric and heroic narratives from the medieval period. Of the original 1,170 WORKS that once would have existed, 799 would survive nowadays. Likewise, the 3,648 DOCUMENTS that are still observable constitute a sample from a population that originally would have counted 40,614 specimens (Fig. 3C).

We observe considerable inter-vernacular variation (Table 1), ranging from the relatively poorly surviving English WORKS (38.6%) to the relatively intact German tradition (79.0%). Although Dutch and French have a substantially lower survival factor, two of the insular assemblages,
Icelandic and Irish, have sustained similar losses to German, with point estimates of 77.3% / 81.0% and 16.9% / 19.2% for the survival of works and documents respectively (12). Puzzling is that Old and Middle English documents did not travel far during their post-medieval afterlives (Fig. 4), while other literatures survive in a wide manuscript diaspora. The survival estimates for works and documents yield similar rankings (Table 1). In the SM, we compare Chao1 to 3 other estimators with similar results (Fig. S2). Fig. 5, finally, shows the (estimated) evenness profiles and offers further insight into the distributional properties characterizing the assemblages. Here too, we note the atypical nature of Icelandic and Irish: in comparison to the highly uneven distribution of e.g. French, these two insular literatures feature a much more even distribution of documents over works.

**Fig 4.** Heatmap of the geolocations of the repositories where documents are kept for 4 vernaculars. Made with Leaflet 1.7.1.

**Fig 5.** Normalized evenness profiles () for the six individual vernaculars, plotting as a function of order . The values on the Y-axis reflect the estimated evenness in the reconstructed assemblages.

<table>
<thead>
<tr>
<th>language</th>
<th>Chao1</th>
<th>MS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dutch</td>
<td>0.492</td>
<td>0.075</td>
</tr>
<tr>
<td>English</td>
<td>0.386</td>
<td>0.049</td>
</tr>
<tr>
<td>French</td>
<td>0.535</td>
<td>0.054</td>
</tr>
<tr>
<td>German</td>
<td>0.790</td>
<td>0.145</td>
</tr>
<tr>
<td>Icelandic</td>
<td>0.773</td>
<td>0.169</td>
</tr>
<tr>
<td>Irish</td>
<td>0.810</td>
<td>0.192</td>
</tr>
<tr>
<td>union</td>
<td>0.683</td>
<td>0.090</td>
</tr>
</tbody>
</table>
Table 1. Point estimates of survival ratios in 6 traditions: for works, using Chao1 (i.e., sample completeness at) and documents (MS) using the minimum sampling extension, including the number of works (), documents (), singletons () and doubletons ()

Regarding documents, our results confirm the severity of the losses, with estimates ranging from 4.9% (English) to 19.2% (Irish). This corroborates previous estimates from book history, positing an overall survival factor of 7%, i.e. slightly lower than our point estimate for the union (9.0% CI[7.5%–10.7%]). Contrary to previous analyses (16, 17), these results are therefore compatible with the historical evidence from book history. It remains to be seen whether these estimates will scale to other cultural domains, but this analysis reveals important relative differences in the persistence of medieval heroic and chivalric narrative across Europe. Some of these differences have not been noticed before and challenge existing assumptions. For example, our results suggest that Irish and Icelandic literature have been preserved comparatively well compared to some of the more canonical mainland literatures (12).

In ecology, island ecosystems stand out: despite being comparatively species-poor for their land surface, they feature a higher endemic species richness compared to mainland regions (30). Additionally, insular assemblages demonstrate a higher species evenness, due to the lack of predators etc. A parallel emerges with some of the cultural diversity profiles for island regions reconstructed here: if land-isolated areas preserve biological heritage more effectively, the same might hold true for cultural heritage. Previous discussions about the survival of historic literature have focused on factors such as library fires or collectors’ interests (1). We identify an additional key aspect that is typically overlooked: the evenness with which documents were originally distributed over works fundamentally affected an assemblage’s stability (29). Medieval French literature, for instance, was sizable, but its long tail of low-abundance works rendered it more
susceptible to immaterial loss. Thus, while the loss figures for Icelandic and Irish are considerable, their distributional characteristics likely made them more robust to post-medieval losses.

Which societies produce a highly even cultural output to safeguard the retention of their diversity? The role of demography, especially population size, has been hotly debated in cultural evolution (6, 7, 31). Smaller, isolated social groups can be more susceptible to the random loss of cultural traits because of stochastic drift (6), although these communities can adopt fitness-improving behavior to guard against such information loss. The topology of social networks seems crucial: a low degree (or interconnectedness between individuals) can counter the impact of drift and promote the retention of cultural complexity (32). For the remote island of Rapa Nui, a model-based account showed how structural constraints in social interactions might have stimulated the retention of diversity (8). We have extended these simulations (21) to show that a lower network degree, under neutral models of transmission, invariably leads to a more evenly distributed cultural production (Fig. S3).

References and Notes

1. E. Buringh, Medieval manuscript production in the latin west, explorations with a global database (Brill, 2011).


21. Materials and methods are available as supplementary materials.


Acknowledgments: We would like to thank Dirk Schoenaers, Georgia Henley, Jean-Baptiste Camps, Bernd Bastert, Daniel Könitz, and Jeroen Deploige for their help, as well as the five anonymous referees.

Funding: EDB’s work was funded through a postdoctoral fellowship from FWO Flanders; KAK’s contribution was funded a Carlsberg Foundation Visiting Fellowship CF20-225 and a Carlsberg Foundation H. M. Queen Margrethe II Distinguished Research Fellowship CF18-500).

Author contributions: Conceptualization: MK, FK, EDB, MD, KAK, POM, DS, RS, AC. Data curation: MK, FK, EDB, MD, KAK, POM, DS, RS, AC. Formal analysis: MK, FK. Investigation: MK, FK, EDB, MD, KAK, POM, DS, RS, AC. Methodology: MK, FK, AC. Software: MK, FK. Validation: MK, FK, EDB, MD, KAK, POM, DS, RS, AC. Visualization: MK, FK. Writing – original draft: MK, FK, EDB, MD, KAK, POM, DS, RS, AC. Writing – review & editing: MK, FK, EDB, MD, KAK, POM, DS, RS, AC.

Competing interests: Authors declare that they have no competing interests.

Data and materials availability: In addition to the publication of the underlying data and code in an open access repository, this paper is released with a Python software package (COPIA, available from PyPI), all under a CC-BY-SA licence, to replicate our findings and stimulate the adoption of this approach in other domains.

Supplementary materials

Materials and Methods