Finding biotic anomalies described in specimen label text is a challenge that artificial intelligence can address

Austin Mast[1,2], Shubo Tian[3], Zhe He[4], Erica Krimmel[2], Fritz Pichardo-Marcano[1], Mikayla Buckley[1], Sophia Gomez[1], Ashley Hennessey[1], Allyson Horn[1], Olivia Howell[1]

Biodiversity specimen collectors are on the front lines of observing biotic anomalies, some of which herald early stages of significant changes (e.g., the arrival of a new disease; Pearson and Mast 2019). Online data sharing has opened new possibilities for the discovery of anomaly descriptions on collectors' labels, but it remains a challenge to find these needles in the haystack of many millions of specimen records available at aggregators like iDigBio and Global Biodiversity Information Facility. In a recent community survey, over 200 collectors identified 170 unique words and phrases (e.g., atypical) that they would use to describe six types of anomaly (Pearson and Mast 2019). Left unanswered was the relative efficiency with which anomaly descriptions can be found using the simple presence of these words. Here, we address that question with a focus on one type of anomaly (phenological; related to the timing of life history events) and ask a second question: can we further improve the efficiency of anomaly description discovery by engaging artificial intelligence (AI)?

We focused on six words that we expect to be used in most descriptions of phenological anomalies: early, earlier, earliest, late, later, and latest. We examined every use of those words in 50 metadata fields (those in Fig. 2 of Pearson and Mast 2019) in the 125 million records aggregated by iDigBio as of early 2022. Every text string in which a focal word occurred was independently classified by two technicians as either an anomaly description, not an anomaly description, ambiguous, or uninterpretable (e.g., in a non-English language). An example of an anomaly description is „aberrantly late flowering individual"; that of a non-anomaly is „Herbarium of the late East India Company"; and that of an ambiguity is „extremely early individual" (which could reference phenology or a portion of a life history stage). When the two technicians disagreed on a classification, Mast made a final decision.

Our six focal words appeared in 516 129 text strings in 43 of the metadata fields. Only six fields (dynamicProperties, occurrenceRemarks, eventRemarks, habitat, locationRemarks, and locality) had >10 records describing an anomaly. We reduced our focus to the 194 377 text strings that were deemed interpretable in these six high-value fields, then distilled it down further to the 110 922 unique text strings among them. We discovered that only 3 % of these unique text strings described an anomaly or potentially did so (i.e., were ambiguous).

Keywords
Anomaly detection, Artificial intelligence, Biodiversity specimens, Global change, Phenology

Addresses of the authors
[1] Department of Biological Science, Florida State University, Tallahassee FL / USA,
[2] Institute for Digital Information and Scientific Communication, Florida State University, Tallahassee FL / USA
[3] Department of Statistics, Florida State University, Tallahassee FL / USA
[4] College of Communication and Information, Florida State University, Tallahassee FL / USA

Contact
amast@fsu.edu

Accepted
25. 1. 2023

DOI
https://doi.org/10.12685/bauhinia.1374

To explore whether artificial intelligence could introduce new efficiencies to the discovery of these relatively rare anomaly descriptions, we split the data into training (63.7 %), validation (11.3 %), and test (25%) sets. We encoded the data using two alternative approaches: (1) term frequency multiplied by inverse document frequency (TF-IDF) of n-grams 1 to 5 words in length and (2) the pre-trained language model Bidirectional Encoder Representations from Transformers (BERT). We processed the TF-IDF encoding using, in turn, the XGBoost machine learning (ML) model and a deep learning (DL) model of a feedforward neural network with one hidden layer of 256 neurons, dropout, and ReLU activation function. We processed the BERT-encoding using DL alone.

Performance of all three approaches produced accuracies greater than 97 % (97.2 % for TF-IDF + ML; 97.7 % for TF-IDF + DL; and 98.6 % for BERT + DL). However, the false negative rate for the methods, where a text string classified as describing an anomaly or as ambiguous is deemed a non-anomaly by the approach, was relatively high (48.6 %, 38.0 %, and 25.1 %, respectively).

The simple presence of words likely to be used to describe phenological anomalies has a low rate of return of text describing anomalies (3 %). In contrast, our early results classifying text strings containing six other anomaly terms (aberrant, abnormal, atypical, odd, unusual, and weird) produce a much higher rate of return (> 50%), but the type of anomaly being described is less consistent. We demonstrate that artificial intelligence approaches introduce valuable efficiencies to discovery. In the most effective approach, AI finds many (75 %) of the needles (i.e., anomaly descriptions) in the haystack. This work moves us closer to being able to flag and deliver high-value anomaly descriptions to interested stakeholders as the data is shared at aggregators, a potential next step.

Acknowledgements | This work was funded by grants from the US National Science Foundation (DBI-1547229 and DBI-2027654). Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

References

Pearson KD & Mast AR (2019) Mobilizing the community of biodiversity specimen collectors to effectively detect and document outliers in the Anthropocene. American Journal of Botany 106(8): 1052–1058. doi:10.1002/ajb2.1335