Supplementary Material

The changing uses of herbarium data in an era of global change: an overview using automated content analysis

J. Mason Heberling, L. Alan Prather and Stephen J. Tonsor

Supplementary Methods

We identified all the published journal articles that directly refer to herbarium specimen data that could be detected using automated literature searches. We first searched the topic field in Web of Science (Clarivate Analytics, formerly ISI), including all years from 1900-2017. Web of Science's topic field includes title, abstract, and keywords: Topic = (herbarium OR herbaria). We restricted search results to journal articles with available abstracts in English contained in the following major databases indexed by the Web of Science: Web of Science Core Collection, BIOSIS Citation Index, and Biological Abstracts. Results were then refined to only include peer-reviewed journal articles (i.e., we excluded document types such as news items, book reviews, and poetry) and excluded irrelevant subject categories. All were included in subsequent quantitative analyses (described below), with the exception of 30 articles that were not relevant (e.g., obituaries of plant collectors removed by searching titles) and two articles published prior to 1920 (sample size too low for meaningful analysis). The initial Web of Science search was conducted on 1 November 2016. Publications for 2016 and 2017 were added 4 January 2017 and 10 May 2018, respectively. This search returned publications based on herbarium specimens themselves (including label metadata) and any specimen derived metadata (e.g., georeferenced localities, annotations, specimen images, measured traits, etc.).

We performed an additional similar literature search using the Elsevier Scopus database on 12 January 2017 using the same search query. Scopus and Web of Science results were merged, followed by removal of all duplicates. The final dataset analyzed consisted of 13,702 abstracts (including article titles) of herbarium-related articles, published from 1923 to 2017 (Figure S1).

Additional background on structural topic models. Automated content analysis, also called topic modeling, broadly refers to a wide set of algorithms and statistical approaches for

synthesizing text using machine learning tools (Blei 2010). Topic models have been widely used in the social sciences, humanities, and medical fields, and more recently applied in ecology and evolutionary biology (Nunez-Mir et al. 2016).

Topic models use text-parsing and machine learning tools developed by computer scientists to discover associations between words in large volumes of documents and organize these associations into clusters of words that tend to occur together. These clusters of co-occurring words are known topics (sometimes called "concepts" or "themes") within a body of literature (corpus). Thus, any number of topics can emerge inductively in a corpus (i.e. unsupervised, or not chosen in advance by the researcher) based on word co-occurrence within and between texts and their prevalence across the corpus. To be associated with each other in a topic, words must occur more often in association with each other than they would if drawn at random from the corpus vocabulary, given their individual frequencies in the corpus. Topics are therefore defined by probability distributions for a vocabulary of words that group together more than they group with other word groups. In our review, we used modification of a *latent Dirichlet allocation* (LDA) topic model approach (Blei et al. 2003). The LDA approach is an unsupervised, mixed-membership model, wherein each word within a document belongs to a given topic and each document can include multiple topics. Each document is represented as a vector of topic proportions according to fractions of words assigned to a given topic. Each topic's importance in the corpus as a whole is a function of proportion of all documents in the corpus that are associated with the topic.

More specifically, we used an approach called Structural Topic Modeling (STM; Roberts et al. 2014), which is derived from the earlier developed *latent Dirichlet allocation* (LDA) topic model approach (Blei et al. 2003) to quantitatively synthesize trends in herbarium research through time. As summarized by Roberts et al. (2014), these models differ from LDA in that STMs provide "structure" such that: 1) topics can be correlated to each other, providing some insight into relationships among topics and allowing meta-topics to emerge, 2) each document has a prior distribution across topics, defined by covariate(s) instead of a shared global mean; in our study there is a distribution specific to the year of publication of the document, and 3) word prevalence within each topic can vary with a covariate.

Details on the advantages of structural topic models compared to hand coding are briefly described in the main text, and elaborated elsewhere (Blei 2010; Farrell 2016; Nunez-Mir et al. 2016). Technical details on structural topic modelling can be found in Roberts et al. (2014).

Model selection. As outlined above, we followed the model selection procedure described by Farrell (2016). While structural topic modelling is an unsupervised approach, it still requires user validation in the determination of how many topics to model. One quantitative method to assist in deciding how many topics to include for a given corpus is through the metrics of exclusivity and semantic coherence (Roberts et al. 2014). Semantic coherence measures the frequency that high probability words tend to co-occur in documents, while exclusivity measures the proportion of high probability words that are distinct to a given topic. Theoretically, the "best" model will optimize both of these metrics. However, there is a tradeoff between model exclusivity and semantic coherence. Therefore, in practice, there is not an obvious "best" number of topics to model based on these metrics alone. A comparison of these metrics across many models for the current set of herbarium-based abstracts suggests between 20-30 topics to be most informative (figure S2).

In addition to considering model exclusivity and semantic coherence, we closely interpreted the outputs from a range of models with 10-50 topics included. For each model with a different number of topics included, we read the top abstracts associated with each topic, along with the high probability words that define that topic. This recursive process resulted in the selection of a topic model that provided a comprehensive, yet meaningful, overview of the herbarium-related literature. Since we had no *a priori* prediction on the number of topics that this literature should encompass, we selected the 25-topic model because the resulting topics were logically sound and covered a diversity of herbarium-related research areas. However, our results are robust to the number of topics modelled and did not affect the overall trends presented in the main text of the current review. Many topics were robust to changes in number of topics modeled. For example, topics best described as global change biology, morphometrics, ethnobotany, new species descriptions, and DNA analyses were consistently present and remarkably similar in many model versions. However, including fewer topics resulted in the obvious lumping of otherwise meaningful topic areas that shared certain words and word relationships (e.g., algae and invasion biology) as well as "catch all" topics that lacked meaning

(figure S3, table S1). Similarly, including additional topics resulted in additional topics with clear meaning (e.g., invasion biology), but also resulted in "junk" topics and the breaking up of meaningful topics into a level of detail that was not appropriate for the current review (e.g., Taxonomy I-IV; figure S4, table S2).

Supplementary Tables

Table S1. Structural topic model results (with 20 topics) from 13,702 abstracts of herbarium-related studies*, including topic name, top 15 words with highest probability in each topic, and a general qualitative description of each topic. Topics are numbered chronologically, ranked by proportion of entire corpus belonging to each topic. "Chimera topics" that combine two or more different topics from 25 topic model or are otherwise different are highlighted in red font.*

	Topic name	Top associated words	General description
1.	History of botanists & collections	*collect, botan, garden, research, work, museum, natur, histori, data, inform, flora, centuri, import, public, botanist*	Botanical history, focused on particular botanists, expeditions, or collections
2.	Taxonomic monographs & revisions	*genus, taxonom, key, revis, base, provid, distribut, descript, describ, present, section, new, taxa, includ, illustr*	Comprehensive treatments of taxon groups
3.	Species distributions	*distribut, flora, local, present, found, area, data, part, region, record, materi, literatur, europ, itali, rare*	Distribution of plant, lichen, and bryophyte taxa
4.	Typification/Nomenclature	name, type, materi, nomenclatur, origin, lectotyp, design, describ, publish, synonym, typif, taxa, ident, valid, collect	Designation of type specimens and nomenclatural updates
5.	Conservation biology	*conserv, area, distribut, data, divers, endem, region, habitat, record, invas, rich, flora, use, natur, inform*	Biology, management, and assessments of rare or threatened species
6.	Taxonomic notes on genera/families	*list, note, given, includ, addit, literatur, descript, discuss, synonymi, citat, genus, materi, name, part, form*	Short reports and focused synopses of specific taxa (towards taxonomic monographs)
7.	Neotropical floristics	*famili, state, brazil, forest, genera, collect, present, repres, flora, distribut, area, region, found, brazilian, record*	Biodiversity studies in South America (esp. Amazon region)
8.	New species descriptions	*new, record, collect, australia, report, describ, india, known, flora, previous, fern, south, western, zealand, australian*	Alpha taxonomy
9.	Morphometric studies	*morpholog, charact, group, hybrid, popul, analysi, variat, complex, taxa, differ, distinct, subsp, use, number, variabl*	Inter- and infraspecific studies based on statistical analyses of morphology (phenetics) (numerical taxonomy)
10.	Soil fungi (CHIMERA)	*new, china, collect, lichen, deposit, univers, fungi, institut, record, type, describ, report, scienc, taxa, provinc*	Mostly species descriptions of soil fungi in China
11.	Global change biology	*chang, climat, use, model, increas, rang, differ, data, distribut, time, sampl, environment, signific, show, flower*	Responses to past and future environmental change (esp. atmospheric change: CO_2, climate); phenological change through time; community- and population-level change
12.	Algae (CHIMERA)	*north, island, america, collect, south, distribut, southern, california, northern,*	Mostly Algal floristics & taxonomy

		coast, eastern, occur, found, american, region	
13.	Herbarium methodology, phytopathology (CHIMERA)	*host, isol, cultur, extract, dri, wood, pathogen, method, diseas, caus, found, fungus, fungi, chemic, contain*	Combined topics based on herbarium methodology, phytochemistry, and phytopathology
14.	Lists from collections (CHIMERA)	*var, nov, ssp, comb, subsp, new, follow, varieti, combin, benth, propos, hook, stat, festuca, arg*	Combined topics of lists from specific collections or taxon groups and reference to type specimens
15.	Morphometrics II (CHIMERA)	flower, fruit, leav, differ, long, form, infloresc, branch, stem, floral, leaf, charact, distinguish, margin, character	Redundant to morphometrics topic (topic 2 above)
16.	DNA analyses	*sequenc, dna, phylogenet, molecular, use, morpholog, data, analys, clade, genet, sampl, region, within, analysi, gene*	Extraction, amplification, and analysis of DNA (esp. molecular systematics)
17.	Ethnobotany	*use, medicin, tradit, local, inform, knowledg, identifi, famili, collect, part, method, survey, district, treatment, ethnobotan*	Traditional plant knowledge; economic and medicinal botany
18.	Regional observations & reports	*mexico, moss, peninsula, distribut, taxa, bryophyt, spain, endem, argentina, present, mexican, iberian, provinc, region, chile*	Taxon occurrences at local or regional scales, primarily geopolitical units (esp. county level in North America);
19.	Morphology & anatomy	*pollen, leaf, cell, structur, spore, morpholog, type, use, materi, studi, anatomi, grain, anatom, featur, differ*	Morphology and anatomy of specific structures, esp. at micro-level
20.	CHIMERA (unclear; “junk topic”)	*seed, cultiv, wild, collect, orchid, orchidacea, costa, peru, colombia, crop, solanum, rica, field, ecuador, genet*	Interpretation unclear and not meaningful

***Table S2. Structural topic model results (with 30 topics) from 13,702 abstracts of herbarium-related studies**, including topic name, top 15 words with highest probability in each topic, and a general qualitative description of each topic. Topics are numbered chronologically, ranked by proportion of entire corpus belonging to each topic. Topics that substantially differ in content from any topic in 25 topic model (as presented in main text) are highlighted in red font.*

	Topic name	Top associated words	General description
1.	Taxonomic treatments & revisions (CHIMERA)	*genus, descript, key, revis, taxonom, distribut, provid, section, includ, note, base, given, discuss, present, illustr*	Comprehensive treatments of taxon groups AND taxonomic notes
2.	Morphometric studies	*morpholog, charact, group, taxa, analysi, taxonom, complex, variat, differ, distinct, use, popul, within, base, result*	Inter- and infraspecific studies based on statistical analyses of morphology (phenetics) (numerical taxonomy)
3.	History of botanists & collections	*botan, work, research, collect, botanist, centuri, public, botani, scientif, flora, institut, publish, scienc, year, first*	Botanical history, focused on particular botanists, expeditions, or collections
4.	Species distributions I (CHIMERA)	*distribut, lichen, local, data, europ, present, part, poland, map, occurr, materi, rare, literatur, record, found*	Distribution of plant, lichen, and bryophyte taxa (especially in Europe)
5.	Floristics I (CHIMERA)	*flora, taxa, famili, genera, list, vascular, record, includ, present, total, checklist, fern, moss, number, peninsula*	Species checklists and community descriptions at regional to local scales (preserves, parks, cities, physiographic regions, etc.)
6.	Floristics II (CHIMERA)	*area, forest, region, veget, endem, divers, distribut, rich, habitat, high, florist, mountain, tree, park, soil*	Nearly identical interpretation to Floristics I topic, but narrower geographic scope
7.	Typification/Nomenclature I (CHIMERA)	*name, nomenclatur, lectotyp, design, synonym, origin, materi, publish, typif, propos, valid, ident, lectotypif, correct, accept*	Designation of type specimens and nomenclatural updates
8.	Biodiversity informatics & conservation	*data, conserv, use, inform, databas, assess, biodivers, distribut, status, can, provid, research, system, need, threaten*	Biodiversity conservation; Biodiversity databases; digitization of collections; statistical methods for biodiversity analysis
9.	Lists from collections I (CHIMERA)	*collect, known, made, also, found, repres, recent, sever, survey, mani, number, first, addit, previous, present*	Summaries of single collections, esp. those of particular botanists or type collections in a given natural history museum (redundant to Topic 10)
10.	Lists from collections II (CHIMERA)	*botan, museum, garden, list, given, histori, includ, nation, natur, note, univers, collect, british, collector, kew*	Summaries of single collections, esp. those of particular botanists or type collections in a given natural history museum (redundant to Topic 9)
11.	Soil fungi (CHIMERA)	*new, record, china, deposit, provinc, report, univers, first, describ, time, institut, chines, scienc, illustr, korea*	Mostly species descriptions of soil fungi in China
12.	Herbarium methodology & phytochemistry	*use, sampl, method, extract, dri, materi, concentr, chemic, result, differ, test, content, acid, leav, activ*	Techniques for specimen preservation; herbarium pest management; chemical properties of specimens
13.	Taxonomy I (CHIMERA)	*hybrid, cultiv, form, origin, may, natur, materi, probabl, possibl, appear, show, wild, seem, suggest, indic*	Taxonomic studies with focus on hybridization and/or taxonomic revisions (species splitting)
14.	New species descriptions	*new, south, describ, africa, australia, southern, western, african, endem,*	Alpha taxonomy

		australian, tropic, occur, northern, region, east	
15.	Global change biology	*chang, climat, distribut, rang, model, use, environment, increas, differ, data, temperatur, respons, time, pattern, signific*	Responses to past and future environmental change (esp. atmospheric change: CO_2, climate); phenological change through time; community- and population-level change
16.	Taxonomic notes on genera/families	*var, nov, ssp, new, comb, varieti, follow, combin, festuca, descr, propos, minor, stat, note, gray*	Short reports and focused synopses of specific taxa (towards taxonomic monographs)
17.	Neotropical floristics	*brazil, state, brazilian, rio, genera, present, distribut, forest, famili, atlant, collect, repres, survey, found, record*	Biodiversity studies in South America (esp. Amazon region)
18.	Floristics III (CHIMERA)	*state, north, mexico, america, counti, california, american, distribut, unit, usa, eastern, mexican, report, rang, canada*	Similar interpretation to Floristics I and II topics, but focused on geopolitical regions
19.	Species distributions II (CHIMERA)	*subsp, itali, subspeci, iran, flora, mediterranean, taxa, turkey, boiss, chorolog, italian, report, investig, spain, belong*	Very similar to Topic 4
20.	Typification/Nomenclature II (CHIMERA)	*type, materi, describ, sheet, list, kept, taxa, present, collect, label, preserv, origin, holotyp, belong, institut*	Similar to Topic 7 (and also lists of types from collections)
21.	Morphology & anatomy I (CHIMERA)	*leaf, cell, spore, structur, anatomi, surfac, light, wall, featur, type, anatom, studi, electron, microscop, shape*	Similar to Topic 20
22.	Ethnobotany	*use, medicin, tradit, local, famili, knowledg, inform, identifi, part, district, ethnobotan, interview, survey, treatment, peopl*	Traditional plant knowledge; economic and medicinal botany
23.	Fungi & phytopathology	*fungi, host, pathogen, isol, cultur, fungus, diseas, fungal, rust, caus, infect, parasit, associ, strain, collect*	Studies on fungi and pathogens and their plant hosts
24.	Invasion biology	*popul, nativ, invas, histor, record, natur, spread, introduc, weed, alien, introduct, habitat, site, year, centuri*	Studies on non-native species biology, introduction history, and spread
25.	DNA analyses	*sequenc, dna, phylogenet, molecular, use, genet, clade, data, region, analys, gene, within, sampl, divers, support*	Extraction, amplification, and analysis of DNA (esp. molecular systematics)
26.	Taxonomy II (CHIMERA)	*flower, fruit, leav, seed, infloresc, long, differ, floral, stem, leaf, branch, distinguish, hair, length, small*	Taxonomic studies with focus on morphology
27.	Taxonomy III (CHIMERA)	*india, indian, rubiacea, madagascar, endem, new, philippin, panama, thailand, borneo, himalaya, malaysia, costa, psychotria, guiana*	Taxonomic studies with focus on regional taxonomic treatments
28.	Algal floristics & taxonomy	*island, coast, alga, zealand, pacif, marin, record, sea, archipelago, atlant, ocean, japan, algal, rhodophyta, report*	Marine biology (esp. macrophytes)

29.	Taxonomy IV (CHIMERA)	*argentina, america, peru, venezuela, american, neotrop, colombia, chile, south, ecuador, solanum, bolivia, del, cuba, potato*	Taxonomic studies with focus on South America and other tropical regions
30.	Morphology & anatomy II (CHIMERA)	*pollen, chromosom, number, orchid, popul, diploid, grain, orchidacea, cytolog, societi, tetraploid, sexual, count, linnean, found*	Morphology and anatomy of specific structures, esp. at micro-level

Supplementary Figures

Figure S1. Flow chart detailing process of literature compilation for inclusion in topic models. *Note: Web of Science search included the following databases: Web of Science Core Collection, BIOSIS Citation Index, and Biological Abstracts.*

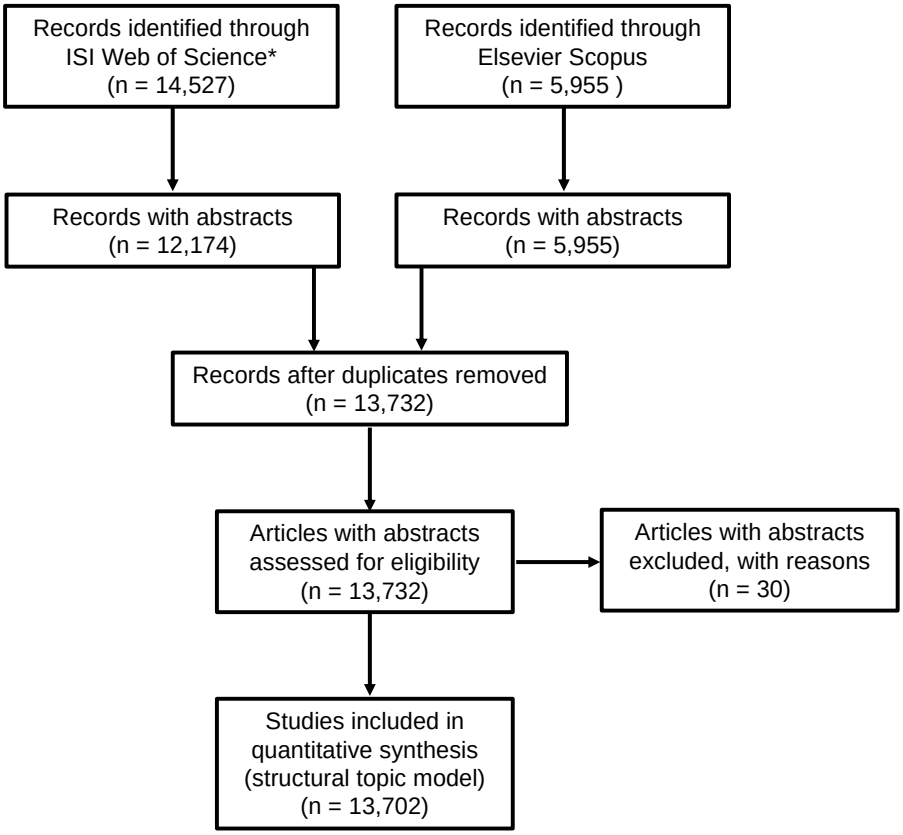


Figure S2. Relationship between model exclusivity and semantic coherence. *Each number refers to a separate topic model that included that number of topics. The number's location on the graph corresponds to the average values for that topic model for exclusivity and semantic coherence.*

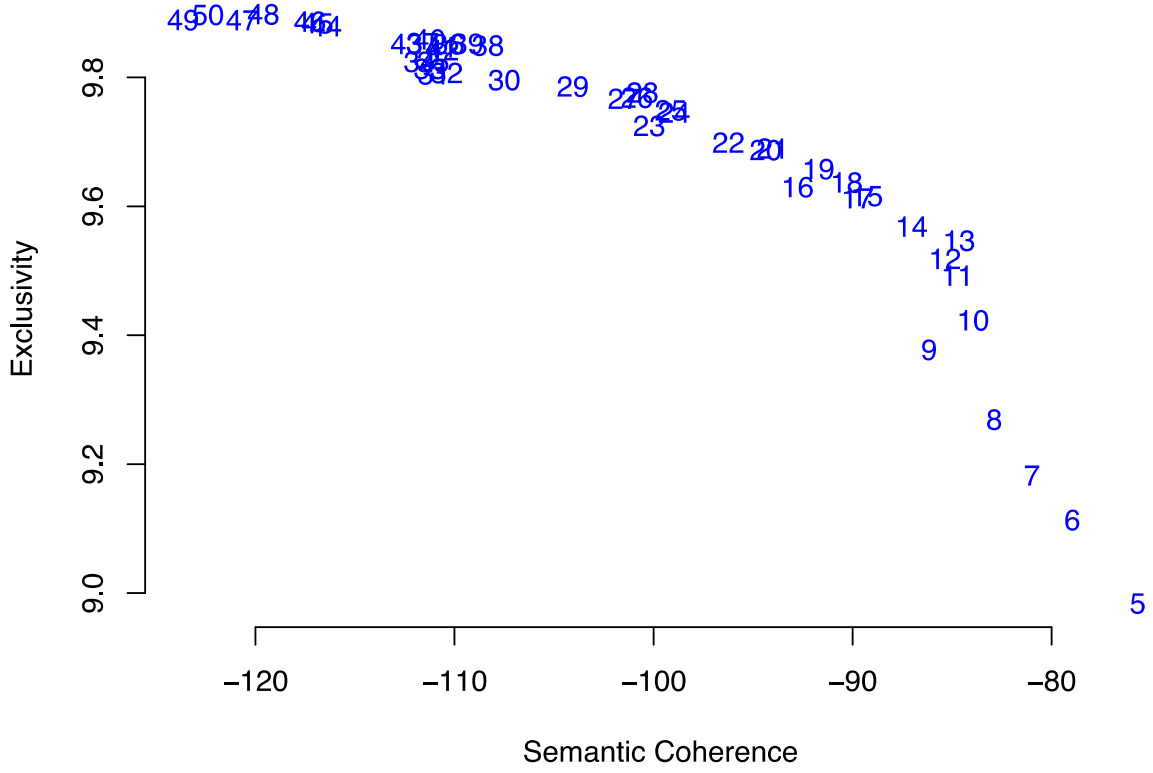


Figure S3. Topic proportions from a 20-topic structural topic model of herbarium-related literature. *Topic proportions are the percentage of the total corpus that belongs to each topic. Topic names were defined from top words associated with each topic and holistic themes across the top abstracts related to each topic. "Chimera topics" that combine two or more different topics from the 25-topic model presented in the main text are denoted with an *.*

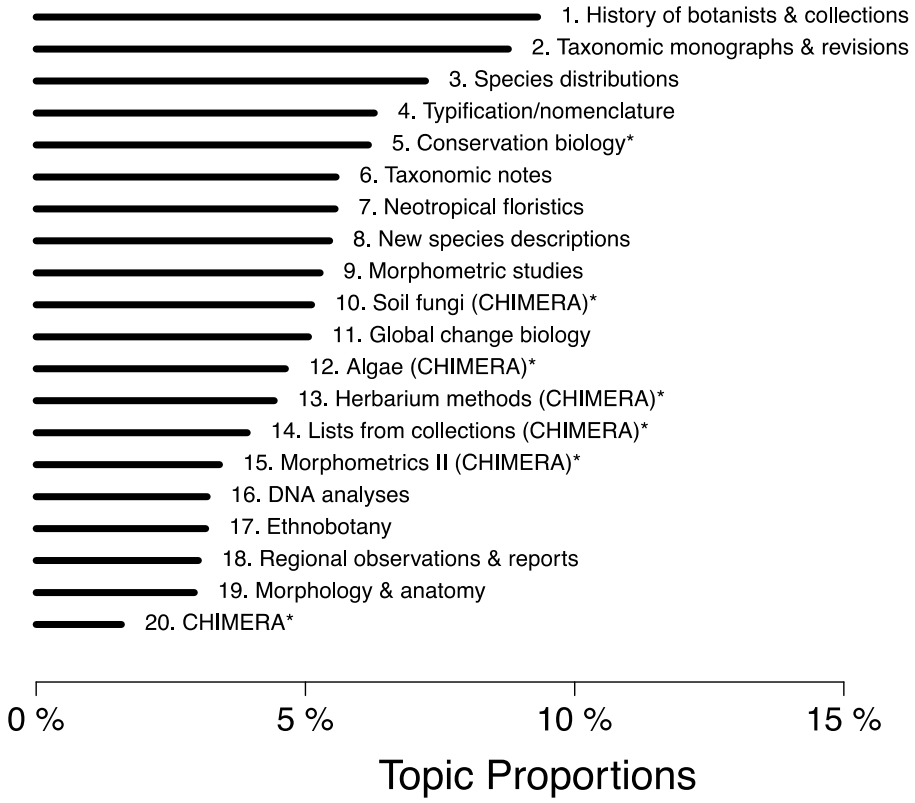
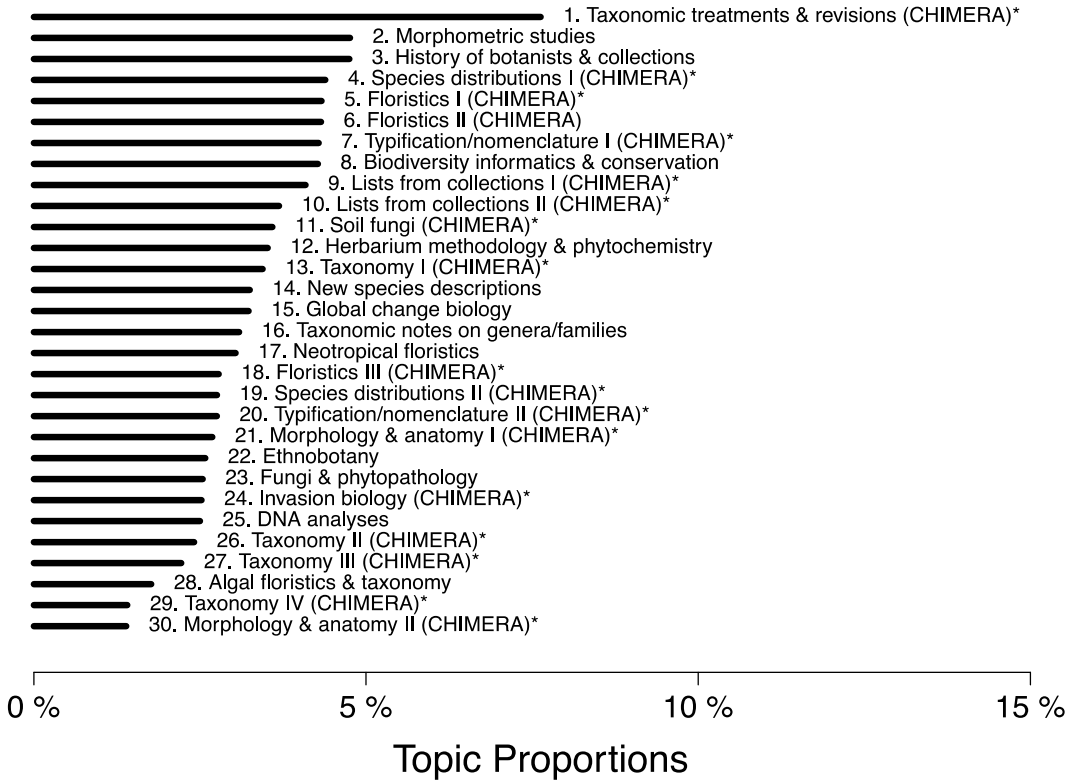


Figure S4. Topic proportions from a 30-topic structural topic model of herbarium-related literature. *Topic proportions are the percentage of the total corpus that belongs to each topic. Topic names were defined from top words associated with each topic and holistic themes across the top abstracts related to each topic. "Chimera topics" that combine two or more different topics from the 25-topic model presented in the main text are denoted with an *.*



References Cited (Supplementary Materials)

Blei DM. 2010. Probabilistic topic models. Commun ACM. 55(4):77–84. doi:10.1109/MSP.2010.938079.

Farrell J. 2016. Corporate funding and ideological polarization about climate change. Proc Natl Acad Sci. 113(1):92–97. doi:10.1073/pnas.1509433112.

Nunez-Mir GC, Iannone B V., Pijanowski BC, Kong N, Fei S. 2016. Automated content analysis: Addressing the big literature challenge in ecology and evolution. Methods Ecol Evol. 7(11):1262–1272. doi:10.1111/2041-210X.12602.

Roberts ME, Stewart BM, Tingley D, Lucas C, Leder-Luis J, Gadarian SK, Albertson B, Rand DG. 2014. Structural topic models for open-ended survey responses. Am J Pol Sci. 58(4):1064–1082. doi:10.1111/ajps.12103.