Evaluating the feasibility of automating dataset retrieval for biodiversity monitoring (#101416)

First submission

Guidance from your Editor

Please submit by **28 Jun 2024** for the benefit of the authors (and your token reward) .

Structure and Criteria
Please read the 'Structure and Criteria' page for guidance.

Author notes
Have you read the author notes on the guidance page?

Raw data check
Review the raw data.

Image check
Check that figures and images have not been inappropriately manipulated.

If this article is published your review will be made public. You can choose whether to sign your review. If uploading a PDF please remove any identifiable information (if you want to remain anonymous).

Files
Download and review all files from the materials page.

6 Figure file(s)
2 Table file(s)
2 Other file(s)

Structure and Criteria

Structure your review

The review form is divided into 5 sections. Please consider these when composing your review:

1. **BASIC REPORTING**
2. **EXPERIMENTAL DESIGN**
3. **VALIDITY OF THE FINDINGS**
4. General comments
5. Confidential notes to the editor

You can also annotate this PDF and upload it as part of your review

When ready submit online.

Editorial Criteria

Use these criteria points to structure your review. The full detailed editorial criteria is on your guidance page.

BASIC REPORTING

- Clear, unambiguous, professional English language used throughout.
- Intro & background to show context. Literature well referenced & relevant.
- Structure conforms to PeerJ standards, discipline norm, or improved for clarity.
- Figures are relevant, high quality, well labelled & described.
- Raw data supplied (see PeerJ policy).

EXPERIMENTAL DESIGN

- Original primary research within Scope of the journal.
- Research question well defined, relevant & meaningful. It is stated how the research fills an identified knowledge gap.
- Rigorous investigation performed to a high technical & ethical standard.
- Methods described with sufficient detail & information to replicate.

VALIDITY OF THE FINDINGS

- **Impact and novelty is not assessed.** Meaningful replication encouraged where rationale & benefit to literature is clearly stated.
- All underlying data have been provided; they are robust, statistically sound, & controlled.
- Conclusions are well stated, linked to original research question & limited to supporting results.

Standout reviewing tips

The best reviewers use these techniques

Tip	*Example*
Support criticisms with evidence from the text or from other sources	*Smith et al (J of Methodology, 2005, V3, pp 123) have shown that the analysis you use in Lines 241-250 is not the most appropriate for this situation. Please explain why you used this method.*
Give specific suggestions on how to improve the manuscript	*Your introduction needs more detail. I suggest that you improve the description at lines 57- 86 to provide more justification for your study (specifically, you should expand upon the knowledge gap being filled).*
Comment on language and grammar issues	*The English language should be improved to ensure that an international audience can clearly understand your text. Some examples where the language could be improved include lines 23, 77, 121, 128 – the current phrasing makes comprehension difficult. I suggest you have a colleague who is proficient in English and familiar with the subject matter review your manuscript, or contact a professional editing service.*
Organize by importance of the issues, and number your points	*1. Your most important issue* *2. The next most important item* *3. …* *4. The least important points*
Please provide constructive criticism, and avoid personal opinions	*I thank you for providing the raw data, however your supplemental files need more descriptive metadata identifiers to be useful to future readers. Although your results are compelling, the data analysis should be improved in the following ways: AA, BB, CC*
Comment on strengths (as well as weaknesses) of the manuscript	*I commend the authors for their extensive data set, compiled over many years of detailed fieldwork. In addition, the manuscript is clearly written in professional, unambiguous language. If there is a weakness, it is in the statistical analysis (as I have noted above) which should be improved upon before Acceptance.*

Evaluating the feasibility of automating dataset retrieval for biodiversity monitoring

Alexandre Fuster-Calvo Corresp., [1] , **Sarah Valentin** [2] , **William C. Tamayo** [1] , **Dominique Gravel** [1]

[1] Biology department, University of Sherbrooke, Sherbrooke, Quebec, Canada

[2] Joint Research Unit Land, Remote Sensing and Spatial Information (UMR TETIS), French Agricultural Research Centre for International Development (CIRAD), Montpellier, France

Corresponding Author: Alexandre Fuster-Calvo
Email address: alexfuster7@gmail.com

Aim. Effective management strategies for conserving biodiversity and mitigating the impacts of Global Change rely on access to comprehensive and up-to-date biodiversity data. However, manual search, retrieval, evaluation, and integration of this information into databases presents a significant challenge to keep pace with the rapid influx of large amounts of data, hindering its utility in contemporary decision-making processes. The automation of these tasks through advanced algorithms holds immense potential to revolutionize biodiversity monitoring. **Innovation.** In this study, we investigate the potential for automating the retrieval and evaluation of biodiversity data from Dryad and Zenodo repositories. We employ automated algorithms to identify potentially relevant datasets and perform a manual assessment to gauge the feasibility of automatically ranking their relevance. We have designed an evaluation system based on various criteria. Additionally, we compare our results with those obtained from a scientific literature source, using data from Semantic Scholar for reference. Our evaluation centers on the database utilized by a national biodiversity monitoring system in Quebec, Canada. **Main conclusions.** The algorithms retrieved 90 (56%) relevant datasets for our database, showing the value of automated dataset search in repositories. Additionally, we find that scientific publication sources offer broader temporal coverage and can serve as conduits guiding researchers toward other valuable data sources. However, our manual evaluation highlights a significant challenge to distinguish datasets by their relevance—scarcity and non-uniform distribution of metadata, especially pertaining to spatial and temporal extents. We present an evaluative framework based on predefined criteria that can be adopted by automated algorithms for streamlined prioritization, and we make our manually evaluated data publicly available, serving as a benchmark for improving classification techniques. Finally, our study advocates for the implementation of metadata standards tailored for automated retrieval systems by repositories and sources of scientific

literature. This, coupled with the rapid evolution of classification algorithms, holds transformative potential to advance in biodiversity monitoring and decisively steering the course of well-informed decision-making processes.

Evaluating the feasibility of automating dataset retrieval for biodiversity monitoring

Alexandre Fuster-Calvo[1], Sarah Valentin[2] , William C. Tamayo[1], Dominique Gravel[1]

[1] Département de biologie, Université de Sherbrooke, Sherbrooke, Québec, Canada
[2] Joint Research Unit Land, Remote Sensing and Spatial Information (UMR TETIS), French Agricultural Research Centre for International Development (CIRAD), Montpellier, France

Corresponding Author:
Alexandre Fuster-Calvo[1]
C/ Dona Amalia 44, Alcoi, Alicante, 03801, Spain
Email address: alexfuster7@gmail.com

Abstract

Aim. Effective management strategies for conserving biodiversity and mitigating the impacts of Global Change rely on access to comprehensive and up-to-date biodiversity data. However, manual search, retrieval, evaluation, and integration of this information into databases presents a significant challenge to keep pace with the rapid influx of large amounts of data, hindering its utility in contemporary decision-making processes. The automation of these tasks through advanced algorithms holds immense potential to revolutionize biodiversity monitoring.

Innovation. In this study, we investigate the potential for automating the retrieval and evaluation of biodiversity data from Dryad and Zenodo repositories. We employ automated algorithms to identify potentially relevant datasets and perform a manual assessment to gauge the feasibility of automatically ranking their relevance. We have designed an evaluation system based on various criteria. Additionally, we compare our results with those obtained from a scientific literature source, using data from Semantic Scholar for reference. Our evaluation centers on the database utilized by a national biodiversity monitoring system in Quebec, Canada.

Main conclusions. The algorithms retrieved 90 (56%) relevant datasets for our database, showing the value of automated dataset search in repositories. Additionally, we find that scientific publication sources offer broader temporal coverage and can serve as conduits guiding researchers toward other valuable data sources. However, our manual evaluation highlights a significant challenge to distinguish datasets by their relevance—scarcity and non-uniform distribution of metadata, especially pertaining to spatial and temporal extents. We present an evaluative framework based on predefined criteria that can be adopted by automated algorithms for streamlined prioritization, and we make our manually evaluated data publicly available, serving as a benchmark for improving classification techniques. Finally, our study advocates for the implementation of metadata standards tailored for automated retrieval systems by repositories and sources of scientific literature. This, coupled with the rapid evolution of classification algorithms, holds transformative potential to advance in biodiversity monitoring and decisively steering the course of well-informed decision-making processes.

Introduction

Biodiversity is undergoing rapid and unprecedented transformations driven by the relentless forces of anthropogenic Global Change (Parmesan & Yohe, 2003; Millennium Ecosystem Assessment, 2005; Newbold et al., 2015; IPBES, 2019; Pyšek et al., 2020). These changes pose a direct threat to the delicate balance of ecosystems and therefore multitude of species that inhabit them, including humans. Recognizing the urgency of the situation, the recent fifteenth meeting of the Conference of the Parties (COP 15) witnessed a landmark moment with the adoption of the Kunming-Montreal Global Biodiversity Framework (GBF) (CBD, 2023). This ambitious framework has set forth action-oriented global targets aimed at urgently bending the curve of biodiversity loss by 2050 (Leadley et al., 2022).

Target 21 of the CBD 2023 underscores the critical importance of providing decision makers, practitioners, and the public with access to comprehensive data, information, and knowledge. This necessity arises because the pursuit of these conservation objectives, data from various ecological disciplines and origins must be seamlessly integrated, spanning an extensive spectrum of spatiotemporal scales (Kelling et al., 2009; Wieczorek et al., 2012; Hampton et al., 2013; Heberling et al., 2021). Achieving this goal relies on two fundamental pillars. Firstly, a robust global framework must be established to systematically collect, standardize, harmonize, and provide timely data on the ever-changing landscape of biodiversity. The Group on Earth Observations Biodiversity Observation Network (GEO BON) has played a pivotal role in this regard by developing the concept of Essential Biodiversity Variables (EBVs) (Pereira et al., 2013). These EBVs now serve as the cornerstone of monitoring programs worldwide, facilitating the quantification of biodiversity changes across diverse ecosystems (Vihervaara et al., 2017; Schmeller et al., 2018; Jetz et al., 2019).

Secondly, to effectively track and address the challenges posed by biodiversity change, the development of comprehensive global databases and sophisticated bioinformatic systems becomes essential (Collen & Nicholson, 2014). These databases and tools are tasked with the colossal mission of collecting, cataloging, integrating, and meticulously analyzing vast volumes of datasets derived from disparate sources. However, despite remarkable strides in the establishment of global macro-ecological databases [e.g. genetic and phylogenetic data (GenBank), species interactions (GloBI, www.globalbioticinteractions.org), traits (TRY Plant Trait Database, www.try-db.org), or abundance (BioTime - Dornelas et al., 2018)] and georeferenced information infrastructures that use them to monitor biodiversity [e.g., Group on Earth Observations–Biodiversity Observation Network (GEO-BON) initiative, www.geobon.org; Global Biodiversity Information Facility, www.gbif.org], significant impediments persists. It is still difficult to find historical and contemporary biodiversity data, particularly for less-studied taxa and less-explored regions (Jetz et al., 2012; Conde et al., 2019). The available data, though valuable, falls short of providing a comprehensive overview of the state and dynamics of biodiversity on a global scale. This limitation poses a significant impediment to the advancement of our understanding of biodiversity and, consequently, its conservation (Hortal et al., 2015).

A major impediment for achieving comprehensive databases is the ever-increasing volume of data published annually (Hendriks & Duarte, 2008; Stork & Astrin, 2014). Managing and staying current with this expanding wealth of information becomes increasingly challenging. This is largely because the labor-intensive nature of locating and evaluating the pertinence of data within scientific publications for integration into global databases remains a predominantly manual process (Guralnick & Hill, 2009; Wen et al., 2017). The task is further exacerbated by the rapid acceleration in research output, rendering it increasingly impractical to maintain real-time coverage.

In addressing the pressing issue of data scarcity in biodiversity studies, the development of automated systems capable of identifying relevant datasets from diverse sources may well mark a pivotal turning point. These systems bridge text-mining techniques with the interdisciplinary field of Natural Language Processing (NLP) in computer science, integrating methodologies from linguistics, computer science, statistics, and artificial intelligence. Toolsets commonly employ frequency analysis, rule-based algorithms, or artificial intelligence methods (Farrell et al., 2024, preprint).

While automatic content analysis is still in its incipient stages for ecological studies (Nunez-Mir et al., 2016), it has already demonstrated success in streamlining literature reviews (Heberling et al., 2019; McCallen et al., 2019), retrieving fossil data (Kopperud et al., 2019), monitoring data for endangered species (Kulkarni & Minin, 2021), and detecting species co-occurrences and interactions (Farrell et al., 2022, preprint) nford et al., 2020 showcased the effectiveness of such approaches in identifying relevant articles for specific databases. Their research highlights the capability of algorithms, trained using data from two distinct databases, to analyze a collection of articles. Impressively, these algorithms can discern between relevant and irrelevant articles for these databases with an accuracy rate exceeding 90%, all based solely on the content of titles and abstracts. Recently, prompt-based approaches leveraging Large Language Models such as GPT have demonstrated promising effectiveness in extracting biodiversity data, offering further advancements in automated systems for biodiversity research (Castro et al., preprint).

While these findings mark significant progress, further refinements are essential to realize truly effective automatic retrieval systems. Firstly, an automated system should extend beyond merely distinguishing between relevant and non-relevant publications; it should also have the capacity to assign varying degrees of relevance based on metadata, hence aiding in prioritizing the most pertinent studies and expediting their integration into databases. Secondly, relying solely on information from titles and abstracts may lead to either over- or underestimating a publication's relevance. If a combination of different features determines the degree of relevance of a publication for a database, it may be necessary to search for these in different sections (article text, tables, supplementary materials, dataset files…). Lastly, it is worth noting that an increasing number of scientific journals now mandate authors to make their data publicly available without restrictions in online repositories upon publication. This shift means that retrieving data could be significantly streamlined by searching these repositories, where data is readily accessible (Fig. 1).

Here, we use automated retrieval of datasets from both Dryad and Zenodo repositories and embark on a comprehensive manual evaluation with the primary objective of assessing the potential for automated classification into various relevance tiers. e same time, we introduce a versatile classification framework engineered to enable algorithms to allocate relevance levels based on features extracted from the publication text. This adaptable system considers global-level attributes of biodiversity data, facilitating its seamless integration with a diverse range of databases. Furthermore, it remains flexible enough to readily accommodate additional parameters customized to suit specific database contexts. Importantly, this classification system not only addresses current challenges but also serves as a foundational baseline for next-generation algorithms, providing a framework for iterative improvements and refinements in automated literature classification. As the assignment of relevance levels necessitates a thorough analysis of the features, algorithms must be adept at identifying these features. Consequently, our investigation delves into the distribution of this metadata across various publication locations, including the title, abstract, repository text, article, and dataset. This exploration holds significant implications for the design of effective search algorithms. Furthermore, we extend our investigation to encompass datasets sourced from articles available through the Semantic Scholar platform and illuminate the strengths and limitations of data derived from articles when compared to repository-sourced data. Finally, we discuss significant challenges and promising opportunities that lie on the horizon in the quest for reliable and efficient automated systems for biodiversity data regardless of the technological sophistication of future algorithms.

Case study - biodiversity monitoring in Quebec

Canada, as the second-largest country, possesses an extensive array of species and ecosystems, all significantly impacted by human activities. With over 841 species at risk, including 371 classified as Endangered (COSEWIC, 2022), it is concerning that 59% of these species are experiencing population declines (WWF Canada, 2020), and that crucial habitats such as wetlands, which cover 14% of the territory and represent 25% of the world's reserves, suffered significant loss (Environment Canada, 2009). Canada's vast geographical expanse, coupled with its diverse ecoregions and extensive wilderness areas, underscores the urgency of developing prioritization strategies for conserving critical habitats and species. Such strategies are imperative for Canada to align with international biodiversity commitments outlined in the Kunming-Montreal Global Biodiversity Framework (CBD, 2023).

The province of Quebec has recently introduced a dedicated national geographic information system aimed at biodiversity monitoring, known as Biodiversité Québec. Our study is tailored to evaluate the integration of relevant data into this information system, which is designed to play a pivotal role in influencing decision-making processes. Additionally, by situating our research within this dataset, we can showcase the design of specific criteria that are not only globally applicable but also regionally relevant, addressing the unique circumstances at regional scale. In the context of Quebec, and more broadly in Canada, biodiversity records are particularly affected by spatial bias, correlated to the South-North human population density gradient. Thus, tundra-type ecosystems and northern territories are very poorly documented and data from these regions should have higher priority. This information system leverages the ATLAS data infrastructure to integrate and provide access to biodiversity data for the Province of Quebec. It standardizes various data types (abundances, occurrences, surveys, population time-series, and taxonomy) for integration into monitoring and modeling workflows. Data is sourced from open science repositories (e.g., GBIF, eBird, iNaturalist, Living Planet Database) and direct partnerships with local and national organizations. The infrastructure is undergoing active expansion, currently aggregating over 53 million occurrences, covering over 23 thousand species from 616 data sources.

Methods

Corpus retrieval

We retrieved data from Dryad and Zenodo repositories according to two criteria: (1) used in the ecology/biodiversity domain and (2) have an API, allowing their automatic retrieval. Because Zenodo now hosts a preservation copy of Dryad datasets, it enables the retrieval of datasets from both repositories from one single API. We built functions to interact with a corresponding API, iterating through search queries to determine if all keywords within a query are found within the repository page, including the title, abstract, and metadata. Subsequently, specific information is extracted from each record, such as title and keywords, and stored in a structured format. The queries, executed in December 2022, were formulated using Boolean "and" operators between keywords (Fig. 2). Additionally, the search utilized a Zenodo filter for publications of resource type “dataset”.

Dataset annotation

We followed pre-established annotation guidelines for each dataset. These guidelines encompassed all variables of interest and criteria for assigning feature categories (Table 1; refer to the explanations below, and consult the complete annotated dataset in the Supplementary Materials).

Dataset relevance

Datasets were categorized as "High," "Medium," "Low," or "Negligible" in terms of relevance. Our classification system was founded on different criteria, which we divided into two groups. The first group, termed Main Classifiers, are thought to capture universally key features of biodiversity data. These encompass data type, temporal and spatial extent, and data size. This categorization was informed by the literature on Essential Biodiversity Variables (EBVs) and the significance of temporal and spatial dimensions. For example, temporal duration was categorized as follows: <3 years as Low, 3 to 10 years as Moderate, and >10 years as High (as detailed in Table S1). "High" relevance datasets typically featured extensive data sizes, highly relevant data types (e.g., abundance), and substantial spatial or temporal scales. "Moderate" relevance datasets exhibited either large data sizes or highly relevant data types, combined with low to moderate temporal and spatial extents or vice versa, or displayed moderate characteristics across all main classifiers. "Low" relevance datasets were characterized by moderate data sizes and very short temporal and spatial scales or vice versa, or held low ratings across all main classifiers. "Negligible" datasets contained no valuable information generally for biodiversity nor for the ATLAS database (Table 1).

The second group encompassed Modulators—features that, while not as critical as Main Classifiers, also inform about the importance of the data. These can include aspects at a more regional scale, such as a first record for a given taxa in a region, or data from undersampled areas. In the context of Quebec biodiversity monitoring strategies, data from northern regions is prioritized as these are largely unstudied areas, and therefore a modulator that we introduce is the sampling bias north-south Quebec. This illustrates that modulators can accommodate needs of specific databases. Modulators influence the evaluation of datasets by slightly altering the relevance category assigned based on Main Classifiers. Modulators had the capacity to shift the category by one level, for instance, a dataset assigned "Low" relevance that contains data from northern regions of Quebec could change to "Moderate" (Table 2).

Dataset features

We manually identified dataset features essential for evaluating data type categories, spatiotemporal extent, taxon numbers and identities, and other specific attributes (Table 1). Data type categories were assigned following the EBV data type classification. Notably, we marked instances when data were time series, as these held unique significance.

Features could be found in either the abstract or additional text within the repository page (referred to as repository text), the source publication text (i.e., the article), or within the dataset itself. We made these distinctions because the location could influence the feasibility of automated retrieval. To achieve this, we conducted manual searches for features and recorded their locations. It is important to note that features might occasionally be intertwined with content unrelated to the data, a situation particularly prevalent in references to EBV data categories or synonyms (see Table S2). For instance, the term "abundance" might appear in texts referring to species abundances or unrelated non-biological content. Therefore, we assessed the location of features referring to EBV data categories by quantifying their frequency in titles or abstracts using the "str_detect" function from the "stringr" R package.

Geospatial data within the datasets were presented in various forms, ranging from highly detailed to less detailed. We classified this information into the following categories, ranked from more to less detailed: sample, site, or range coordinates, species distribution models (SDMs; thereafter “distribution”), geographic features (e.g., Mont Mégantic), administrative units, maps, and site IDs (where sampling sites were identified, but precise locations were known only to the authors).

Extension to scientific articles

We expanded our dataset retrieval beyond repositories to include articles sourced from Semantic Scholar, covering the period from 1980 to 2022. To ensure a manageable analysis, we limited the retrieved publications to a maximum of 50 positives (top 50 found publications). Due to potential API request limitations, some queries retrieved more publications than others, with an average of 23 publications per query. Subsequently, we assessed the relevance of the datasets and publication years in the same manner as with repositories, facilitating comparisons between these two sources. Our approach generates a random sampling among queries resulting in uneven numbers, like those retrieved from repositories. This is optimal for our objectives as we are interested in comparing temporal depth and the location of information, while broadly assessing their performance for relevance. For a more stringent comparison, standardizing the sampling numbers among queries and conducting statistical tests for comparisons would be necessary.

Scientific articles may not necessarily reference a repository for presenting their data, especially older articles. Instead, data may be embedded in tables within the article text or included in the appendix. Recognizing the prevalence of datasets presented in this manner is vital for devising future strategies for automated retrieval. Consequently, we noted the location of each dataset within retrieved articles.

Moreover, articles may not always present their own data but may use data from other sources, necessitating reference. This additional information can be invaluable for identifying other sources of relevant datasets. To gauge the extent to which we retrieved articles that referred to pertinent datasets, we noted instances where articles used data from other dataset sources, particularly those of high relevance.

Evaluation

Queries

We analyzed the performance of each query by noting the amount of retrieved, not found, and not accessible datasets and their relevance categories. We computed an F-score for each query calculating a precision and recall metric as follows:

$$F\ score = 2 \cdot \frac{Precision\ \cdot Recall}{Precision + Recall}$$

where:

$$Precision = \frac{True\ Positive}{True\ Positive + False\ positive} \qquad Recall = \frac{True\ Positive}{True\ Positive + True\ negative}$$

Positive refers to those datasets assigned a high or moderate relevance, and negative to low or negligible relevance.

Features and accessibility

Our assessment encompassed both the content and accessibility of datasets. To gauge the spatiotemporal extent and the representation of Essential Biodiversity Variable (EBV) categories within the datasets, we conducted feature frequency analyses. In terms of spatial extent, we quantified the frequencies of datasets falling within the low (< 5,000 km²), moderate (5,000-15,000 km²), and high (>15,000 km²) spatial range categories. For datasets containing both temporal and spatial information, we conducted visual inspections to discern the alignment between dataset duration, spatial range, and their assigned relevance categories.

In parallel, we assessed the accessibility of features, a crucial factor for automated retrieval. To this end, we tallied the occurrences of feature locations within the datasets, distinguishing between repository text, articles, and dataset contents, for dataset type, temporal, spatial, and taxon features. Additionally, we cataloged the frequency of occurrences for each geospatial information category to gain insights into the level of detail provided in this regard.

Semantic scholar - repositories comparison

We conducted a comprehensive comparison between Semantic Scholar and the repositories, examining various aspects. This entailed evaluating the relevance, number, and accessibility of datasets obtained from both sources. To delve deeper into the temporal dimension, we quantified the frequencies of publication years for the datasets.

Results

Datasets annotation evaluation

Out of the initial 161 datasets retrieved through our queries, 55 were subsequently excluded for various reasons: 37 due to incorrect locations, 5 categorized as laboratory studies, and 13 for miscellaneous reasons. Notably, many datasets with incorrect locations resulted from the matching of the keyword "Quebec" with the affiliations of the authors.

The classification based on relevance yielded 90 relevant datasets categorized as either highly, moderately, or low, which we will refer to as “relevant datasets”: 20 datasets (18%) categorized as highly relevant, 33 (31%) as moderately relevant, 37 (35%) as having low relevance, and 16 (15%) as negligible in relevance.

Queries

The most simple query, “species”, is the one showing the highest performance (Fcore = 0.33), followed by “population + species” (Fscore = 0.23) and “sites + species” (Fscore = 0.17). “Occurrence + species” was the query with the highest precision (0.44) (Fig. 2). The mean overlap between queries was 11% (Fig. S1).

Features

Among the relevant datasets, presence-only data emerged as the most common EBV data category, encompassing 30 publications (29%), followed by genetic data with 27 (26%) and abundance with 26 (25%). In contrast, data from species distribution models (SDMs), referred to as "distribution" data, was the least common, featured in only 3 publications (3%) (Fig. 3*C*). The temporal ranges of these datasets span from the 1930s to the present, although the majority fall within the last two decades (Fig. 3*A*). Outliers include Favret et al., 2020, offering data on Odonata specimens from various entomological collections dating back to 1875, and Schumacher et al., 2022, providing pollen records for butternut spanning from 20,000 years ago to the present. On average, the temporal duration was 11.1 years, ranging from less than 1 year to 50 years (Fig. 3*B*). Short-term studies, less than 1 year in duration, constituted the most common category, accounting for 12% of the datasets. A total of 12 datasets (13%) contained time series data (Fig. 3*A*). Spatial extents varied widely, ranging from 0.2 to 24.706.834 km², with a mean of 1,388,738 km²: 26 datasets (29%) covered less than 5,000 km², 9 (10%) fell between 5,000 and 15,000 km², and 30 (34%) exceeded 15,000 km² (Fig. 3*BD*).

The datasets cover a diverse array of taxa, spanning 13 distinct classes (not counting those within zooplankton). Among these, 68 (76%) datasets were associated with one to ten species, with mammals (21), fish (13), birds (11), and angiosperms (10) being the most frequently represented. In contrast, 30 (34%) datasets pertained to communities comprising 10 to 180 species, with an average of 49 species per dataset. Notably, datasets of this nature were more prevalent for plants (12 datasets) and insects (7 datasets) (Fig. 3*E*).

Features accessibility

Within the 90 relevant datasets, at least one comprehensive metadata (features) belonging to Main Classifiers were automatically accessible for 88 of them (98%), typically within the repository page's abstract or additional text. These encompassed explicit mentions of temporal range in 25 publications (27%), temporal duration in 14 publications (15%), and spatial range in 7 publications (8%) (Fig. 4*A*). For species-level studies (involving 1 to 10 species), species names were consistently present in the title or abstract. Additionally, dataset types or synonyms were included in the title for 24 of them (27%) and in the abstract for 80 (89%).

Only one publication contained all these features together in the repository text. Furthermore, a total of 5 (6%) did not explicitly report the temporal range, 31 (34%) the spatial range, and 62 (69%) the data duration in any location (repository text, source article, appendix, or dataset) (Fig. 4*A*).

Geospatial information, essential for biodiversity monitoring, was unavailable within the repository text in 68 publications (76%), and no location data, including the dataset, was provided for 33 (37%) of them (Fig. 4*B*). Site IDs were given in the dataset of 24 (27%) publications, but in most cases, it required consulting a map in the article, with no specific coordinates, to interpret them.

Semantic scholar

Our queries yielded 254 datasets from Semantic Scholar without overlap, of which 60 were excluded due to incorrect locations (33) and miscellaneous reasons (27). Classification by relevance resulted in 3 (2%) high, 25 (13%) moderate, and 41 low (21%), and 11 (6%) negligible relevant datasets, alongside 28 (15%) inaccessible datasets and 85 (44%) publications without datasets (Fig. 5*A*). A comparison of publication years between Dryad and Zenodo versus Semantic Scholar revealed that retrieving from repositories yielded datasets published from 2010 onwards, while retrieving from Semantic Scholar included older datasets dating back to 1981 (Fig. 5*B*).

Furthermore, while one highly relevant dataset, the Neotoma Paleoecology Database (www.neotomadb.org), was referenced in publications extracted from repositories, we identified a total of 6 highly relevant datasets referenced in articles retrieved from Semantic Scholar. These datasets originated from sources such as the Canadian National Forest Inventory, the Canadian Wildlife Service, and the Québec Ministry of Environment and Wildlife.

Discussion

Our findings underscore the significant value of automating the retrieval of biodiversity data from repositories, which can substantially augment the volume of pertinent information within databases. We have introduced a classification system, designed to serve as a logical framework for automated algorithms, expediting the evaluation process by categorizing data based on its relevance. This system draws upon globally applicable biodiversity classifiers, making it adaptable to various data types, while also allowing for the incorporation of dataset-specific nuances. We publish our high quality-annotated dataset alongside this paper, with the aim of providing a benchmark for new classifiers.

Assigning relevance categories, whether through our system or alternative approaches, necessitates a meticulous analysis of features (i.e. metadata) within the publication text. Our study highlights that this task poses a considerable challenge for automated processes, often stemming from the absence or scarcity of these features and their sparsity across different sections of the publication (repository page, dataset, article, supplementary materials). This challenge is particularly pronounced in the case of spatio-temporal features, which are pivotal for guiding relevance assessments. Furthermore, our study demonstrates that repositories present a valuable source of readily accessible, publicly available data, surpassing scientific articles in terms of speed and efficiency for data retrieval. However, we also emphasize the significance of designing automated processes for data extraction from articles. These offer substantial temporal depth of publications, and serve as gateways that can guide researchers to other pertinent and valuable data sources.

By employing simple search queries consisting of one to three words, we achieved the retrieval of a substantial number of pertinent datasets, a notable percentage of which fell within the highly relevant category. Remarkably, a high percentage of datasets contained genetic data, an area where greater collection efforts have been advocated (Hoban et al., 2021; Hoban et al., 2022). Moreover, following presence-only data, abundance data was the most frequently encountered, being information that can aid in constructing time-series and elucidating population trends. These are promising findings for the repositories' potential as pivotal resources for automating the retrieval of biodiversity data. Notably, the growing trend among scientific journals to mandate the deposition of data used in publications into online repositories reflects a progressive move toward fostering openness in science. Consequently, repositories are anticipated to witness exponential growth, encompassing an ever-expanding volume of published data.

We have devised a comprehensive classification system aimed at assigning relevance categories to datasets, grounded in a set of criteria. We advocate for the adoption of similar schemes by automated dataset detection algorithms to aid in prioritizing datasets and maintaining pace with the relentless surge in published data. A critical aspect of our system involves the distinction between criteria as either Main Classifiers or Modulators: the former encompass fundamental, overarching aspects pertinent to biodiversity data on a global scale, while the latter encompass secondary criteria that retain significance, potentially at regional scales. This distinction has enabled us to underscore the vital importance of data concerning plants and animals in boreal (e.g. Lait et al., 2013; Thiffault et al., 2016; Martin et al., 2022) and arctic ecosystems (e.g. Leblond et al., 2017; Lamarre et al., 2018; Chagnon et al., 2021), which face heightened vulnerability and remain underexplored in Quebec, along with the inclusion of new species (Anderson et al., 2016).

Nevertheless, it is worth noting that the development of a universally accepted and standardized classification system for dataset relevance could itself be a substantial undertaking, warranting collaborative efforts from a multitude of experts. Such a system, once established, could serve as a common benchmark for dataset retrieval projects and should be periodically revisited and incorporated into the Essential Biodiversity Variables framework to ensure its enduring relevance and utility.

Our assessment, however, has revealed a significant challenge in accessing the metadata essential for automated algorithms to assign categories of relevance to publications. This challenge stems from the absence and dispersion of critical features throughout various sections of the publication, making it particularly problematic for detecting Main Classifiers related to temporal and spatial extents. In the majority of datasets we examined, these details were either entirely missing or exclusively located within the article text, neglecting inclusion in the abstract or repository page. Search algorithms should then be engineered to thoroughly scan all parts of a publication, encompassing the dataset itself, to capture these essential features.

This underscores the urgency of developing general and standardized frameworks within publication guidelines to supply the requisite metadata for automatic detection and extraction of information. Presently, various global frameworks and initiatives, such as the FAIR Data Principles, GBIF guidelines and standards, and Biodiversity Data Journal recommendations, advocate for accompanying data with comprehensive metadata, encompassing methodological details, temporal scope, geographic coverage, and more. However, the manner in which this information is presented poorly reflects information needs and is the biggest obstacle in retrieving relevant biodiversity data (Jones et al., 2019; Löffler et al., 2021).

For the establishment of an automated framework for biodiversity data retrieval, it becomes imperative to take an additional stride. Contemporary formats of online publications should incorporate structures that facilitate access for automated algorithms to evaluate data relevance based on standardized global criteria. This can be achieved by incorporating dedicated sections within publications, both in online repositories and articles, where authors are required to provide a predefined set of metadata, including the Main Classifiers. Implementing such straightforward updates would effectively alleviate the primary challenges we encountered regarding the identification of features necessary for assessing data relevance.

Our findings highlight that repositories like Dryad and Zenodo, as well as scientific literature search engines such as Semantic Scholar, offer distinct advantages and disadvantages when it comes to automated data retrieval. Sources of scientific literature may exhibit a relatively high percentage of publications that lack data or render it inaccessible as it was observed for Semantic Scholar, which can potentially pose challenges when retrieving datasets and assessing their relevance. Conversely, repositories offer readily accessible datasets and prove to be highly efficient sources for data retrieval. However, repositories may exhibit a limited temporal depth of datasets, typically spanning only the past decade, given their contemporary nature. In contrast, sources of scientific literature have the potential to provide a more extensive historical dataset archive. Moreover, articles may make reference to and cite other data sources, such as the six distinct highly relevant databases we found referenced in retrieved articles, thereby facilitating the identification of unknown relevant datasets. These substantial differences underscore the importance of conducting data retrieval from both types of sources to ensure a comprehensive approach.

In the realm of automated algorithm development, manual evaluations, such as the one conducted in this study (publicly available; see Data Availability Statement), will remain invaluable. They serve both as training data for automated algorithms to learn identifying relevant datasets, and as a crucial benchmark that enables the assessment of automated processes by drawing comparisons with the results from manual evaluations. Furthermore, it's imperative to acknowledge the need for specialized strategies when dealing with diverse data sources. For instance, one we did not assess here are the increasingly digitized collections of museums, which might require specific automated search and evaluation approaches. An exciting avenue for future research lies in these sources, which harbor invaluable historical data often absent from contemporary global information systems or databases (Graham et al., 2004; Guralnick et al., 2007; Page et al., 2015; Wen et al., 2015).

While the development of literature classification algorithms is advancing vertiginously, especially with AI systems (see Google Gemini, www.deepmind.google), it is imperative to recognize that the challenges surrounding information structure and metadata organization within scientific literature persist regardless of technological evolution. The effectiveness of automated systems relies not only on the sophistication of algorithms but also on the clarity and consistency of metadata standards, the accessibility of data repositories, and the interoperability of databases. Our work not only identifies significant challenges for forthcoming automated algorithms tasked with dataset retrieval for biodiversity but also underscores the relatively surmountable nature of these challenges. These foundational issues transcend the current state of AI technology and are central to ensuring the long-term viability and utility of automated systems for biodiversity data retrieval. As such, efforts to address these structural challenges as well as standardized schemes for prioritization such as the one we proposed, must remain a priority alongside advancements in AI capabilities. To move forward effectively, prioritizing the establishment of global frameworks and guidelines that streamline the workflow for automated data retrieval systems is essential. Simultaneously, raising awareness among scientists about the importance of publishing data in formats conducive to automatic retrieval and evaluation holds great promise. These initiatives carry the transformative potential to reshape our data acquisition capabilities profoundly, greatly bolstering our capacity for biodiversity monitoring and informed decision-making, with far-reaching positive ecological implications.

Acknowledgments

We thank Vincent Beauregard for his feedback on the algorithms and the ATLAS database.

References

Anderson, Frances; Lendemer, James C. (2016). Data from: *Aspicilia bicensis* (Megasporaceae), a new sterile, pustulose lichen from eastern Canada [Dataset]. *Dryad*. https://doi.org/10.16.../0007-2745-119.1.008

Castro, A., Pinto, J., Reino, L., Pipek, P., & Capinha, C. (2024). Large language models overcome the challenges of unstructured text data in ecology. *bioRxiv*, 2024-01.

Chagnon, Catherine; Simard, Martin; Boudreau, Stéphane (2021). Patterns and determinants of lichen abundance and diversity across a subarctic to arctic latitudinal gradient [Dataset]. *Dryad*. https://doi.org/10.5061/dryad.9w0vt4bg8

Collen, B., & Nicholson, E. (2014). Taking the measure of change. *Science*, 346(6206), 166–167. http://dx.doi.org/10.1126/science.1255772

Conde, D. A., Staerk, J., Colchero, F., da Silva, R., Schöley, J., Baden, H. M., ... & Vaupel, J. W. (2019). Data gaps and opportunities for comparative and conservation biology. *Proceedings of the National Academy of Sciences*, 116(19), 9658-9664. http://dx.doi.org/10.1073/pnas.1816367116

Convention on Biological Diversity (CBD). Kunming-Montreal Global Biodiversity Framework; United Nations Environment Programme: Montreal, QC, Canada, 2023.

Cornford, R., Deinet, S., De Palma, A., Hill, S. L., McRae, L., Pettit, B., ... & Freeman, R. (2021). Fast, scalable, and automated identification of articles for biodiversity and macroecological datasets. *Global Ecology and Biogeography*, 30(1), 339-347. http://dx.doi.org/10.1111/geb.13219

Dornelas, M., Antao, L. H., Moyes, F., Bates, A. E., Magurran, A. E., Adam, D., ... & Murphy, G. (2018). BioTIME: A database of biodiversity time series for the Anthropocene. *Global Ecology and Biogeography*, 27(7), 760-786. http://dx.doi.org/10.1111/geb.12729

Environment Canada (2009). Canada's 4th National Report to the United Nations Convention on Biological Diversity.

Farrell, M. J., Brierley, L., Willoughby, A., Yates, A., & Mideo, N. (2022). Past and future uses of text mining in ecology and evolution. *Proceedings of the Royal Society B*, 289(1975), 20212721. http://dx.doi.org/10.1098/rspb.2021.2721

Farrell, M. J., Le Guillarme, N., Brierley, L., Hunter, B., Scheepens, D., Willoughby, A., ... & Mideo, N. (2024). The changing landscape of text mining-a review of approaches for ecology and evolution. *EcoEvorxiv*. https://doi.org/10.32942/X2VG87

Global Biodiversity Information Facility (GBIF). https://www.gbif.org

Graham, C.H., Ferrier, S., Huettman, F., Moritz, C. and Peterson, A.T. (2004). New developments in museum-based informatics and applications in biodiversity analysis. *Trends Ecol. Evol.*, 19, 497–503. http://dx.doi.org/10.1016/j.tree.2004.07.006

Guralnick, R. P., Hill, A. W., & Lane, M. (2007). Towards a collaborative, global infrastructure for biodiversity assessment. *Ecology letters*, 10(8), 663-672. http://dx.doi.org/10.1111/j.1461-0248.2007.01063.x

Guralnick, R., & Hill, A. (2009). Biodiversity informatics: automated approaches for documenting global biodiversity patterns and processes. *Bioinformatics*, 25(4), 421-428. http://dx.doi.org/10.1093/bioinformatics/btn659

Hampton, S. E., Strasser, C. A., Tewksbury, J. J., Gram, W. K., Budden, A. E., Batcheller, A. L., ... & Porter, J. H. (2013). Big data and the future of ecology. *Frontiers in Ecology and the Environment*, 11(3), 156-162. http://dx.doi.org/10.1890/120103

Heberling, J. M., Prather, L. A., & Tonsor, S. J. (2019). The changing uses of herbarium data in an era of global change: an overview using automated content analysis. *BioScience*, 69(10), 812-822. http://dx.doi.org/10.1093/biosci/biz094

Heberling, J. M., Miller, J. T., Noesgaard, D., Weingart, S. B., & Schigel, D. (2021). Data integration enables global biodiversity synthesis. *Proceedings of the National Academy of Sciences*, 118(6), e2018093118. http://dx.doi.org/10.1073/pnas.2018093118

Hendriks, I. E., & Duarte, C. M. (2008). Allocation of effort and imbalances in biodiversity research. *Journal of Experimental Marine Biology and Ecology*, 360(1), 15-20. http://dx.doi.org/10.1016/j.jembe.2008.03.004

Hoban, S., Bruford, M. W., Funk, W. C., Galbusera, P., Griffith, M. P., Grueber, C. E., ... & Vernesi, C. (2021). Global commitments to conserving and monitoring genetic diversity are now necessary and feasible. *Bioscience*, 71(9), 964-976. http://dx.doi.org/10.1093/biosci/biab054

Hoban, S., Archer, F. I., Bertola, L. D., Bragg, J. G., Breed, M. F., Bruford, M. W., ... & Hunter, M. E. (2022). Global genetic diversity status and trends: towards a suite of Essential Biodiversity Variables (EBVs) for genetic composition. *Biological Reviews*, 97(4), 1511-1538. http://dx.doi.org/10.1111/brv.12852

Hortal, J., de Bello, F., Diniz-Filho, J. A. F., Lewinsohn, T. M., Lobo, J. M., & Ladle, R. J. (2015). Seven shortfalls that beset large-scale knowledge of biodiversity. *Annual Review of Ecology, Evolution, and Systematics*, 46, 523-549. http://dx.doi.org/10.1146/annurev-ecolsys-112414-054400

IPBES. (2019). Global assessment report on biodiversity and ecosystem services of the Intergovernmental Science-Policy Platform on Biodiversity and Ecosystem Services (ed. by E. S. Brondizio, J. Settele, S. Díaz and H. T. Ngo). IPBES secretariat, Bonn.

Jetz, W., McPherson, J. M., & Guralnick, R. P. (2012). Integrating biodiversity distribution knowledge: toward a global map of life. *Trends in ecology & evolution*, 27(3), 151-159. http://dx.doi.org/10.1016/j.tree.2011.09.007

Jetz, W., McGeoch, M. A., Guralnick, R., Ferrier, S., Beck, J., Costello, M. J., ... & Turak, E. (2019). Essential biodiversity variables for mapping and monitoring species populations. *Nature ecology & evolution*, 3(4), 539-551. http://dx.doi.org/10.1038/s41559-019-0826-1

Jones, M. B., O'Brien, M., Mecum, B., Boettiger, C., Schildhauer, M., Maier, M., ... & Chong, S. (2019). Ecological metadata language version 2.2. 0. *KNB* Data Repository. doi:10.5063/F11834T2

Kelling S, Hochachka WM, Fink D, Riedewald M, Caruana R, Ballard G, et al. Data-intensive Science. A New Paradigm for Biodiversity Studies. *BioScience*. 2009; 59: 613–620. http://dx.doi.org/10.1525/bio.2009.59.7.12

Kopperud, B. T., Lidgard, S., & Liow, L. H. (2019). Text-mined fossil biodiversity dynamics using machine learning. *Proceedings of the Royal Society B*, 286(1901), 20190022. http://dx.doi.org/10.1098/rspb.2019.0022

Kulkarni, R., & Di Minin, E. (2021). Automated retrieval of information on threatened species from online sources using machine learning. *Methods in Ecology and Evolution*, 12(7), 1226-1239. http://dx.doi.org/10.1111/2041-210X.13608

Lait, Linda A.; Burg, Theresa M. (2013). Data from: When east meets west: population structure of a high- latitude resident species, the boreal chickadee (*Poecile hudsonicus*) [Dataset]. *Dryad*. https://doi.org/10.5061/dryad.82hs7

Lamarre, Jean-François et al. (2018). Data from: Predator-mediated negative effects of overabundant snow geese on arctic-nesting shorebirds [Dataset]. *Dryad*. https://doi.org/10.5061/dryad.796t8

Leadley, P., Gonzalez, A., Obura, D., Krug, C. B., Londoño-Murcia, M. C., Millette, K. L., ... & Xu, J. (2022). Achieving global biodiversity goals by 2050 requires urgent and integrated actions. *One Earth*, 5(6), 597-603. http://dx.doi.org/10.1016/j.oneear.2022.05.009

Leblond, Mathieu; St-Laurent, Martin-Hugues; Côté, Steeve D. (2017). Data from: Caribou, water, and ice – fine-scale movements of a migratory arctic ungulate in the context of climate change [Dataset]. *Dryad*. https://doi.org/10.5061/dryad.4k275

Löffler, F., Wesp, V., König-Ries, B., & Klan, F. (2021). Dataset search in biodiversity research: Do metadata in data repositories reflect scholarly information needs?. *PloS one*, 16(3), e0246099. http://dx.doi.org/10.1371/journal.pone.0246099

Maiorano, L., Montemaggiori, A., Ficetola, G. F., O'connor, L., & Thuiller, W. (2020). TETRA-EU 1.0: a species-level trophic metaweb of European tetrapods. *Global Ecology and Biogeography*, 29(9), 1452-1457. http://dx.doi.org/10.1111/geb.13138

Martin, M., Leduc, A., Fenton, N. J., Montoro Girona, M., Bergeron, Y., & Valeria, O. (2022). Irregular forest structures originating after fire: An opportunity to promote alternatives to even-aged management in boreal forests. *Journal of Applied Ecology*, 59(7), 1792-1803. http://dx.doi.org/10.1111/1365-2664.14186

McCallen, E., Knott, J., Nunez-Mir, G., Taylor, B., Jo, I., & Fei, S. (2019). Trends in ecology: shifts in ecological research themes over the past four decades. *Frontiers in Ecology and the Environment*, 17(2), 109-116. http://dx.doi.org/10.1002/fee.1993

Millenium Ecosystem Assessment Ecosystems and Human Well-being. World Resources Insitute, 2005

Nunez-Mir, G. C., Iannone III, B. V., Pijanowski, B. C., Kong, N., & Fei, S. (2016). Automated content analysis: addressing the big literature challenge in ecology and evolution. *Methods in Ecology and Evolution*, 7(11), 1262-1272. http://dx.doi.org/10.1111/2041-210X.12602

Newbold, T., Hudson, L. N., Hill, S. L., Contu, S., Lysenko, I., Senior, R. A., ... & Purvis, A. (2015). Global effects of land use on local terrestrial biodiversity. *Nature*, 520(7545), 45-50. http://dx.doi.org/10.1038/nature14324

National Center for Biotechnology Information (NCBI). https://www.ncbi.nlm.nih.gov/

Ocean Biodiversity Information System (OBIS). https://obis.org/

Page, L. M., MacFadden, B. J., Fortes, J. A., Soltis, P. S., & Riccardi, G. (2015). Digitization of biodiversity collections reveals biggest data on biodiversity. *BioScience*, 65(9), 841-842. http://dx.doi.org/10.1093/biosci/biv104

Parmesan, C., & Yohe, G. (2003). A globally coherent fingerprint of climate change impacts across natural systems. *Nature*, 421(6918), 37-42. http://dx.doi.org/10.1038/nature01286

Pereira, H. M., Ferrier, S., Walters, M., Geller, G. N., Jongman, R. H., Scholes, R. J., ... & Wegmann, M. (2013). Essential biodiversity variables. *Science*, 339(6117), 277-278. http://dx.doi.org/10.1126/science.1229931

Pyšek, P., Hulme, P. E., Simberloff, D., Bacher, S., Blackburn, T. M., Carlton, J. T., ... & Richardson, D. M. (2020). Scientists' warning on invasive alien species. *Biological Reviews*, 95(6), 1511-1534. http://dx.doi.org/10.1111/brv.12627

Schmeller, D. S., Weatherdon, L. V., Loyau, A., Bondeau, A., Brotons, L., Brummitt, N., ... & Regan, E. C. (2018). A suite of essential biodiversity variables for detecting critical biodiversity change. *Biological Reviews*, 93(1), 55-71. http://dx.doi.org/10.1111/brv.12332

Stork, H., & Astrin, J. J. (2014). Trends in biodiversity research—a bibliometric assessment. *Open Journal of Ecology*, 4(07), 354. http://dx.doi.org/10.4236/oje.2014.47033

Thiffault, Nelson; Grondin, Pierre; Noël, Jean; Poirier, Véronique (2016). Data from: Ecological gradients driving the distribution of four Ericaceae in boreal Quebec, Canada [Dataset]. *Dryad*. https://doi.org/10.5061/dryad.4767v

Vihervaara, P., Auvinen, A. P., Mononen, L., Törmä, M., Ahlroth, P., Anttila, S., ... & Virkkala, R. (2017). How essential biodiversity variables and remote sensing can help national biodiversity monitoring. *Global Ecology and Conservation*, 10, 43-59. http://dx.doi.org/10.1016/j.gecco.2017.01.007

Wen, J., Ickert-Bond, S. M., Appelhans, M. S., Dorr, L. J., & Funk, V. A. (2015). Collections-based systematics: Opportunities and outlook for 2050. *Journal of Systematics and Evolution*, 53(6), 477-488. http://dx.doi.org/10.1111/jse.12181

Wen, J., Harris, A. J., Ickert-Bond, S. M., Dikow, R., Wurdack, K., & Zimmer, E. A. (2017). Developing integrative systematics in the informatics and genomic era, and calling for a global

Biodiversity Cyberbank. *Journal of Systematics and Evolution*, 55(4), 308-321. http://dx.doi.org/10.1111/jse.12270

Wieczorek, J., Bloom, D., Guralnick, R., Blum, S., Döring, M., Giovanni, R., ... & Vieglais, D. (2012). Darwin Core: an evolving community-developed biodiversity data standard. *PloS one*, 7(1), e29715. http://dx.doi.org/10.1371/journal.pone.0029715

WWF-Canada. 2020. Living Planet Report Canada: Wildlife At Risk. Currie J. Snider J. Giles E. *World Wildlife Fund Canada*. Toronto, Canada. DOI: 10.13140/RG.2.2.16556.49280

Data Accessibility Statement

All the data and code is available on both Zenodo (https://doi.org/10.5281/zenodo.11288378) and Github (https://github.com/Alex-Fuster/automated_datset_retrieval.git).

Figure 1

Sources and availability of datasets relevant for biodiversity.

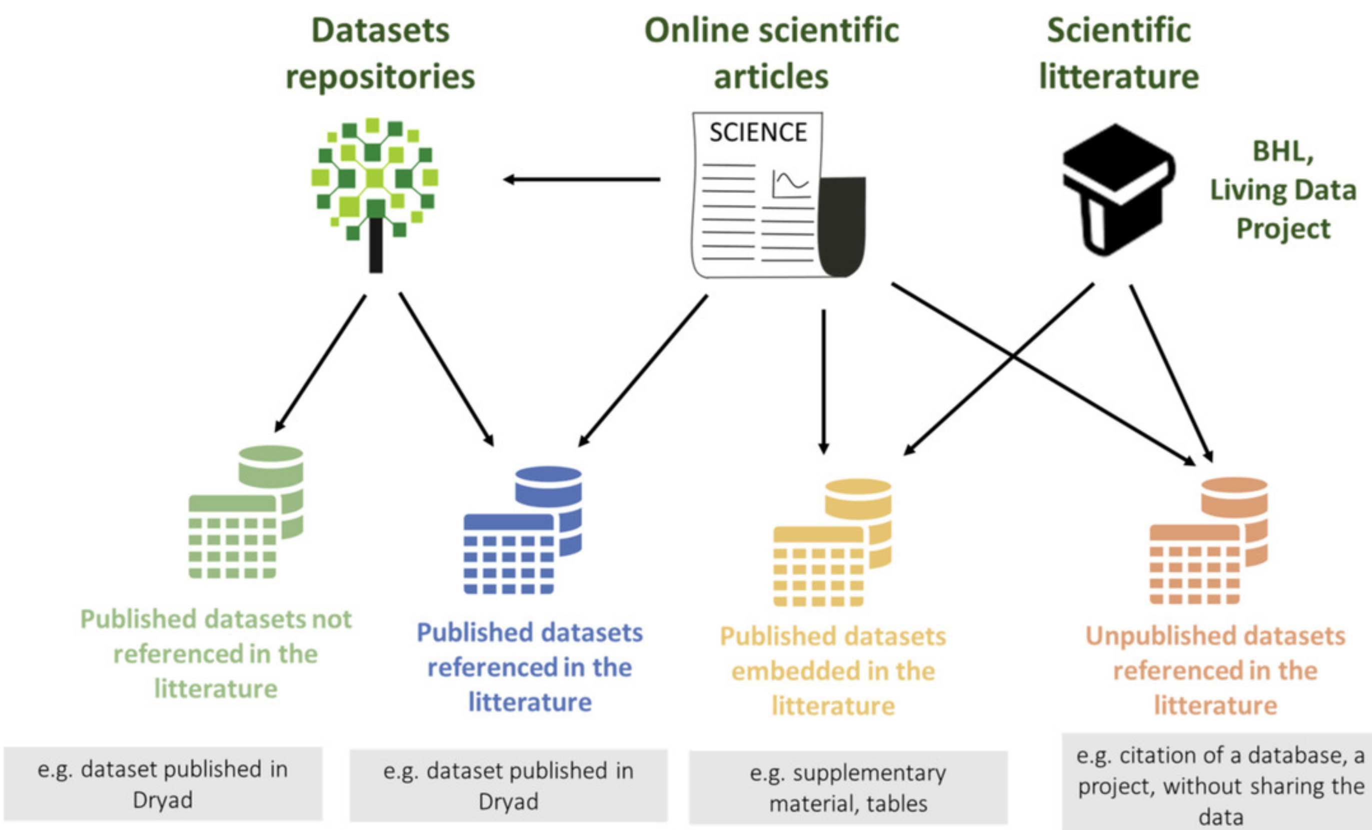

Figure 2

Queries performance retrieving relevant datasets from repositories (Dryad and Zenodo)

(A) Number of publications and counts of relevant categories per query. Colors indicate relevance categories: H (High), M (Moderate), L (Low), and X (Negligible). (B) F scores, precision, and recall metrics for each query. All queries also account for the word “Quebec”.

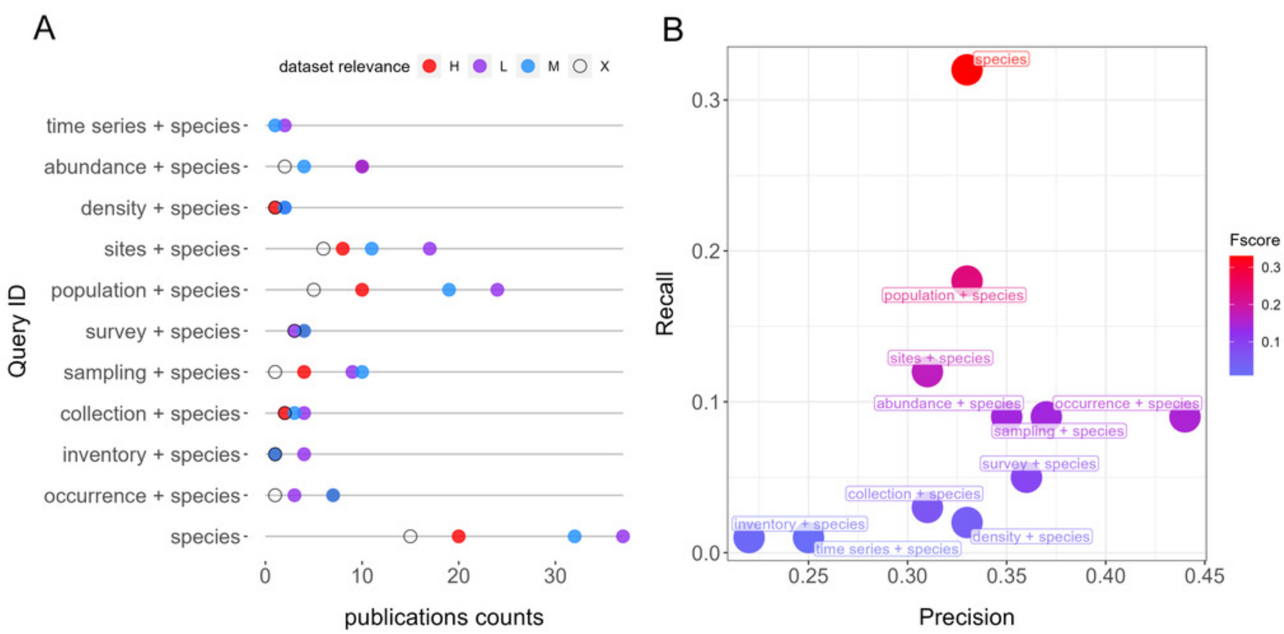

Figure 3

Features of retrieved datasets from repositories Dryad and Zenodo and results of the relevance evaluation.

(A) Temporal duration and ranges. Red range bars indicate datasets with time series data. The first dataset extends to 1875. Not showing outlier dataset from 20.000 ya to the present. (B) Duration, spatial rage, and relevance category. (C) Publication counts by EBV data types. (D) Spatial range counts in the low, medium, and high range categories, respectively. (E) Publication counts by taxa and relevance categories: the left panel shows species-level studies, which contain data for 1 to 10 species, whereas the right panel shows community-level studies with data for more than 10 species. Letters H, M, and L correspond to High, Moderate, and Low dataset relevance categories, respectively.

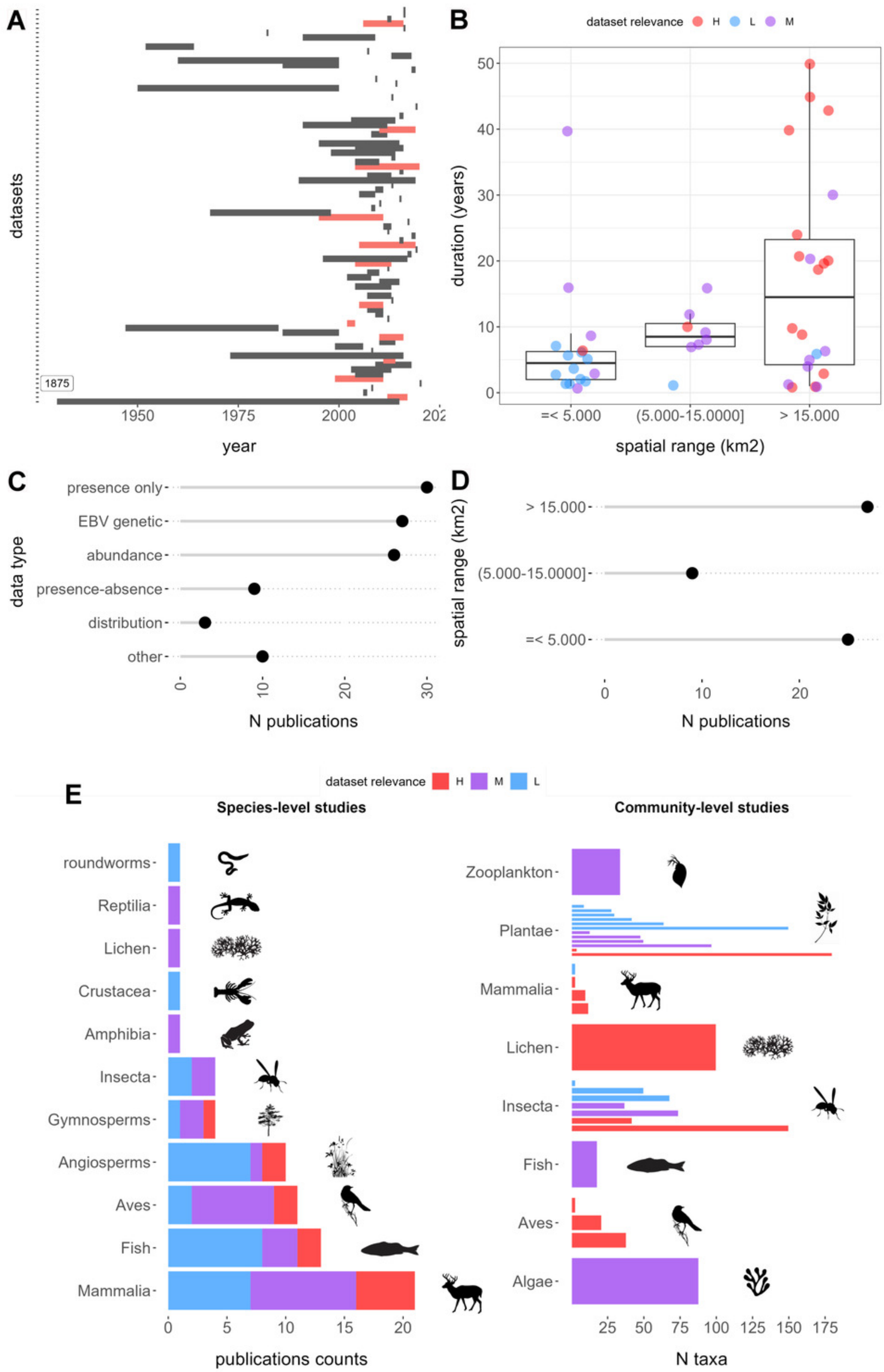
A
datasets
1875
1950
1975
2000
202
year
B
dataset relevance H L M
duration (years)
50
40
30
20
10
0
=< 5.000
(5.000-15.0000]
> 15.000
spatial range (km2)
C
data type
presence only
EBV genetic
abundance
presence-absence
distribution
other
0
10
20
30
N publications
D
spatial range (km2)
> 15.000
(5.000-15.0000]
=< 5.000
0
10
20
N publications
E
dataset relevance H M L
Species-level studies
roundworms
Reptilia
Lichen
Crustacea
Amphibia
Insecta
Gymnosperms
Angiosperms
Aves
Fish
Mammalia
0
5
10
15
20
publications counts
Community-level studies
Zooplankton
Plantae
Mammalia
Lichen
Insecta
Fish
Aves
Algae
25
50
75
100
125
150
175
N taxa

Figure 4

Location of features (i.e. metadata) in the publication spaces.

(A) Spatiotemporal features and (B) geospatial features.

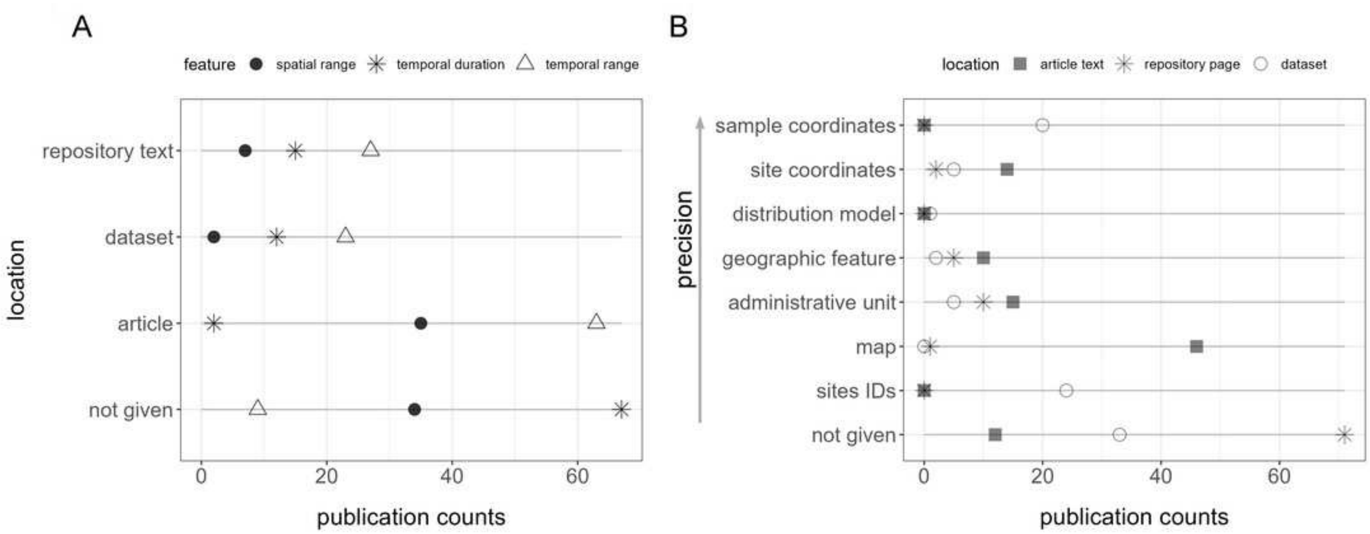

Figure 5

Comparison between Semantic Scholar and repositories (Dryad and Zenodo).

(A) Relevance and accessibility of datasets. (B) Temporal depth of retrieved publications.

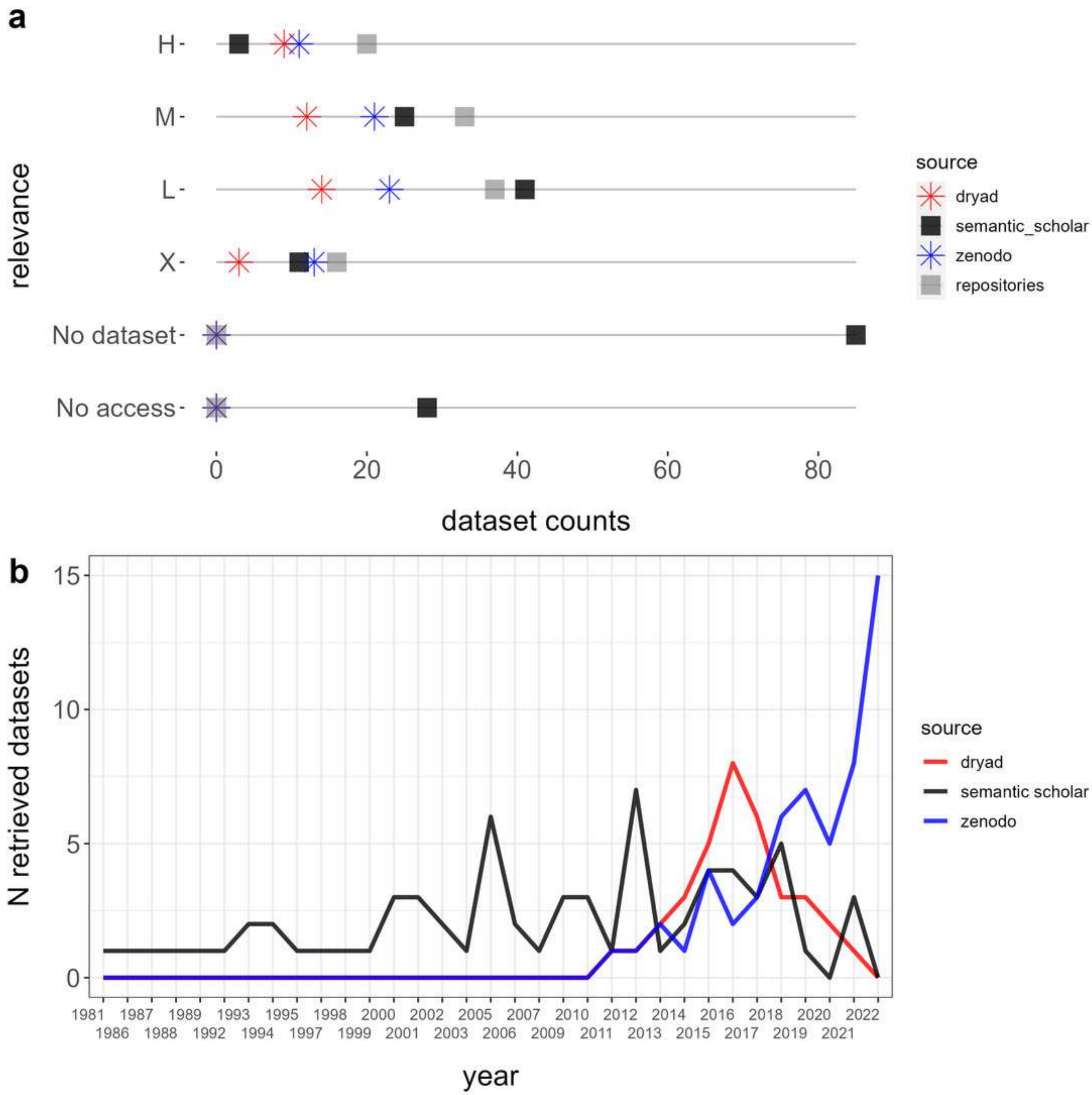

Table 1(on next page)

Manually evaluated features from retrieved datasets.

Feature	**Type**	**Example**
EBV data type	categorical	abundance
Geospatial information	continuous	sample coordinates
Spatial range	continuous	100.000 km2
Temporal range	string	from 1999 to 2008
Temporal duration	continuous	9 years
Taxons	string	black-legged tick
Referred dataset source	string	Ministère des Ressources naturelles et des Forêts
Dataset location	categorical	repository
Dataset type location	categorical	abstract
Geospatial information location	categorical	article text
Spatial range location	categorical	abstract
Temporal range location	categorical	dataset
Temporal duration location	categorical	article text

Table 2 (on next page)

Evaluation criteria used to assign the relevance category to the datasets.

First, the main classifiers determine whether the dataset relevance is “High”, “Medium”, “Low”, or “Negligible” (“Relevance by Main Classifiers” column), and then Modulators can increase or decrease the relevance category by one (e.g. “Low” to “Medium” but not “Low” to “High”). The “Organization level” modulator takes into account whether the data is about individuals, populations, or species; the “Bias North-South” modulator is evaluated by establishing an arbitrary latitudinal threshold to separate northern from southern areas of Québec, with northern areas having higher relevance. In the example below, data on an endangered species may have relevance “Moderate” according to the main classifiers but its conservation status and priority location would increase it to “High”.

	Main Classifiers				Modulators				
Data type	Data size	Spatial range	Temporal range	**Relevance by Main Classifiers**	Organization level	Conservation status	New regional species	Bias North-South Quebec	**Relevance**
Presence only	Moderate	Moderate	Low	**Moderate**	individual	EN	FALSE	North	**High**