CULTURAL HERITAGE

Forgotten books: The application of unseen species models to the survival of culture

Mike Kestemont[1]*†, Folgert Karsdorp[2]†, Elisabeth de Bruijn[1,3], Matthew Driscoll[4], Katarzyna A. Kapitan[5,6,7,8,9], Pádraig Ó Macháin[10], Daniel Sawyer[11], Remco Sleiderink[1], Anne Chao[12]

The study of ancient cultures is hindered by the incomplete survival of material artifacts, so we commonly underestimate the diversity of cultural production in historic societies. To correct this survivorship bias, we applied unseen species models from ecology to gauge the loss of narratives from medieval Europe, such as the romances about King Arthur. The estimates obtained are compatible with the scant historic evidence. In addition to events such as library fires, we identified the original evenness of cultural populations as an overlooked factor in these assemblages' stability in the face of immaterial loss. We link the elevated evenness in island literatures to analogous accounts of ecological and cultural diversity in insular communities. These analyses call for a wider application of these methods across the heritage sciences.

Historical studies of human culture are hindered by the fact that they must work with incomplete samples of material artifacts (books, paintings, statues, etc.) that still survive (*1*, *2*) but do not necessarily represent the original population faithfully. Because of this survivorship bias, we risk underestimating the diversity of the cultural production of past societies. In response to this risk, 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 underdetection of species that are difficult to observe during bioregistration campaigns, creating a detection bias that must be accounted for quantitatively. Following recent studies (*4*, *5*) pointing to parallels between cultural and ecological diversity, we show that unseen species models can be applied to manuscripts preserving 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 endeavor resonates with a broader interest in the persistence of cultural information in human societies, particularly in the domain of cultural evolution (*5*–*9*).

Narrative fiction was a mainstay of medieval culture (~600 to 1450 CE). The courtly chivalric romances concerning King Arthur

[1]University of Antwerp, Antwerp, Belgium. [2]KNAW Meertens Institute, Amsterdam, the Netherlands. [3]Ruhr-Universität Bochum, Bochum, Germany. [4]Arnamagnæan Institute, University of Copenhagen, Copenhagen, Denmark. [5]Linacre College, University of Oxford, Oxford, UK. [6]Department of Nordic Studies and Linguistics, University of Copenhagen, Copenhagen, Denmark. [7]Vigdís Finnbogadóttir Institute of Foreign Languages, University of Iceland, Reykjavík, Iceland. [8]The National Museum of Iceland, Reykjavík, Iceland. [9]The Museum of National History, Frederiksborg Castle, Hillerød, Denmark. [10]University College Cork, Cork, Ireland. [11]Merton College, University of Oxford, Oxford, UK. [12]Institute of Statistics, National Tsing Hua University, Hsin-Chu, Taiwan.
***Corresponding author. Email: mike.kestemont@uantwerp.be**
†These authors contributed equally to this work.

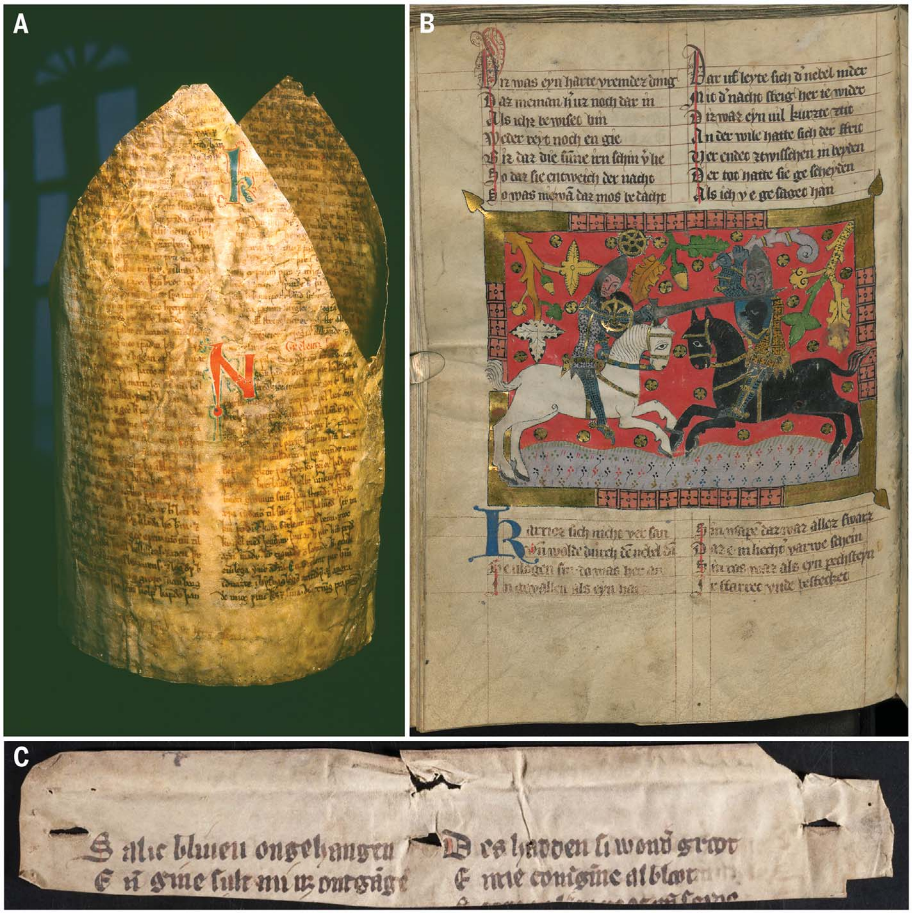


Fig. 1. Narrative fiction survives in a diverse range of medieval text carriers. (**A**) Fragment of *Strengleikar* repurposed to stiffen a bishop's miter (Copenhagen, Denmark, Arnamagnæanske Samling, AM 666 b 4to; used with permission). (**B**) Intact, lavishly illustrated codex (*Wigalois*; Leiden, University Library, Ltk. 537, f. 72v, CC-BY). (**C**) Fragment (binding waste) of an unidentified Dutch romance (KU Leuven Libraries, Special Collections, manuscript no. 1488; public domain).

and the Knights of the Round Table, for example, have had a long-lasting impact. Before movable-type printing in Europe (~1450 CE), handwritten documents (manuscripts) were used for the sustainable storage of text (*10*). In some places, such as 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 artifacts, typically in the form of parchment or, later, 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 that survive at all now only exist in manuscripts that are fragmentary, lacking leaves or bearing damage from tearing, insects, overuse, etc. Because of parchment's durability, books were often recycled for more everyday practical uses (Fig. 1A) such as small boxes or used as tailors' measures or even packing material for meat. Additionally, strips of parchment were frequently used by binders 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 preserved it is known to have survived (*13*). A theoretical distinction must be made between documents that have been destroyed and those that have not been recovered yet, for example, because of inadequate cataloging; sources in the latter category might still reemerge. Different survival scenarios are represented in Fig. 2. We adopt a distinction between the (nonmaterial) work as listed in preexisting scholarly repertories and the (material) documents in which these works are attested (*14*). Although medieval narratives also circulated orally, the present analysis is necessarily limited to written production.

The survival rates for medieval documents are traditionally estimated based on medieval library catalogs: If the listed specimens can still be identified, then 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 upward to ~20% for higher-end codices (*1*, *11*, *15*). Such estimates are nevertheless problematic because they depend on a small sample of catalogs from protected collection environments, with catalogers 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, for example, in library catalogs (*13*), but many lost works will not have been mentioned. Egghe and Proot

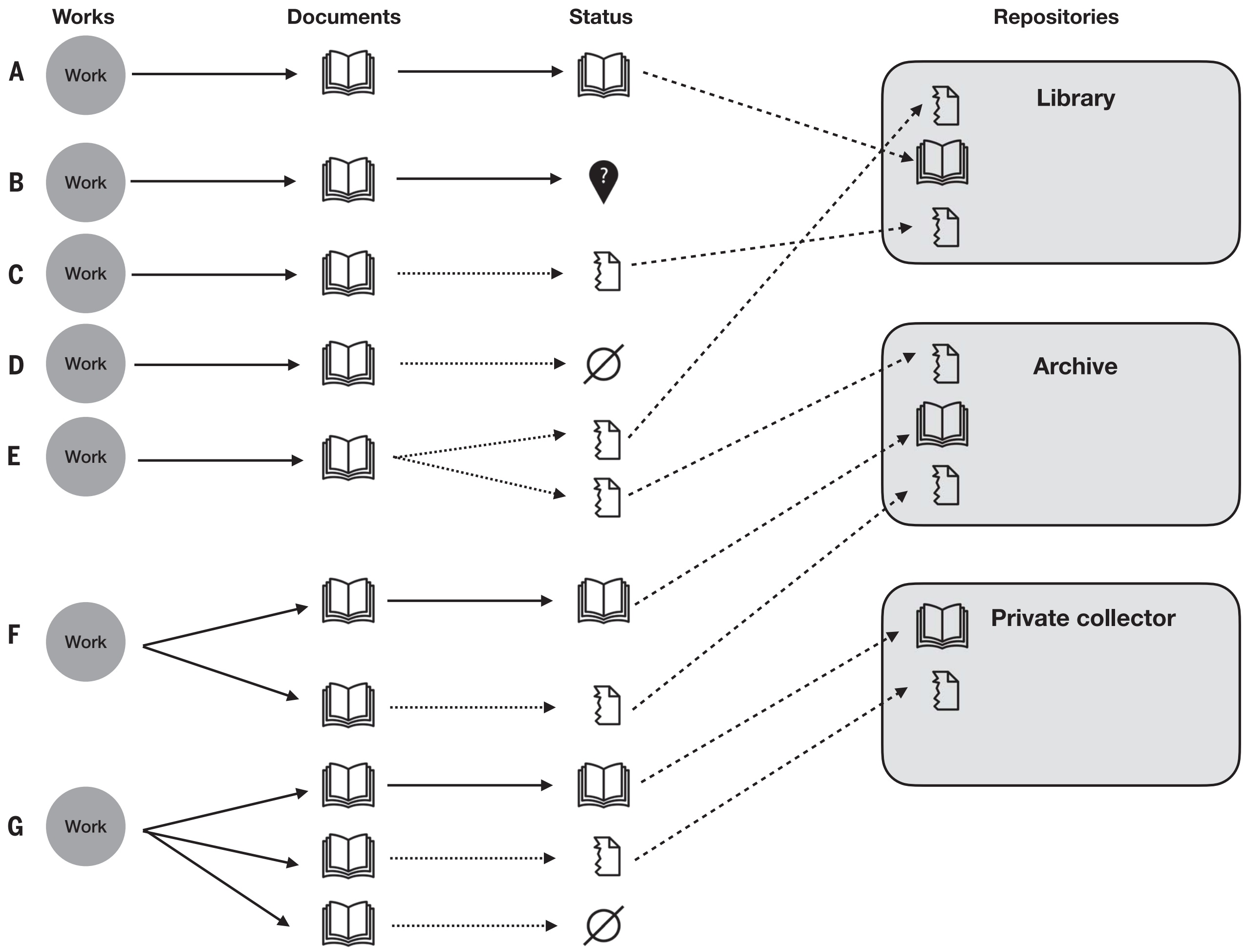


Fig. 2. Schematic representation of example survival scenarios for medieval literature. Individual works were copied into one [(**A**) to (**E**)] or more [(**F**) and (**G**)] documents, the survival status of which varies from intact codices (A) to fragments [(C) and (E)] residing in repositories such 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) and (D)], none of the original documents has been recovered.

published a pioneering estimator for the loss of multicopy 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*), then 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, and English) and three continental (Dutch, French, and German). For all works, we have listed the number of handwritten medieval documents in which they survive (Table 1). Next, we applied nonparametric methods to estimate the original richness of these traditions. For a given assemblage, let $(X_1, X_2, \ldots, X_{S_{obs}})$ represent the abundance-based frequencies for S_{obs} unique works that were observed in n documents.

Chao1 is a method to estimate a lower bound on $\hat{f}_0$, or the number of undetected species in an assemblage, based on the number of singletons (f_1, species sighted only once) and doubletons (f_2, species sighted exactly twice) in a sample of n individuals. The original number of works ($\hat{S}$) can then be estimated as $S_{obs} + \hat{f}_0$ (*22*). 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 hyperdiverse, undersampled 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 (*24*). In this framework, the survival ratio for the works can be quantified as the sample completeness or $S_{\text{obs}}/\hat{S}$: the ratio of the number of unique observed works (S_{obs}) over the estimated true species abundance $\hat{S}$ (*25*). Species richness is an intuitive measure to quantify species diversity, but there are alternative measures, such as 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 q, a scalar corresponding to different diversity measures at specific points (e.g., $q = 0$ for richness, $q = 1$ for Shannon, $q = 2$ for Simpson). Hill numbers are now 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 m of additional observations that are required to observe each of the $\hat{f}_0$ 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 $n + m$. 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, and 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 nonparametric, the lower bound is valid for any distribution of entities among types and it should be robust to differences in survival across document types (*15*).

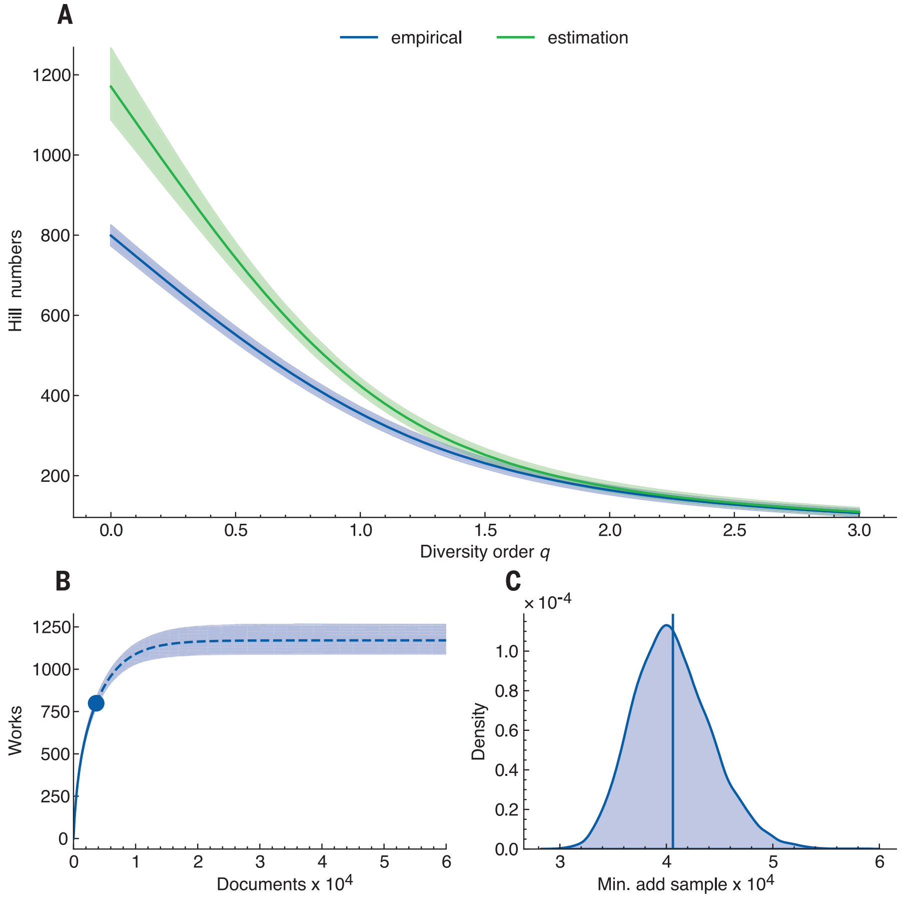


Fig. 3. Estimates for the union of the six assemblages. (**A**) Hill number curves (for 0 ≤ q ≤ 3), empirical and estimated, showing the absolute underestimation of the original diversity of works. (**B**) Species accumulation curve plotting the number of works as a function of the number of documents. The filled circle shows the observable data, the solid line the rarefaction for sample sizes <n, and the dashed line the extrapolation to sample sizes >n. (**C**) Kernel-density plot for the estimated number of documents.

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 (E_3) of evenness measures. 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 (*21*).

The results for the union of the corpora (Table 1 and table S2) suggest an overall survival ratio with a 68.3% confidence interval (CI) of 63.2 to 73.5% for works and a 9.0% CI of 7.5 to 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*). Figure 3A shows the empirical and estimated Hill number profiles. At $q = 0$, the curves indicate the absolute size of our current underestimation of the original diversity in the

Fig. 4. Heatmap of the geolocations of the repositories where documents are kept for four vernaculars. Figure was made with Leaflet version 1.7.1 software.

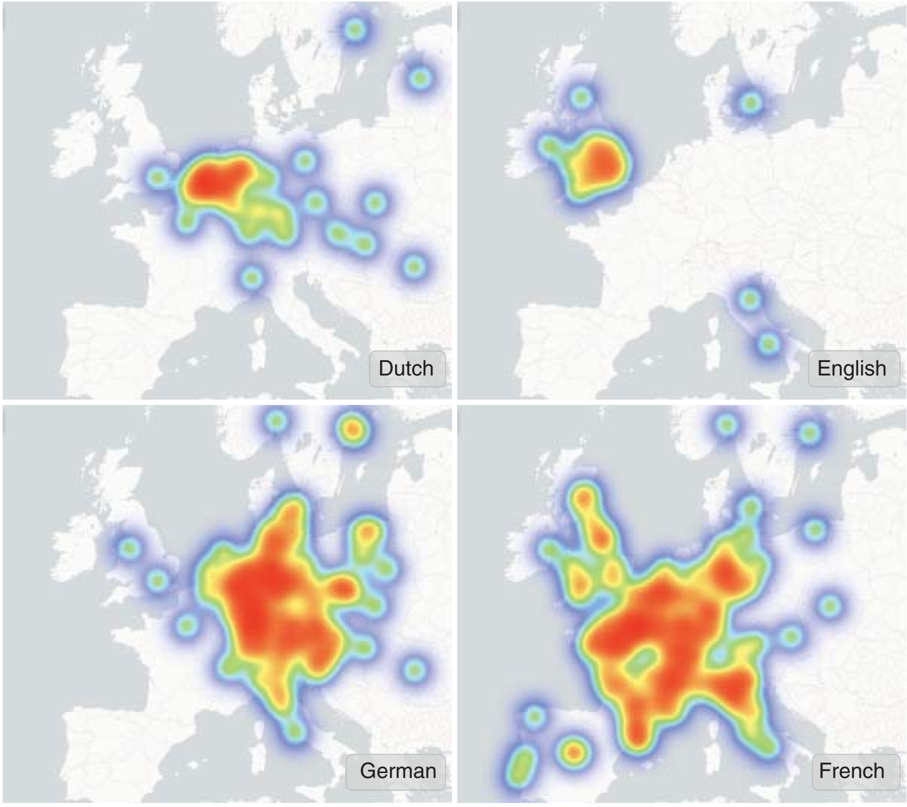
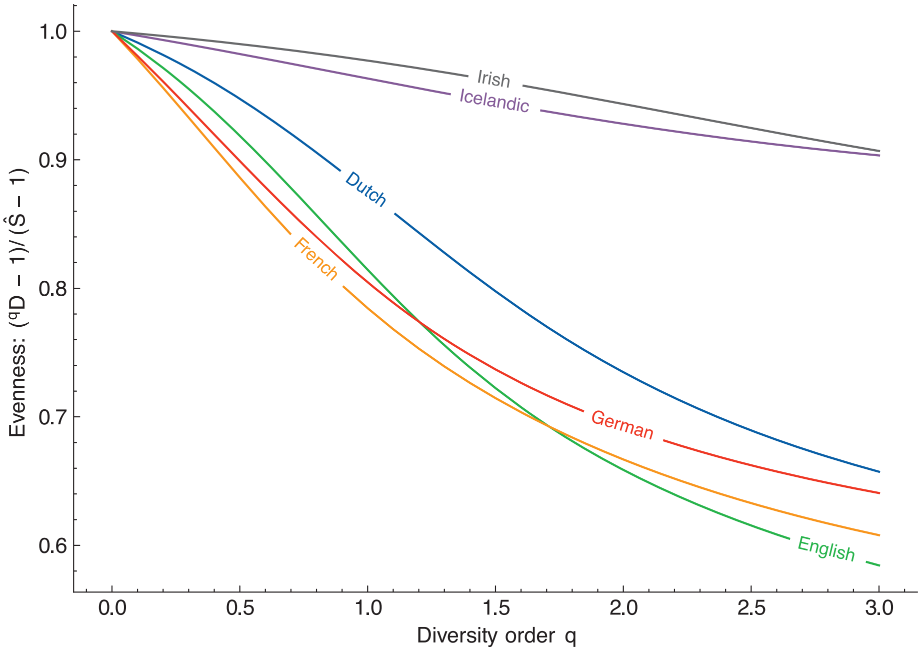


Fig. 5. Normalized evenness profiles (E_3) for the six individual vernaculars, plotting ^qE as a function of order $0 \le q \le 3$. The values on the *y*-axis reflect the estimated evenness in the reconstructed assemblages.



combined assemblage of chivalric and heroic narratives from the medieval period. Of the original ~1170 works that once would have existed, 799 would survive today. Likewise, the 3648 documents that are still observable constitute a sample from a population that originally would have counted ~40,614 specimens (Fig. 3C).

We observed considerable intervernacular variation (Table 1), ranging from the relatively poorly surviving English works (38.6%) to the relatively intact German tradition (79.0%). Dutch and French have a substantially lower survival factor than German, whereas two of the insular assemblages, Icelandic and Irish, have sustained similar losses to German, with point estimates of 77.3 and 81.0% and 16.9 and 19.2% for the survival of works and documents, respectively (*12*). It is puzzling that Old and Middle English documents did not travel far during their postmedieval afterlives (Fig. 4), yet other literatures survive in a wide manuscript diaspora. The survival estimates for works and documents yield similar rankings

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

(Table 1). In the supplementary materials, we compare Chao1 with three other estimators with similar results (fig. S1). Figure 5 shows the (estimated) evenness profiles and offers further insight into the distributional properties characterizing the assemblages. The profiles (fig. S2) for additional evenness classes ($E_1 - E_5$) yield consistent findings. Here, too, we note the atypical nature of Icelandic and Irish: Compared with the highly uneven distribution of French, for example, these two insular literatures feature a much more even distribution of documents over works.

Regarding documents, our results confirm the severity of the losses, with survival ratio 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 to 10.7%). Contrary to previous analyses (*16*, *17*), these results are therefore compatible with 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 observed before and challenge existing assumptions. For example, our results suggest that Irish and Icelandic literature has been preserved comparatively well compared with 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 with mainland regions (*30*). Additionally, insular assemblages demonstrate a higher species evenness because of the lack of predators and other factors. A parallel emerges with some of the cultural diversity profiles for island regions reconstructed here: If land-isolated areas preserve biological heritage more effectively, then 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 have identified 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, whereas the loss figures for Icelandic and Irish are considerable, their distributional characteristics seem to have made them more resistant 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 network 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, for example, 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).
2. F. Bruni, A. Pettegree, *Lost Books: Reconstructing the Print World of Pre-Industrial Europe* (Brill, 2016).
3. N. J. Gotelli, R. K. Colwell, in *Biological Diversity: Frontiers in Measurement and Assessment*, A. E. Magurran, B. J. McGill, Eds. (Oxford University Press, 2011), pp. 39–54.
4. L. J. Gorenflo, S. Romaine, R. A. Mittermeier, K. Walker-Painemilla, *Proc. Natl. Acad. Sci. U.S.A.* **109**, 8032–8037 (2012).
5. H. Zhang, R. Mace, *Evol. Hum. Sci.* **3**, e30 (2021).
6. J. Henrich, *Am. Antiq.* **69**, 197–214 (2004).
7. J. Henrich *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **113**, E6724–E6725 (2016).
8. C. P. Lipo, R. J. DiNapoli, M. E. Madsen, T. L. Hunt, *PLOS ONE* **16**, e0250690 (2021).
9. A. Acerbi, J. Kendal, J. J. Tehrani, *Evol. Hum. Behav.* **38**, 474–480 (2017).
10. E. Kwakkel, *Books Before Print* (Arc Humanities Press, 2018).
11. U. Neddermeyer, *Von der Handschrift zum gedruckten Buch. Schriftlichkeit und Leseinteresse im Mittelalter und in der frühen Neuzeit. Quantitative und qualitative Aspekte* (Harrassowitz, 1998).
12. D. Ó Corráin, *Peritia* **22–23**, 191–223 (2011–2012).
13. R. Wilson, *The Lost Literature of Medieval England* (Methuen, ed. 2, 1970).
14. P. Eggert, *The Work and the Reader in Literary Studies: Scholarly Editing and Book History* (Cambridge Univ. Press, 2019).
15. H. Wijsman, *Luxury Bound. Illustrated Manuscript Production and Noble and Princely Book Ownership in the Burgundian Netherlands (1400-1550)* (Brepols, 2010).
16. J. L. Cisne, *Science* **307**, 1305–1307 (2005).
17. G. Declercq, *Science* **310**, 1618 (2005).
18. N. D. Pyenson, L. Pyenson, *Science* **309**, 698–701 (2005).
19. M. S. Cuthbert, *Musica Disciplina* **54**, 39–74 (2009).
20. L. Egghe, G. Proot, *J. Informetrics* **1**, 257–268 (2007).
21. Materials and methods are available as supplementary materials.
22. A. Chao, *Scand. J. Stat.* **11**, 265–270 (1984).
23. A. Chao, C. H. Chiu, in *Methods and Applications of Statistics in the Atmospheric and Earth Sciences*, N. Balakrishnan, Ed. (Wiley, 2012), pp. 76–111.
24. M. I. Eren, A. Chao, W.-H. Hwang, R. K. Colwell, *PLOS ONE* **7**, e34179 (2012).
25. A. Chao *et al.*, *Ecol. Res.* **35**, 292–314 (2020).
26. M. O. Hill, *Ecology* **54**, 427–432 (1973).
27. A. Chao, R. K. Colwell, C.-W. Lin, N. J. Gotelli, *Ecology* **90**, 1125–1133 (2009).
28. A. Chao, C. Ricotta, *Ecology* **100**, e02852 (2019).
29. I. Donohue *et al.*, *Ecol. Lett.* **19**, 1172–1185 (2016).
30. R. J. Whittaker, J. M. Fernández-Palacios, *Island Biogeography: Ecology, Evolution, and Conservation* (Oxford Univ. Press, 2006).
31. A. Acerbi, R. A. Bentley, *Evol. Hum. Behav.* **35**, 228–236 (2014).
32. M. Cantor *et al.*, *Proc. Biol. Sci.* **288**, 20203107 (2021).

ACKNOWLEDGMENTS

We thank D. Schoenaers, G. Henley, J.-B. Camps, B. Bastert, D. Könitz, and J. Deploige for their help, as well as the five anonymous referees. **Funding:** E.D.B. was supported by a postdoctoral fellowship from FWO Flanders; K.A.K. was supported by the Carlsberg Foundation (Visiting Fellowship CF20-225 and H. M. Queen Margrethe II Distinguished Research Fellowship CF18-500). **Author contributions:** Conceptualization: M.K., F.K., E.D.B., M.D., K.A.K., P.O.M., D.S., R.S., A.C.; Data curation: M.K., F.K., E.D.B., M.D., K.A.K., P.O.M., D.S., R.S., A.C.; Formal analysis: M.K., F.K.; Investigation: M.K., F.K., E.D.B., M.D., K.A.K., P.O.M., D.S., R.S., A.C.; Methodology: M.K., F.K., A.C.; Software: M.K., F.K.; Validation: M.K., F.K., E.D.B., M.D., K.A.K., P.O.M., D.S., R.S., A.C.; Visualization: M.K., F.K.; Writing – original draft: M.K., F.K., E.D.B., M.D., K.A.K., P.O.M., D.S., R.S., A.C.; Writing – review and editing: M.K., F.K., E.D.B., M.D., K.A.K., P.O.M., D.S., R.S., A.C. **Competing interests:** The authors declare no competing interests. **Data and materials availability:** Data and code have been deposited to an open access repository (33). This paper is released with a Python software package (Copia, available from PyPI), all under a CC-BY-SA license, to replicate our findings and stimulate the adoption of this approach in other domains.

SUPPLEMENTARY MATERIALS

science.org/doi/10.1126/science.abl7655
Materials and Methods
Supplementary Text
Figs. S1 to S3
Tables S1 and S2
References (*33*–*99*)

4 August 2021; accepted 18 November 2021
10.1126/science.abl7655

Science

Forgotten books: The application of unseen species models to the survival of culture

Mike KestemontFolgert KarsdorpElisabeth de BruijnMatthew DriscollKatarzyna A. KapitanPádraig Ó MacháinDaniel SawyerRemco SleiderinkAnne Chao

Science, 375 (6582), • DOI: 10.1126/science.abl7655

Ecological methods for cultural history

Much of the narrative literature from the European Middle Ages has been lost over the ages because of manuscript physical degradation and destruction, including library fires. Kestemont *et al*. show that established methods from ecology for estimating the numbers of unseen species can be applied to abundance data representing cultural artifacts to estimate the losses that ancient cultural domains have sustained over the centuries. The authors obtain estimates that not only corroborate existing hypotheses from book history, but also reveal unexpected geographic differences that have gone unnoticed so far. For example, insular literatures, such as those from Iceland and Ireland, combine a surprisingly strong cultural persistence with an elevated distributional evenness. —AMS

Science (ISSN) is published by the American Association for the Advancement of Science. 1200 New York Avenue NW, Washington, DC 20005. The title *Science* is a registered trademark of AAAS.