Trends in
Ecology & Evolution

CellPress
OPEN ACCESS

Forum

# AI-based discovery of habitats from museum collections

Christopher B. Jones [1,*], Kristin Stock [2] and Sarah E. Perkins [3]

**Museum collection records are a source of historic data for species occurrence, but little attention is paid to the associated descriptions of habitat at the sample locations. We propose that artificial intelligence methods have potential to use these descriptions for reconstructing past habitat, to address ecological and evolutionary questions.**

## Habitat description in natural history collections

The value of museum and other natural history archival records for ecological research has been highlighted in several studies [1–3]. Currently the **Global Biodiversity Information Facility (GBIF)** (see Glossary) reports more than two billion digital records, the earliest of which date back to the 18th century. Most studies of such records focus on exploitation of species occurrence data for biodiversity research and species distribution models [1]. These written records, however, also frequently contain accompanying textual descriptions of the environment in which specimens were collected (Box 1), that, to date, have rarely been utilised, (but see [4] using habitat data, and [5] ecological traits). Artificial intelligence (AI) methods for **natural language processing (NLP)** now provide the potential for automated reconstruction of past habitats from the evidence of these descriptions in combination with species occurrence data.

## What can habitat descriptions tell us?

Large extent mining of written habitat descriptions could provide evidence of the recent past that complements existing map surveys and studies of physical specimens [6]. It could yield immense quantities of historical data which, when linked to the species records, would deepen our understanding of changes in habitat and the interactions between species and habitat that could provide ecological and evolutionary insight. We acknowledge though that collection bias may limit sample size and inference [7].

Habitat data could supplement existing AI-driven biodiversity research (specifically species distribution data) to reveal mechanisms facilitating the spread of invasive non-native species (see for example [8]); climate-induced species shifts; shifts in community interactions and help understand plant and animal disease dynamics from a **One Health** perspective, in which habitat change is an integral part of pathogen spread. Habitat data can provide important evolutionary insight into drivers of biogeographical patterns at a global scale [9]. However, understanding what shapes the distribution and formation of species has, to date, been rather limited by a lack of empirical data. Exploiting historical habitat data and linking this with the field of **museomics** (genomic data from museum samples) over time could provide new opportunities to test evolutionary drivers of species change in the Anthropocene [3].

Georeferenced habitat data alone could provide an indication of the extent of habitat loss over time and also, when used in comparison with present day records, of land conversion (change in land use), where that information may otherwise be unknown, so providing insight into whether a **shifting baseline syndrome** is occurring; the idea that each generation perceives the environment based on their

## Glossary

**CORINE:** a land cover classification data set for the EU, published as snapshots every 6–10 years since 1990.
**Deep learning:** applications of neural networks that have multiple internal layers of neurons.
**Extended specimen:** concept that expands beyond traditional understanding of a physical specimen due to associated data and resources for the individual organism.
**Georeferencing:** assignment of coordinate location (e.g., latitude and longitude, or easting and northing on a recognised coordinate system) to place names or text descriptions of location.
**Global Biodiversity Information Facility (GBIF):** a global network and infrastructure (www.gbif.org) that provides an aggregated data set created by contributions from collection agencies around the world, published in a standardised format.
**Ground truth data:** data for which we know the classification, or correct answer, used to train and evaluate machine learning models.
**Habitat conversion:** the process of transforming natural habitats into other land uses; for example, agriculture or urbanisation.
**Machine learning classification:** use of algorithms that learn statistical patterns from existing data to make predictions for new data in order to classify data objects (e.g., land cover spatial units) into one of two or more classes.
**Museomics:** the study of museum genomics; ancient DNA and historic DNA specimens in museum collections.
**Natural language processing (NLP):** computational methods for parsing the grammatical structure and semantics of human language (e.g. text), extracting meaning and identifying patterns.
**Neural network:** a machine learning algorithm consisting of layers of artificial neurons, that output values which are a function of connected input neurons in adjacent layers. There are many different types of neural network.
**One Health:** an interdisciplinary approach that recognises the interconnectedness of human health, animal health, and the environment (https://www.who.int/health-topics/one-health).
**Shifting baseline syndrome:** a gradual change in the accepted norms for the condition of the natural environment because each generation perceives the environment based on their own experiences and expectations, without considering (or knowing) the historical changes that have occurred.
**Species distribution modelling (SDM):** also known as habitat suitability modelling, is a model to predict the potential distribution of a species in geographic space using georeferenced species data alongside environmental data.
**Transformer language models:** a type of neural network often used for natural language processing. Transformers are good at detecting interactions between words in text. They are pre-trained on very large corpora. Some are used for classifying text;

others generate text in response to input. BERT and GPT-n are examples.
**Word embedding:** a multidimensional vector that captures the meaning or semantics of a particular word based on analysis of its co-occurrence with other words among millions of documents.

own experiences and expectations, without considering historical changes [10]. Where written records exist for a species prior to extirpation or adaption, for example, due to **habitat conversion**, written records could offer insight into otherwise unknown habitat requirements. This information is a necessity if a species reintroduction is planned. The International Union for Conservation of Nature (IUCN) criteria for reintroduction or translocation of a regionally extinct species requires the former range and habitat of the species to be known.

The addition of habitat data to species distributions allows inference of habitat suitability and preference (**species distribution modelling**), and assessment of whether the habitat preferences of a given species have changed over time. The historical spread of species could be tracked (as assessed by [4]) by use of museum records to reconstruct how habitat played a role in the spread of a common pest species. This finding is contrary to previous research that found that roadside verge habitat contributed to successful spread of the species. Such insight to historical habitat associations could be especially fruitful to assess how non-native invasive species establish.

## Evidence of habitat in species occurrence records

In museum databases, information about the environment in which specimens were collected can be recorded with various labels. A preliminary analysis of 2.3 billion GBIF records, of observations and museum specimens, found 345 million records containing habitat-related data in one or more of the fields occurrenceRemarks, eventRemarks, fieldNotes, and habitat. Datasets differ markedly in which such fields contain information. The habitat descriptions vary greatly in level of detail and the nature of the description (Box 1). They can be also expected to vary in reliability. Most of the habitat-related records we detected are for animals (61%) and plants (28%) (see Figure I in Box 2). About 23% of all plant records had habitat data

### Box 1. Examples of descriptions of habitat in specimen collection records

Specimen collection records in natural history collections such as in museums and herbaria can include a description of the environment or habitat in which the recorded species was collected. These descriptions vary greatly in level of detail and type of information that they include, such as co-occurring species, vegetation, soil type, aspect, or hydrological features (Table I). We propose that the text of these descriptions, and other fields with information such as phenology, other specimen characteristics, co-occurring species, and elevation, can be input to a machine learning classifier to infer a standard class of habitat or land cover that that can then be used to construct historic maps of habitat (when georeferenced) and assist in studies of species–habitat interactions and associated implications for evolution.

Table I. Examples of habitat data[a]

| Examples of habitat-related text | Location, GBIF ID, species |
|---|---|
| In association with a small bryophyte mound, deep in thicket. Mosaic communities of dense heath-forming shrubs to 3 m tall, subalpine herbs and dwarf heaths to 0.5 m tall, dominated by stunted *Leptospermum scoparium* (manuka) and *Dracophyllum* and a ground tier including *Empodisma*. | New Zealand; 2828180673; *Acromastigum mooreanum* |
| Clayey bank amongst scrub and dead fallen *Pinus* branches | New Zealand; 1091305468; *Lindsaea linearis* |
| Wet, mossy subalpine mountain beech forest under loose bark of fallen beech trees | New Zealand; 2427260858; *Oopterus patulus* |
| Limber pine and Douglas fir forest with patches of sage grassland. Associated taxa: Douglas fir, limber pine, sagebrush, grass | Wyoming; 4069686930; *Townsendia parryi* |
| Grasses/juniper/rabbitbrush, dry grassy hillside | Wyoming; 2467783608; *Apamea burgessi* |
| On marshy creek banks | Wyoming; 4404950377; *Juncus bufonius* |
| With cottonwoods. Growing in loamy to gravelly soil on SE facing, 0–2% slope of riparian lowland | Wyoming; 4073770476; *Agrostis stolonifera* |
| Clay soil on basalt ledge; with *Riccardia*, *Fossombronia*; beneath *Holcus lanatus*, *Blechnum vulcanicum*. Basalt roadside ledges and *Melycytus ramiflorus Fuchsia excorticata* scrub with small stream over basalt bedrock. | New Zealand; 1091258790; *Pohlia ochii* |
| Margin of moderately forested bog savannah with creeping juniper, pitcher plant, spruce, white cedar, *Cladium*, *Phragmites*, shrubby cinquefoil, tamarack, orchids, gentians, goldenrod, other forbes & sedges | Wyoming; 2432354910; *Xestia imperita* |
| Plot is dominated by dispersed matagouri with pasture in between. On a gentle sloping alluvial fan/ river terrace. Shrubs give way to pasture to the west. Relatively low biodiversity for pasture. On Terrace. Aspect: 350 degrees. Slope: 8 degrees | New Zealand; 1091243779; *Trifolium pratense* |
| Disturbed boggy area, gravel base, moss on top. | New Zealand; 2436615306; *Euchiton lateralis* |
| Low bank of small pasture creek, water only in pools. Sandy, gravel, clay soil. | Wyoming; 1930852452; *Sporobolus airoides* |
| On warm sinter soil in a thermal barren in full sun | Wyoming; 3503303306; *Ceratodon purpureus* |

[a]The listed examples were obtained from one of four data fields in GBIF records (see text) relating to New Zealand and the USA state of Wyoming. The second column records the region, the unique GBIF identifier and the species name for the respective record containing the text. For access to these records see https://doi.org/10.15468/dl.cx4m5h.

## Box 2. AI-based habitat discovery

AI methods have the potential to automate reconstruction of past habitats, including the creation of time-specific historic habitat maps. The ability to predict land cover classes from textual data has been proven [11]. The pipeline in Figure I applies related methods to generate georeferenced habitat data. Step 1 is digitisation of records, where the bar chart indicates proportions of all records with habitat data for best represented taxa of Animals, Plants, Bacteria and Fungi. Step 2 extracts the textual descriptions of location and of habitat from digitised records. These are input to Step 3 which applies a machine learning classifier, trained to determine the association between habitat-related language and specific habitat classes (see main text for elaboration of aspects of that process). If there are no existing coordinates, the step infers coordinates from the location description using georeferencing methods. The latter currently exist but will be refined in future to interpret a wider range of spatial language. Step 4 outputs the determined habitat class and coordinates. Applications, when applied to multiple specimens for the same time period and location, include quantifying loss of particular habitats, and refining species distribution models. Example specimen from Manaaki Whenua, Landcare Research, New Zealand (https://scd.landcareresearch.co.nz/Specimen/CHR%20171955). Created with BioRender.com.

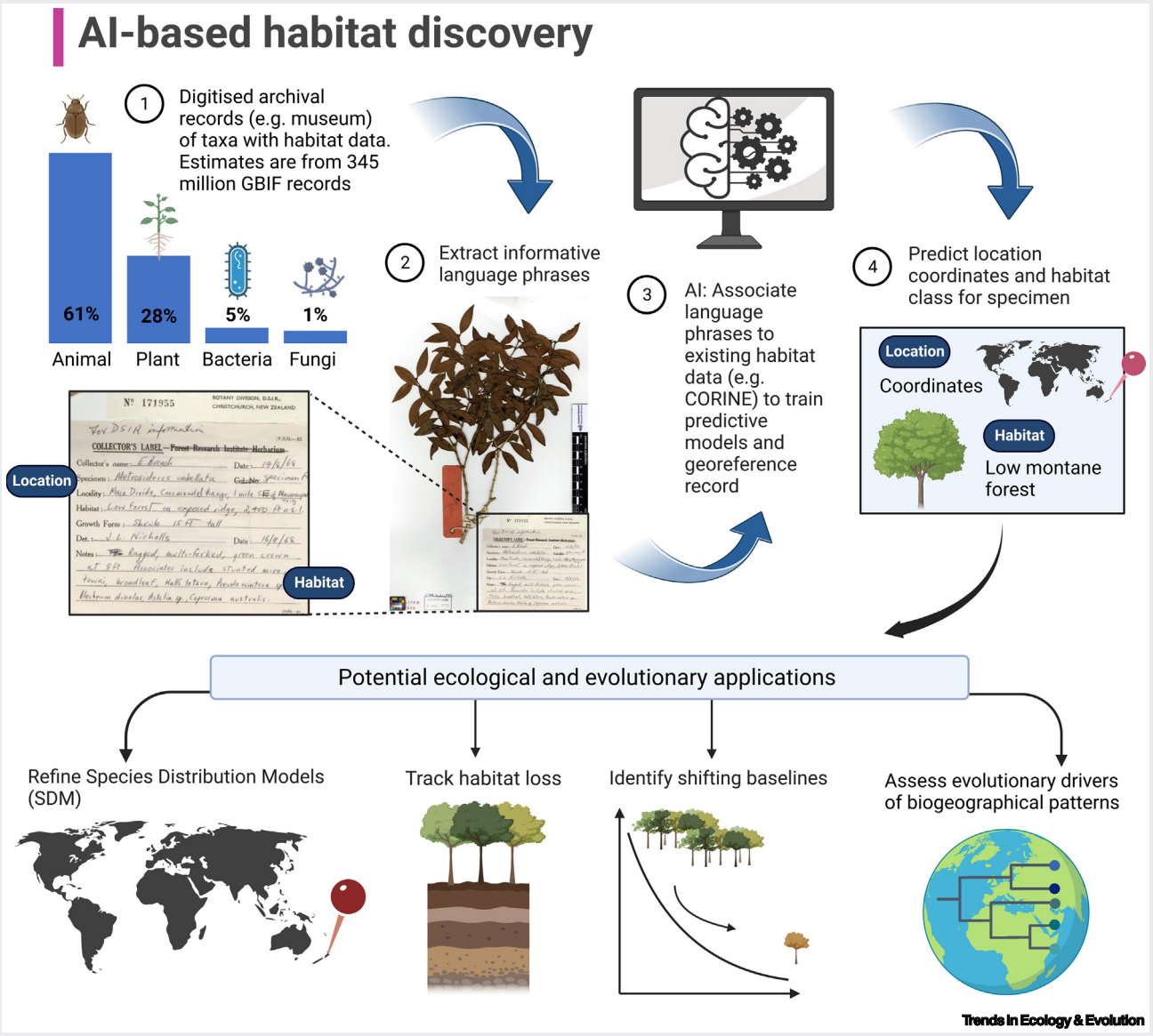


Figure I. Pipeline for generating and using habitat data. Abbreviations: AI, artificial intelligence; GBIF, Global Biodiversity Information Facility.

(11% of animal records and about 95% for archaea and bacteria) but these proportions vary regionally, being 71% for New Zealand plants. In a sample of 700 million GBIF records with habitat data about 62% of animal records were for birds, while 21% were for insects.

It is possible to distinguish several aspects of descriptions of the sampled environment that have the potential to infer habitat. These include vegetation, agriculture, geomorphology (land forms), soil and shallow (recent) geology, and older geology. Notably these descriptions commonly mention species co-occurring with the specimen (Box 2), which has potential for biodiversity studies [11]. References to vegetation include terms such as alpine grass, sphagnum, beech and *Coprosma*. Examples of agricultural descriptions include pasture, orchards, and was cropped as oats. Descriptions of geomorphology features include old meander channels, floodplain, and glacial outwash terrace. Soil type can be described by phrases such as marton loam, Kawatau stony silt loam, while geological terms include greywacke, schist and mudstone.

## Challenges in inferring habitat from museum records

While habitat-related data in museum records provide a rich resource, its automated exploitation presents challenges due to the diverse use of terminology alluded to above. Broad scale generation of maps of previous habitat will require translation of arbitrary terminology of museum records to a controlled vocabulary. It is also the case that digitisation campaigns sometimes omit the transcription of such data. In subsequent sections we propose practical solutions and recommendations for the resolution of these issues.

## How can AI help?

There is a clear route to the application of AI **machine learning classification** methods to exploit the existing wide range of habitat-related description in georeferenced biological records to infer standard classes of habitat for their respective locations (Box 2). An example of accepted habitat-related terminology is the EU **CORINE** classification. This scheme is hierarchical with 44 classes at the third, most detailed level. Each category has a number of descriptive terms such as salt marshes, burnt areas, sclerophyllous vegetation, broad-leaved forest, nonirrigated arable land, green urban areas. In practice there could be multiple terms in a specimen record to indicate that such categories apply (Box 1). When provided with appropriate **ground truth data** AI methods can learn to associate the various habitat terms, and associated species, with such a standard set of categories. This is elaborated upon below, but it should be noted that machine learning has already been demonstrated to infer land cover categories from text, as in [12] using social media.

Most state-of-the-art AI methods depend upon the use of so-called **word embeddings** that represent the meaning or semantics of a word, as a vector of numbers, based on its association with other words. When implementing a machine learning classifier, such as a **neural network**, the input to a classifier would include the word embeddings of each word within the habitat descriptions. Thus, when a descriptive word is encountered in a specimen description but has not been seen in data used to train the classifier, it can still be exploited because the classifier will recognise its embedding as having similar meaning to the words that have been used to train the classifier. **Deep learning** methods referred to as **transformer language models** (Note that current web-based question answering systems such as ChatGPT are based on transformer deep learning methods that give rise to large language models such as the various versions of GPT.) refine the embeddings, when training the classifier, according to their context and hence improve their predictive power. In addition to textual descriptions of habitat and species, collections records can contain numeric information, such as altitude, that can also be input to habitat classification models.

## Obtaining examples of ground truth

To train a habitat classifier, it is necessary to obtain example ground truth habitat data for the locations to which selected training specimen descriptions apply. Such data are readily available for many regions as a result of systematic, usually remote sensing-based, land cover mapping conducted over more than 30 years. While EU maps use the CORINE classification, in the USA, variations on the Anderson classification are often used. Advances in remote sensing continue to improve the quality of such data [13] but are limited to recent decades. Going back some hundreds of years, there are other potential sources of ground truth such as historical topographic maps, and written accounts. Old maps can portray aspects of land cover, such as woodland, water bodies, marshes [14], built-up areas, and of marine environments such as coral reefs [15].

Deep learning AI models applied to specimen habitat descriptions provide the opportunity to predict habitat classifications for historic periods not covered by existing land classification surveys, or in low resourced areas without such mapping programmes. The degree to which known land surveys can train models to predict classifications in other times and places will however need to be considered, particularly in light of the bias that can occur in sampling [1,2,7].

## The way forward

In seeking to exploit the rich resources of habitat data in collections records it will be necessary to conduct experiments to test the power of AI, using approaches such as outlined above, to create historic habitat

maps that can be used for studies such as of habitat change and loss, past species–habitat interactions and the nature of evolutionary processes. Application of such methods for periods before the publication of modern, satellite imagery-driven, land cover maps will be dependent on identifying historic maps and documents that can serve as ground truth to train the classifiers. Despite the opportunity for habitat data to create an **extended specimen,** longer term exploitation is hindered by a lack of standardisation across digitisation programmes; they can omit habitat information (as in the current digitisation project at Kew Gardens; https://www.kew.org/science/our-science/projects/digitising-kews-collection). Resources are needed to fund such digitisation, but citizen science campaigns could also help. The current, sometimes confusing, mix of habitat descriptions could be addressed by introducing protocols for application of a controlled habitat vocabulary. More research is also needed to understand and compensate for the forms of bias in collection practices. Finally, all habitat mapping requires **georeferencing** (https://docs.gbif.org/georeferencing-best-practices/1.0/en/) but more effective and efficient automated methods are needed to understand the often complex language of historic textual location descriptions.

## Acknowledgements

The authors would like to acknowledge assistance from Dr John Wieczorek who provided some data derived from GBIF records that contributed to some of the statistics on habitat data quoted in the article.

## Declaration of interests

No interests are declared.

[1]School of Computer Science and Informatics, Cardiff University, Cardiff, UK
[2]Massey Geoinformatics Collaboratory, Massey University, Palmerston North, New Zealand
[3]School of Biosciences, Cardiff University, Cardiff, UK

*Correspondence:
jonescb2@cardiff.ac.uk (C.B. Jones).

https://doi.org/10.1016/j.tree.2024.01.006



## References

1. Hedrick, B.P. *et al.* (2020) Digitization and the future of natural history collections. *BioScience* 70, 243–251
2. Meineke, E.K. *et al.* (2018) Biological collections for understanding biodiversity in the Anthropocene. *Philos. Trans. R. Soc. B Biol. Sci.* 374, 20170386
3. Davis, C.C. (2023) The herbarium of the future. *Trends Ecol. Evol.* 38, 412–423
4. Lavoie, C. *et al.* (2007) How did common ragweed (*Ambrosia artemisiifolia* L.) spread in Québec? A historical analysis using herbarium records. *J. Biogeogr.* 34, 1751–1761
5. Heberling, J.M. (2022) Herbaria as big data sources of plant traits. *Int. J. Plant Sci.* 183, 87–118
6. Fordham, D.A. *et al.* (2020) Using paleo-archives to safeguard biodiversity under climate change. *Science* 369, eabc5654
7. Daru, B.H. *et al.* (2018) Widespread sampling biases in herbaria revealed from large-scale digitization. *New Phytol.* 217, 939–955
8. Daru, B.H. *et al.* (2021) Widespread homogenization of plant communities in the Anthropocene. *Nat. Commun.* 12, 6983
9. Ringelberg, J.J. *et al.* (2023) Precipitation is the main axis of tropical plant phylogenetic turnover across space and time. *Sci. Adv.* 9, eade4954
10. Pauly, D. (1995) Anecdotes and the shifting baseline syndrome of fisheries. *Trends Ecol. Evol.* 10, 430
11. Pearson, K.D. (2018) Rapid enhancement of biodiversity occurrence records using unconventional specimen data. *Biodivers. Conserv.* 27, 3007–3018
12. Jeawak, S.S. *et al.* (2020) Predicting environmental features by learning spatiotemporal embeddings from social media. *Ecol. Inform.* 55, 101031
13. Francis, E.J. and Asner, G.P. (2019) High-resolution mapping of redwood (*Sequoia sempervirens*) distributions in three Californian forests. *Remote Sens.* 11, 351
14. Bromberg, K.D. and Bertness, M.D. (2005) Reconstructing New England salt marsh losses using historical maps. *Estuaries* 28, 823–832
15. McClenachan, L. *et al.* (2017) Ghost reefs: nautical charts document large spatial scale of coral reef loss over 240 years. *Sci. Adv.* 3, e1603155