InsectGapMap

A project to semi-automate research gap identification

Final project report, November 2020

Neal R Haddaway,
Senior Research Fellow at SEI

Eliza Grames,
PhD candidate, University of Connecticut

Table of contents

Executive summary

Background:

Computational tools show particular promise as means of rapidly and relatively crudely scoping literature on a topic. Such assessments, whilst not sufficiently rigorous to act as a basis for policy or practice decision making, could be particularly useful as decision-support tools to help identify topics that warrant further attention in the form of primary or secondary research, and research funding.

This report documents the use of a type of text analysis known as *topic modelling* to automatically extract information (such as relevant conservation actions) from a set of research records (i.e. titles and abstracts) and then catalogue how much evidence exists across key variables (such as actions, publication year, taxon, and location) thus, highlighting knowledge gaps in a semi-automatic way.

Objectives and methods:

The project's goals were to: identify potential research gaps that could warrant funding for primary research; identify gaps using semi-automated data science tools; and, develop an easily replicable methodology for semi-automated research gap identification that could be readily applied to other topics. The methods involved the following procedures: assembly of a longlist of insect conservation actions; conversion of this list into a hierarchical ontology of actions based on their key characteristics; application of this ontology as 'seed' points in semi-supervised topic modelling; combination of 'dictionary methods' for classifying documents (based on presence of specific terms) with 'unsupervised topic modelling' to classify studies (based on terminology found in the document abstracts); visualisation of the volume of evidence identified interactively across conservation actions, biomes, regions, and taxa.

Results:

Approximately 50,000 documents were identified that focused on insect conservation. Over 800 insect conservation actions were identified from over 9,000 literature reviews and from Conservation Evidence. Ontologies for biomes, regions and insect taxa were supplemented by a novel ontology of conservation actions co-designed with c. 40 expert stakeholders. These ontologies were then applied using dictionary methods and iterative topic modelling to 'tag' 11,492 documents with conservation actions, geographical locations, insect taxa, and studied biomes.

Conclusions:

This map reveals patterns that correspond to what many subject experts suspected was true of the literature on insect conservation actions—in many cases there is substantial evidence of effectiveness for insect conservation actions, but limited research on the broader theories of change that involve institutional, political, and social contexts that affect the efficacy of insect conservation actions. The major causes of population declines (habitat loss and habitat degradation, invasive species, pollution and pesticide application, and climate change) and ways to mitigate or reverse their effects have received the most study, as revealed by the clusters of

knowledge in our map on the topics of habitat management and species management, such as reducing pesticide contamination of insect habitats, controlling vegetation succession, and increasing habitat complexity. Our map revealed large gaps in all areas of the ontology related to human dimensions of insect conservation: habitat protection, education and awareness, law and policy, and livelihood, economic, and other incentives. This highlights the greater need for integration of social science in conservation research to test whether actions that work in theory can also work in the 'real world', embedded in complex human contexts. Our map identifies a large number of highly specific topics that may represent gaps and warrant further funding. For example: habitat protection (creating insect reserves, using taxonomic checklists for resource allocation, establishing riparian reserves, protecting forests, preventing grassland conversion), and habitat management (controlling and preventing invasive and non-native species). This work highlights a significant gap related to interdisciplinary research teams spanning natural and social science that bring together entomologists, conservation biologists, and social scientists with experts in environmental law, business and industry certification boards, inclusive education, communications and public relations, and urban planning and community development. Future research on insect conservation should focus on the human dimensions of solutions, such as community development, public education, and law and policy.

Partially automated methods can help to identify research clusters and gaps that are consistent with patterns expected by subject experts in a way that is reproducible and avoids confirmation bias. In this project, we have developed a semi-automated method to rapidly map the knowledge in a field of research by combining expert opinions and text mining approaches, and applied it to the topic of insect conservation to identify which insect conservation actions have been studied and which are urgent priorities for future research. Because the methods, R package, and ontologies from this project are being made publicly available and fully replicable, other research teams can adopt the tools to conduct similar rapid assessments of the state of knowledge on a topic. Although this approach is not recommended for full evidence syntheses such as those that would be used to inform policy, it could be used for other efforts to make recommendations for funding, or to rapidly identify research gaps that could merit further consideration.

Background

Systematic maps (SMs) are rigorous, methodological processes for identifying, collating, and describing evidence pertinent to a particular question (James, Randall & Haddaway 2016), based closely on the related method of systematic review (The Collaboration for Environmental Evidence 2018). SMs are used to answer questions about what research exists on a topic, which features have been studied, and who has conducted the research, where and how. SMs are vital and robust tools to identify knowledge gaps (topics within the evidence base where there exists an underrepresentation or lack of research) and knowledge clusters (topics where sufficient quantity and quality of studies exist to allow full synthesis of study findings). However, they require substantial resources to undertake well, with one estimate placing the person days required to complete a SM at 211 days (Haddaway & Westgate 2019).

Evidence synthesis technology (tools and techniques for improving the efficiency, accessibility, rigour and legacy of systematic reviews and maps) is a particularly popular topic of discussion and methodological research at present (Westgate *et al.* 2018). The use of machine learning and other text analysis tools seems a promising means to replace repetitive and laborious tasks within systematic reviews and maps, offering opportunities to significantly increase the efficiency of review conduct (Marshall & Wallace 2019). However, we are still a long way from the possibility of a complete, reliable, computer-driven systematic review or map.

Despite this, computational tools show particular promise as means of rapidly and relatively crudely scoping literature on a topic (Stansfield, Thomas & Kavanagh 2013). Such assessments, whilst not sufficiently rigorous to act as a basis for policy or practice decision making, could be particularly useful as decision-support tools to help identify topics that warrant further attention in the form of primary or secondary research, and research funding.

Here, we propose the use of a type of text analysis known as *topic modelling* (Blei & Lafferty 2009) to automatically extract information (such as relevant interventions) from a set of research records (i.e. titles and abstracts) and then catalogue how much evidence exists across key variables (such as interventions, publication year, taxon, and location) thus, highlighting knowledge gaps automatically. An intermediate stage would require human verification of the intervention list (a so-called ontology of interventions) before the evidence is catalogued. Potential knowledge gaps would then need interpretation by subject experts. We outline the proposed methods below.

Although this method could be applied to any set of search results, we explain the process in the context of insect biodiversity conservation interventions, since this topic was suggested by Mistra as a potential priority area, and because of the shared objectives and overlapping timeline with the ongoing EntoGEM project. EntoGEM aims to conduct a community-driven systematic mapping of evidence pertaining to insect populations trends globally, by assembling a group of researcher volunteers to manually screen 144,000 search results to find relevant evidence.

Goals and objectives

This project had the following goals and methodological objectives:

Project goals

- To identify potential research gaps that could warrant funding for primary research
- To identify gaps using semi-automated data science tools
- To develop an easily replicable methodology for semi-automated research gap identification that could be readily applied to other topics

Methodological objectives

- To assemble a longlist of insect conservation actions
- To convert this list into a hierarchical ontology of actions based on their key characteristics
- To apply this ontology as ‘seed’ points in semi-supervised topic modelling
- To combine ‘dictionary methods’ for classifying documents (based on presence of specific terms) with ‘unsupervised topic modelling’ to classify studies (based on terminology found in the document abstracts)
- To visualise the volume of evidence identified interactively across four key axes:
    - Conservation actions
    - Biomes
    - Regions
    - Taxa

Methods

Methodological steps

Table 1 outlines the key steps used in this project and how the outputs from each relate to one another.

Table 1. Summary of the project approaches, objectives and methods.

Approach	Objective	Methods
Collated insect conservation actions	Systematic map of literature reviews	Development (and publication) of an *a priori* protocol; systematic searches for literature reviews on insect conservation; screening of titles and abstracts against predefined inclusion criteria; extraction of conservation actions from abstracts; screening of full texts; extraction of additional meta-data from studies; critical appraisal of review validity using CEESAT2; reporting of findings.
	Collated actions from Conservation Evidence	'Scrape' conservation actions from relevant syntheses conducted by the organisation Conservation Evidence; integrate actions from an ongoing aquatic invertebrate conservation synopsis.
Ontology development	Cluster actions in concise hierarchy	Longlist of actions subsetted by subject; recruitment of expert stakeholder panel; conduct of workshops to cluster actions into draft ontology; refinement of ontology; presentation of ontology on Open Access, web-based platform; reporting of methods and findings.
Semi-supervised topic modelling	Assembly of evidence base	Filtering of extensive evidence base from the EntoGEM project (containing >144,000 research records relating to insect populations) based on conservation action search terms.
	'Dictionary methods' to tag studies containing conservation actions (and other ontologies)	Application of the ontology to detect strict mentioning of actions and their synonyms (identified from the review of reviews and refined into the ontology); application of other established ontologies to tag studies with geographical location, insect taxa and biome.
	'Semi-supervised topic modelling' to classify remaining studies	Use of topic modelling to cluster studies not tagged with dictionary methods; automated creation of clusters of topics based on conservation actions in the ontology.
Visualisation and analysis	Produce interactive media to communicate the findings and allow interrogation of the data	Creation of visualisations of the volume of studies identified across all ontologies (actions, biomes, regions, taxa); development of interactive platform to support interrogation of the findings and identification of research gaps (under-representations of primary research) and synthesis gaps (sufficient studies to warrant evidence synthesis).

In the following pages, we describe each process in detail. This process is summarised in Figure 1.

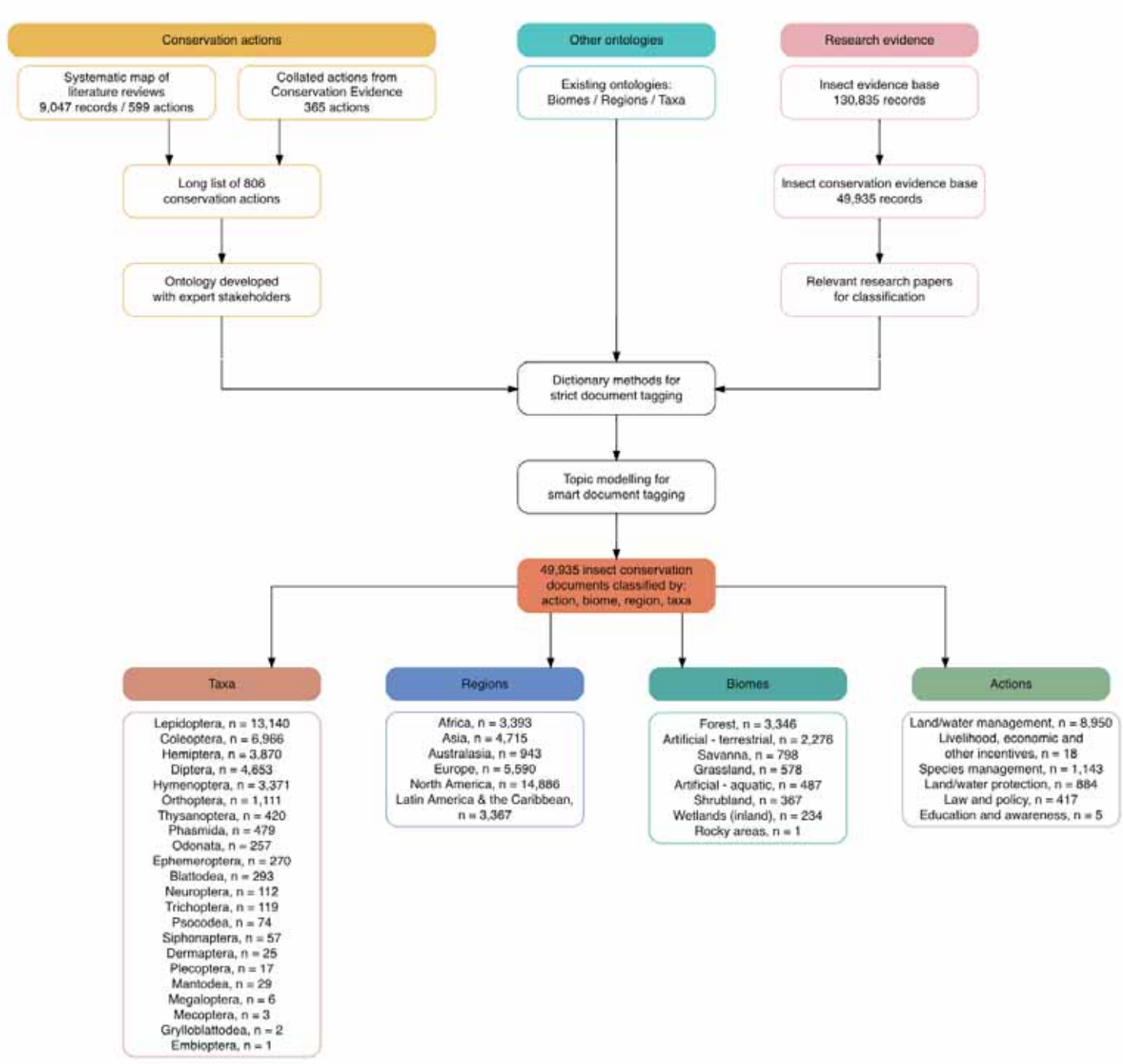


Figure 1. Flow of information through the project across three types of data (conservation actions, other ontologies, and research evidence).

Collate insect conservation actions

Systematic map of literature reviews

We developed an *a priori* protocol that outlines the planned methods for a systematic map of literature reviews on insect conservation (Haddaway *et al.* 2020). The systematic map was conducted in accordance with guidance on evidence synthesis set out by the Collaboration for Environmental Evidence (The Collaboration for Environmental Evidence 2018), and the protocol and final review are reported in accordance with the ROSES reporting standards for systematic maps (Haddaway *et al.* 2018). We provide a summary of the methods used here. For further details, see the protocol (Haddaway *et al.* 2020).

We searched for evidence reviews across seven large, generic bibliographic databases, a database of environmental reviews, and five grey literature resources using a search string consisting of an insect substring, a biodiversity or population response substring, an evidence synthesis substring, and a conservation substring.

The 27,283 results were deduplicated and then screened at title and abstract (concurrently) against predefined inclusion criteria. The public project page for the review is accessible on the SysRev review management platform (https://sysrev.com/u/371/p/31569). We initially performed consistency checking on a subset of abstracts to ensure the inclusion criteria were sufficiently clear and understood by multiple reviewers.

A total of 9,047 unique records were screened at title and abstract level. We extracted a total of 599 unique insect conservation actions and supplementary notes from 840 relevant review abstracts.

The authors of this protocol plan to continue extracting data from full texts and publish the systematic map on a voluntary basis. This output will not directly feed into the outputs of this project, but any additional conservation actions discovered from reading full texts will be highlighted and passed to a refinement of the models described below in a 'living' analysis.

We will extract a suite of descriptive meta-data from relevant reviews, including a description of the action and information on each review's focal taxa, biomes, and locations. If resources allow, we will apply the CEESAT critical appraisal tool for evidence reviews to assess validity of individual records and the evidence base as a whole.

Collated actions from Conservation Evidence

The organisation Conservation Evidence (https://www.conservationevidence.com/) lists conservation actions across a wide variety of environmental management topics (e.g. bat conservation (Berthinussen, Richardson & Altringham 2020), amphibian conservation (Smith & Sutherland 2014), bee conservation (Dicks, Showler & Sutherland 2010), forest conservation (Agra *et al.* 2016), and natural pest control (Wright *et al.* 2013)). These synopses are somewhat systematic approaches to collating and summarising representative research on conservation and environmental management actions. Whilst there are no synopses that focus on insects

collectively, some are directly relevant (e.g. bee conservation), whilst others contain specific actions relevant to insects.

We constructed a tool to scrape conservation actions from the Conservation Evidence website in R. We downloaded 2,400 actions catalogued by Conservation Evidence as a plain text file and used text mining to extract the summary for any actions mentioning insect taxa or pollinators. We manually screened these 279 action summaries and extracted any actions that were potentially relevant and had not been identified mistakenly (e.g. actions about other pollinating groups such as bats), resulting in 164 conservation actions that were potentially relevant to insects.

In addition, through contacts with researchers at Conservation Evidence, we became aware of an ongoing synopsis focusing on terrestrial macroinvertebrates (with a particular emphasis on Lepidoptera) led by Andrew Bladon at the University of Cambridge. We were therefore able to obtain an additional set of candidate actions from this ongoing, relevant work and merge these with the actions scraped from published Conservation Evidence summaries.

A total of 365 actions were collated from across Conservation Evidence.

Ontology development

The long list of conservation actions described above was refined to remove redundant terms and subsetted into 7 subject categories prior to assembling a draft ontology. These categories were necessary in order to provide experts at the workshop with a manageable number of actions and to allow experts to be provided with actions relevant to their expertise as far as possible. The following groups were created: aquatic; farmland; forest; grassland; policy; urban; and other.

Expert stakeholders were recruited in three main ways: systematic searching; a public call; and snowballing suggestions from existing experts. This combination approach was selected in the hopes of minimising risks of bias, such as: selection bias (choosing people we were aware of); confirmation bias (choosing people with similar views and opinions); subject bias (an overrepresentation of experts working on certain actions, biomes, regions or taxa); geographical bias (an overrepresentation of people from some regions, e.g. Europe/North America); language bias (an overrepresentation of English or US/British English speakers). We are still very aware of a lack of balance in our final set of stakeholders, but we tried hard to minimise this.

Firstly, a highly specific search of Web of Science Core Collections[1] was conducted[2], resulting in 84 relevant research articles on insect conservation in the last 3 years (in order to increase the

[1] Consisting of the following indexes: Science Citation Index Expanded (SCI-EXPANDED) --1900-present, Social Sciences Citation Index (SSCI) --1900-present, Arts & Humanities Citation Index (A&HCI) --1975-present, Conference Proceedings Citation Index- Science (CPCI-S) --1990-present, Conference Proceedings Citation Index- Social Science & Humanities (CPCI-SSH) --1990-present, Emerging Sources Citation Index (ESCI) --2015-present.

[2] The following search string was used on 25th June 2020: *"(TI=("insect conservation" OR "conservation of insects") OR AB=("insect conservation" OR "conservation of insects") OR AK=("insect conservation" OR "conservation of insects")) AND PY=(2017 OR 2018 OR 2019 OR 2020) AND WC=(ECOLOGY OR*

likelihood that emails were still viable). From these results, we obtained 76 unique email addresses for corresponding authors. Secondly, we invited known expert stakeholders from our own networks and made use of social media to widen the reach of the call, emphasising the need for representation from low- and middle- income country participants. Together, recruitment from our own networks, social media and snowballing resulted in an additional 34 contacts. A total of 110 experts were therefore contacted with a request to participate.

A total of 45 expert stakeholders participated in two 3-hour workshops held at two different times to account for time zones on the 11th and 13th August 2020. The workshops were coordinated by the authors. During the workshops, participants were provided with the subsetted list of actions for their group and were asked to cluster them based on similarity in the approach used. Some experts were given more than one group where the number of actions were smaller (see Table 2 for further details). Following the workshop, a small minority of actions remained to be nested into the ontology. We finalised the ontology based on the existing structure.

Table 2. Groups of actions, the number of actions within each group, and the number of experts assigned to each group.

Group	Actions	Expert stakeholders
Aquatic	112	5
Farmland	205	8
Forest	93	4
Grassland	105	8
Policy	69	3
Urban	40	4
Other	182	4

The final ontology is publicly available: https://doi.org/10.5281/zenodo.4383213. This resource is a highly useful resource for conservation frameworks and text analysis and has therefore been made Open Access.

R package development

To make our method reproducible and accessible to other research teams, we developed an open source R package containing functions, workflows, and example datasets to replicate the analyses

ENTOMOLOGY OR BIODIVERSITY CONSERVATION OR ENVIRONMENTAL SCIENCES)"; where TI = title words, AB = abstract words, AK = author keywords, PY = publication year, WC = Web of Science subject category.

done in this project for a subset of the data. The R package *topictagger* (Grames & Haddaway 2020) can be used to extract and tag meta-data and topics, such as location of study or primary outcome, for evidence syntheses. Given a hierarchical ontology of relationships between entities, a set of articles with known topics, or a set of terms associated with a topic, topictagger will tag the specified topics in a set of user-supplied documents. If the topics contained in a set of documents are unknown, topictagger can also run unseeded topic models to discover possible topics and clusters of similar documents within the dataset. The package has documentation for all functions and is accompanied by a vignette that functions as a tutorial for using the package by demonstrating how the functions can be applied to the example dataset of determining which conservation actions and biomes are represented in a sample of entomology papers (i.e. a subset of the current project).

Semi-supervised topic modelling

Assembly of evidence base

We aimed to assemble an evidence base for automated research gap analysis that was comprehensive enough to have a high likelihood of being representative of the research landscape as a whole. However, there is a trade off with the need for a manageable set of records to allow complex topic modelling, which can be computationally demanding. We were able to start from the set of >138,000 search results from the EntoGEM project as a basis.

EntoGEM (https://entogem.github.io/) is an ongoing project that aims to conduct a community-driven systematic map of all research literature on insect population trends. As part of this work, systematic searches for relevant evidence were already conducted across 18 bibliographic databases and other resources. The resultant records focused on all aspects of insect populations, not just insect conservation actions.

In order to obtain a more manageable number of records, we filtered this database using a set of conservation terms[3], which resulted in 49,935 records.

'Dictionary methods' to tag studies containing conservation actions (and other ontologies)

We used the expert ontology to create a nested dictionary of insect conservation actions, where each entry was composed of key terms and synonyms extracted from the entries in the ontology. More specific action classifications (e.g. those at the tip of the ontology) were included in the definitions of higher-level actions. For example, the action "restrict the sale of problem species in garden centres and pet shops" is defined by the terms "problem species", "garden centres", and "pet shops". This action and its terms are included in the definition of the higher-level action "enforce invasive species control", along with its sister action "increase biosecurity checks" which

[3] Search terms used within the EntoGEM database to identify research focusing on conservation actions: *conserv*, manag*, restor*, preserv*, action*, interven*, policy, policies, practice*, protect*, reintroduc*, scheme, regulat*, legislat*, rule, rules, subsidi*, tarif*, reform*, red list, recovery, communication, awareness, endangered, threatened, vulnerable*

has its own nested actions and their associated terms. Any article classified as "restrict the sale of problem species in garden centres and pet shops" will also automatically be classified as "enforce invasive species control" and the higher-level actions "Compliance and enforcement" and "Law & policy".

Using the same nested dictionary structure, we constructed a dictionary of biomes based on the IUCN Red List Habitats Classification Scheme (Version 3.1.); for example, "caatinga", "miombo", and "mulga" are all terms included in the definition of "dry savanna" which is itself included in the definition of "savanna" along with "moist savanna" and its descendants. We also created a nested dictionary of cities, regions, countries, and continents based on the World Cities Database (https://simplemaps.com/data/world-cities) and the UN country classification scheme. We removed cities that are common words in English (e.g. Of, Turkey; Same, Timor-Leste; Same, Tanzania; Young, Australia) or within the insect conservation literature (e.g. Male, Maldives and Pest, Hungary). For common regional names shared across countries (e.g. "Northern", "Central"), we appended the region to the name of the country to avoid ambiguity in the dictionary. To tag insect taxa, we used a dictionary of insect orders based on the GBIF Taxonomy, with synonyms including common names and genus names.

Using the evidence base and our four dictionaries (conservation actions, biomes, geographic regions, and insect taxa), we created four document-feature matrices where each row represents one of the articles in the evidence base, and each column in the matrix represents a dictionary entry and its associated terms. Counts of how many terms from a dictionary entry appeared in an article were used to classify each article. Although many articles were matched to multiple topics, the maximum count was used as the primary classification for actions, countries, habitats, and insect taxa for each article.

'Semi-supervised topic modelling' to classify remaining studies

To classify articles that were not tagged with a conservation action using the dictionary approach, we trained a model to probabilistically assign each remaining article to one of the top-level conservation actions (i.e. Land/water protection, Land/water management, Species management, Education & awareness, Law & policy, or Livelihood, economic, and other incentives). Because the dictionary approach mapped too few papers to some approaches to provide adequate sample sizes for training the model, several actions related to education, law, and livelihood were grouped as 'Unclassified Human Dimensions' for the model training dataset. To reduce the risk of articles being mistakenly classified to one of our conservation actions, we randomly sampled 1500 articles from the EntoGEM database that were not included in our subset of conservation articles and added these unrelated articles to those in our evidence base. By adding this "noise" to the model input, we were able to add a topic into which the model could classify articles that did not match any of our main conservation actions.

To build the topic model, we extracted key phrases from articles using the *mine_terms()* function from the R package topictagger (Grames & Haddaway 2020). We refined the list of possible phrases to only include phrases which appeared at least three times across the evidence base in order to reduce the chance of the model identifying key phrases as being unique to a topic due to their rarity (e.g. the name of a specific national park). We also filtered out terms that are generic and uninformative (e.g. "results indicate" or "positively correlated"), which can lead to higher probability of an article being included in a common topic simply because the phrase is common.

We also removed terms not associated with article content that are the result of the evidence base being pulled from bibliographic databases (e.g. "full abstract" or "copyright holder").

We used the final list of terms to create a document-feature matrix where each row represented an article from the evidence base and the additional noise, and each column represented one of the final phrases. This matrix was used to fit a symmetric multinomial model using the *glmnet* function in the R package *glmnet* v.4.0.2 (Friedman, Hastie & Tibshirani 2010), which we used to predict classifications for articles that were not tagged by the dictionary approach. The output of is a matrix of probability of classification to each topic for each previously unclassified article. We classified articles that had at least a 0.30 probability of inclusion in a topic as belonging to it, and marked all articles that were classified as "noise" or with a lower than .30 probability of inclusion in a topic as unclassified.

Visualisation and analysis

General approach

The core output from this project is a 'backend' database that holds >40,000 bibliographic research records, each one 'tagged' according to four key dimensions (actions, biomes, regions and taxa). Many articles could not be tagged with one of our key dimensions because the data was missing from the title, abstract, and keywords (e.g. in the case of taxa) or the model was uncertain how to classify an article (e.g. in the case of conservation actions) (Table 3). In order to convey the nature of this evidence base, we must visualise patterns in terms of the absolute and relative volume of evidence.

Table 3. Cross-tabulations for articles tagged with key dimensions.

	Actions	**Biomes**	**Taxa**	**Regions**
Actions	**11,492**			
Biomes	2,336	**8,090**		
Taxa	8,645	6,201	**35,279**	
Regions	8,197	6,687	25,397	**32,894**

In order to help identify research gaps (and clusters), we aimed to plot the volume of evidence identified across multiple dimensions concurrently. This is challenging given the four dimensions we have investigated (actions, biomes, regions, taxa), and given the hierarchical, complex nature of some of these (in particular, taxa, regions and actions). As a result of this complexity and

multidimensionality, we supplement our static figures here with an interactive platform, where users can decide the scale of each dimension and which dimensions to plot together. The plots in our results are therefore intended to act as an indication for the kind of information held within our database.

Evidence atlases and heat maps

We present the volume of evidence identified in our initial set of studies using a set of ontologies. The methods that we have developed can be very easily refined to incorporate different ontologies (e.g. author affiliations), using further refinements to the existing ontologies (e.g. functional groups of taxa), and based on modified sets of research records (e.g. incorporating biodiversity action plans and other grey literature). The results are therefore just an indication of what is possible. Further consultation with expert stakeholders would be necessary to both: 1) confirm whether a lack of records corresponds to a sensible research gap; and 2) establish additional sources of information and ontologies on which further refinements could be based. Such refinements could be conducted rapidly with minimal human inputs.

Results

Summary

The numbers of records and actions obtained is presented in Figure 1, above. In addition to the descriptions in the text, here, we have developed an interactive web site that allows the user to diver deeper into various aspects of the ontologies to examine the evidence across the following dimensions and at any desired level of detail. This evidence portal will continue to develop in terms of its interactivity and user experience to support learning and understanding. The website is available here: https://insectconservation.github.io/.

In the following pages, we focus on the results of the semi-supervised topic modelling and the clustering of studies across the four key dimensions: actions, biomes, regions, and taxa.

Actions

Our action tagging algorithm was able to identify actions in 23.0% of the evidence base at the two highest levels of the conservation actions ontology (Figure 2). The second level includes broad classes of actions (e.g. "Habitat & natural process restoration") without concrete action steps or management recommendations. We were able to tag 8.1% of articles with more specific conservation strategies (e.g. "increase aquatic habitat complexity") in the third level of the ontology (Figure 3a-f), and 7.9% at the fourth level of specificity in conservation actions (e.g. "drill holes in concrete river beds and river banks"). Beyond the third and fourth levels, the number of actions that were tagged is highly dependent on the number of ontology items at that level, which tapers off rapidly after the third level as actions get more specific and have fewer nested actions. There are 541 ways to be tagged at the third level, 355 at the fourth level, 207 at the fifth level, and only 84 at the sixth level where actions are highly specific (e.g. "undersow spring cereals with floral resources such as clover").

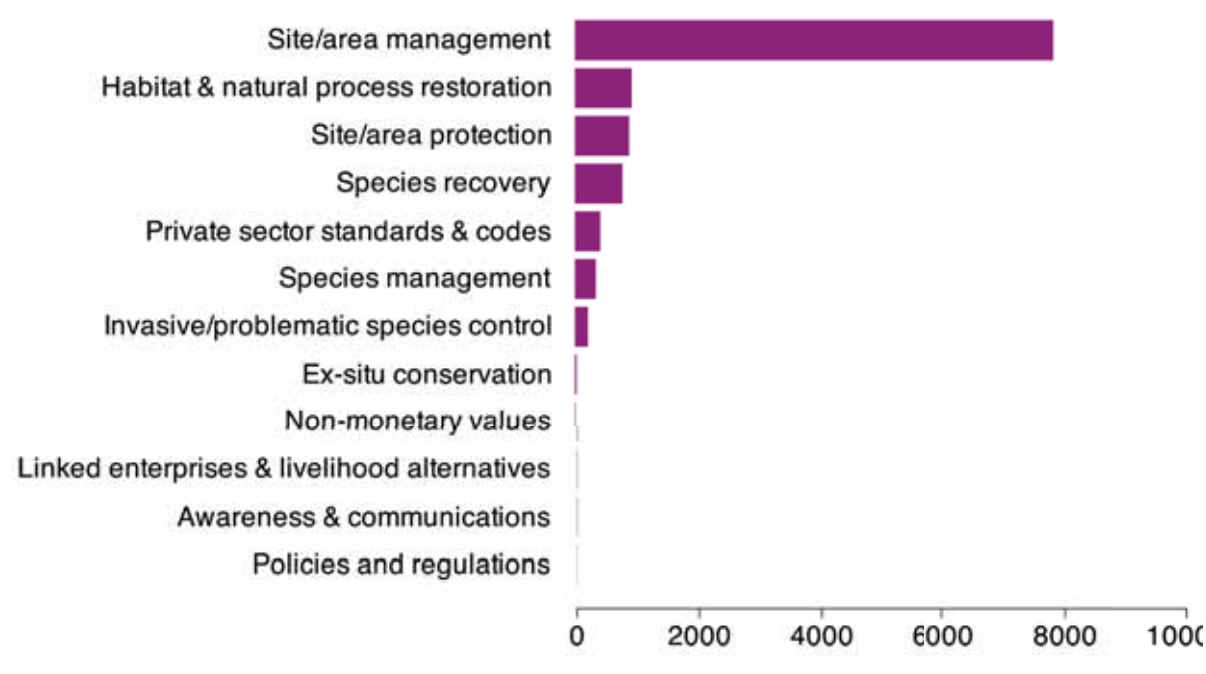


Figure 2. Number of studies tagged at the second highest level of the insect conservation actions ontology as a result of semi-supervised topic modelling.

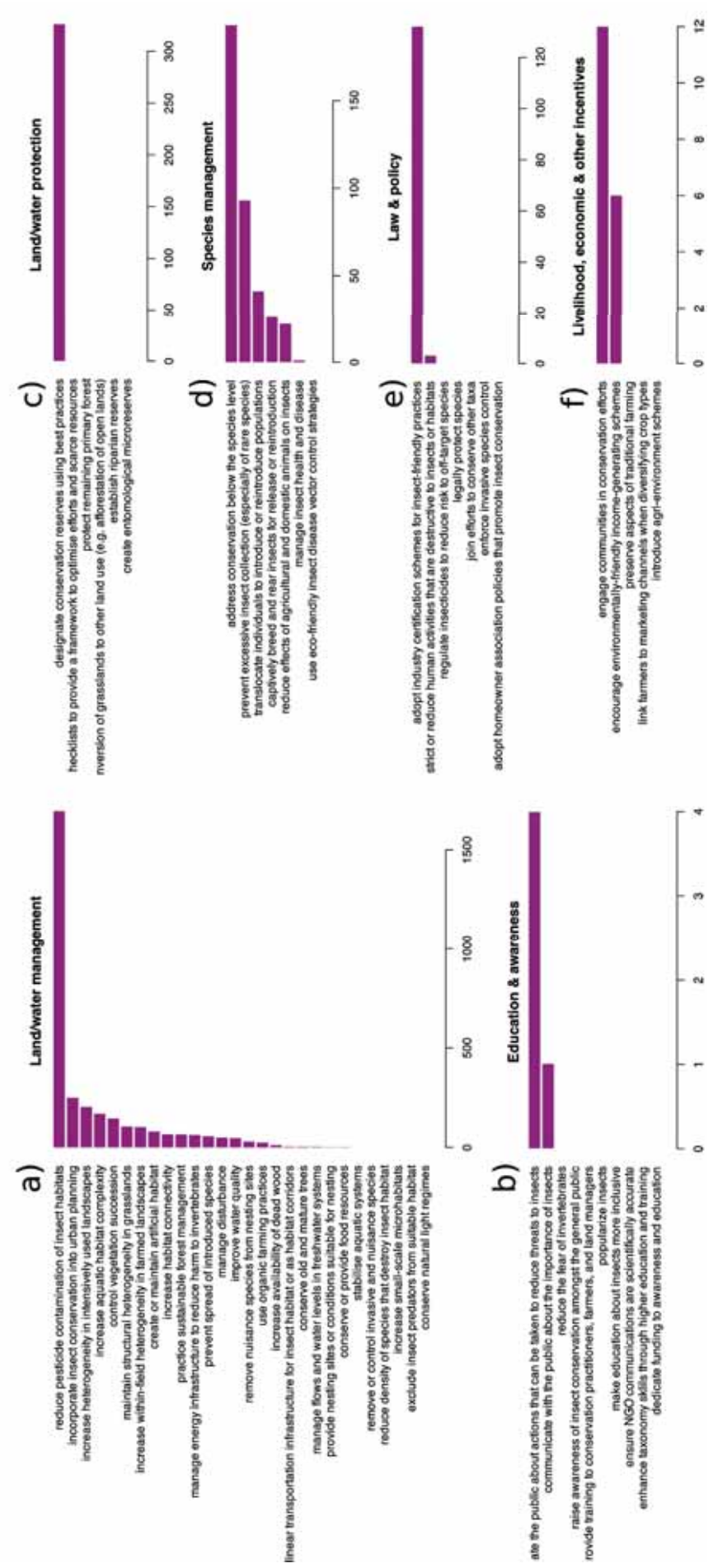


Figure 3. Number of studies tagged at the third level of the insect conservation ontology for those grouped at the highest level of the ontology as a) Land/water management, b) Education & awareness, c) Land/water protection, d) Species management, e) Law & policy, and f) Livelihood, economic & other incentives.

Biomes

We were able to classify records to biomes in 16.2% of cases (Figure 4). Most records were tagged as studies taking place in forests (41.4%) or terrestrial artificial habitats (28.1%) such as arable land, pastures, and urban areas, with fewer records tagged as savanna (9.9%), grassland (7.1%) artificial aquatic habitats (6.0%), shrubland (4.5%), or wetlands (2.9%). When considering records that were tagged with both a conservation action and a biome, there was a relative increase in the number of studies taking place in artificial aquatic (+5.2%) and terrestrial (+2.2%) habitats, and fewer studies taking place in forests (-3.4%) and wetlands (-1.4%) compared to the prevalence of all studies in those habitats. Cross-tabulations of insect taxa and biomes (Figure 5) corresponded fairly well to where subject experts would expect taxa to be most represented.

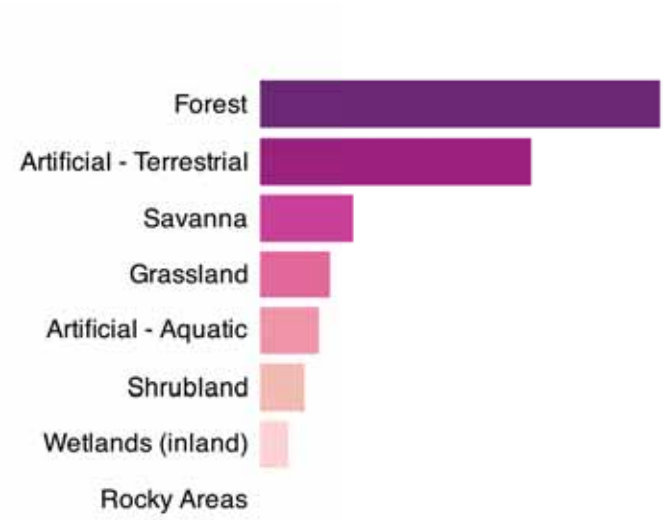
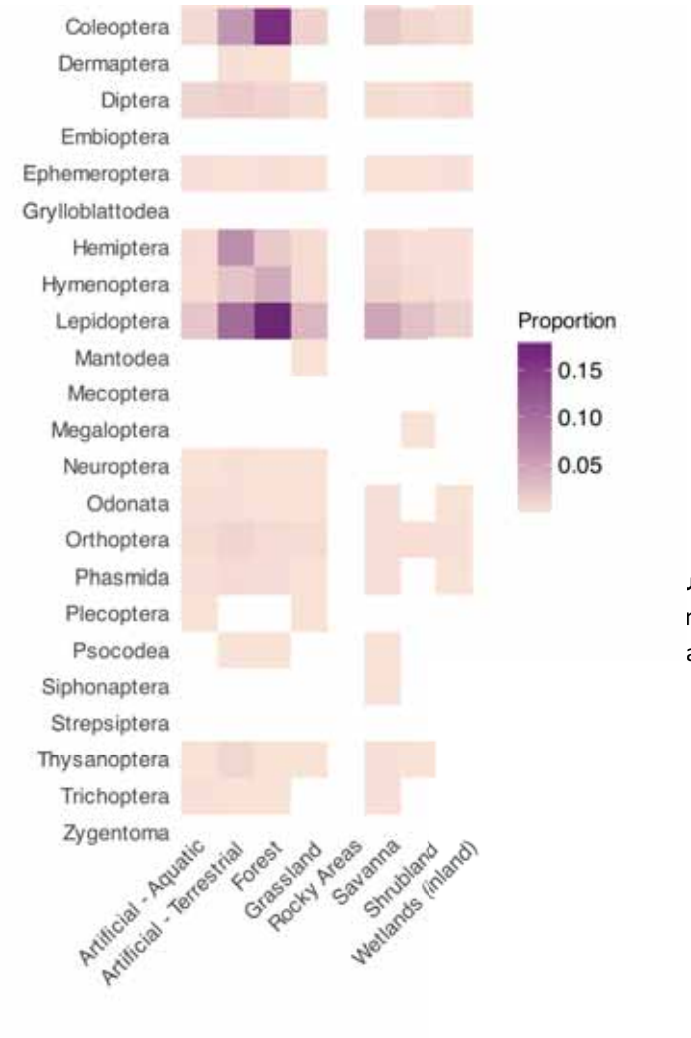


Figure 4. Number of records tagged to each biome in the subset of the EntoGEM database identified as conservation articles.



ure 5. Heatmap showing which taxa and me pairs occurred most frequently in the abase.

Regions

Our algorithm to parse geographical information tagged 65.4% of records with a study location at the continental level and 59.1% at the country level. Only 14.6% of all records were tagged with both a country and a conservation action (Figure 6) across 91 different countries, though the majority of studies tagged with both an action and country were from the United States (52.1%).

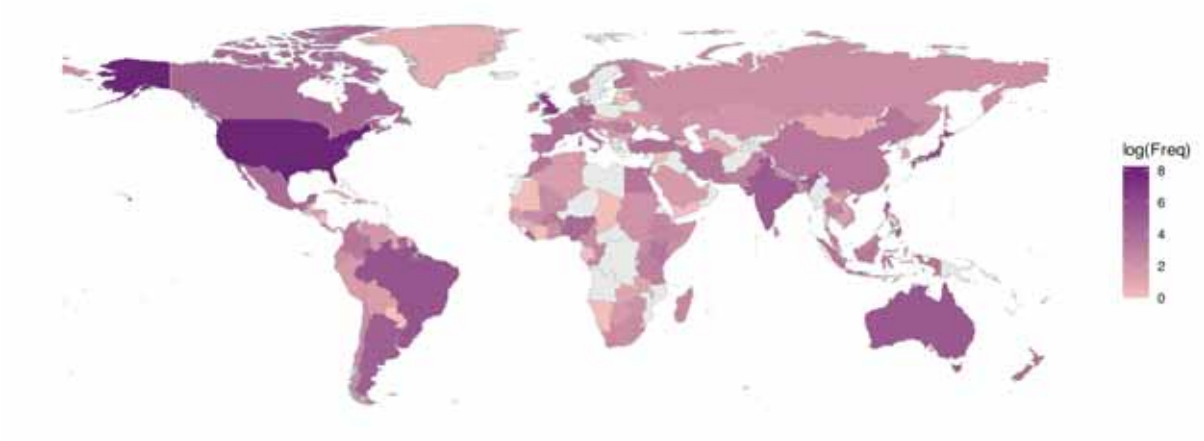


Figure 6. Map of countries coloured by log-frequency of studies tagged with both geographic information and conservation actions.

Taxa

Our taxonomy tagging approach was able to identify insect taxa at the order level in 70.6% of records. The most well-studied taxa (Figure 7a) were butterflies and moths (Lepidoptera, 37.2% of studies), beetles (Coleoptera, 19.7%), flies and mosquitos (Diptera, 13.2%), true bugs (Hemiptera, 11.0%), bees, wasps, and ants (Hymenoptera, 9.6%), and grasshoppers (Orthoptera, 3.1%), with all other taxa represented by less than 2% of studies. Although some taxa were less well-represented in the subset of articles tagged with both conservation actions and taxa, such as flies and mosquitos (Diptera, -4.6%), or increased in representation such as true bugs (Hemiptera, +2.4%) and bees and wasps (Hymenoptera, +2.1%), most remained fairly consistently represented in the conservation actions subset, likely due to the high proportion of studies that were tagged with taxa.

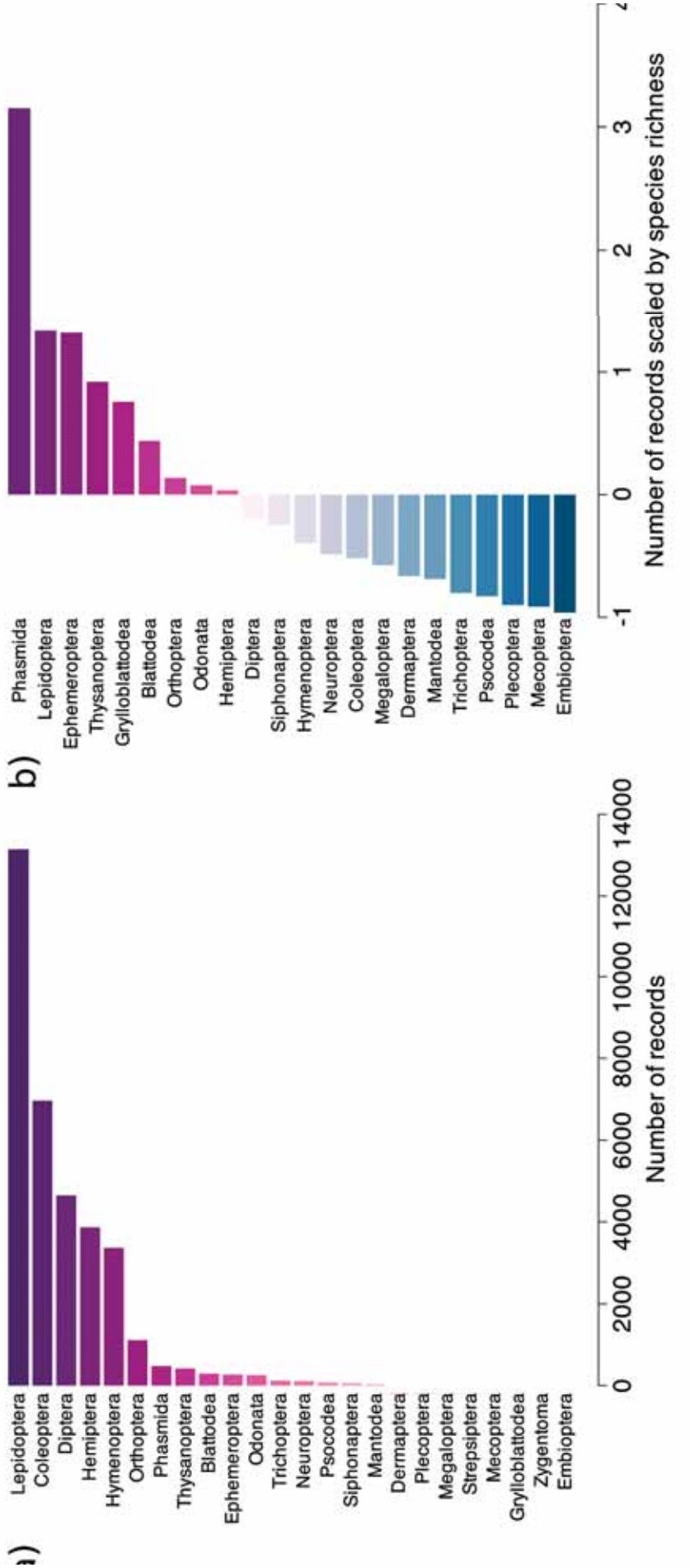


Figure 7. Number of records tagged to each insect taxonomic order within the subset of EntoGEM articles matching conservation terms as a) raw counts, and b) proportional to estimated species richness of each order.

Proportional to the estimated species richness of each insect order (Figure 7b), the most over-studied taxa were walking sticks (Phasmida), butterflies and moths (Lepidoptera), mayflies (Ephemeroptera), and thrips (Thysanoptera), and the most under-studied taxa were webspinners (Embioptera), scorpionflies (Mecoptera), stoneflies (Plecoptera), and caddisflies (Trichoptera). The discrepancies in the number of studies compared to species richness is likely due to the charismatic nature (e.g. walking sticks, butterflies) of the most over-studied taxa, their association with biological phenomena (e.g. mayflies) or their status as agricultural pests (e.g. many moths and thrips). Many of the taxa that were understudied are aquatic for at least part of their life cycle (e.g. Trichoptera, Plecoptera) and may be described in the literature as "aquatic invertebrates" rather than by scientific name due to sampling methods, and therefore were missed by our dictionary tagging method because "aquatic invertebrates" cannot be uniquely identified to a taxonomic group.

Heat maps

Figures 8 to 10 show heat maps for the corpus: identifying where across conservation actions and, respectively, biome, taxa, and regions the evidence is clustered. Of particular note are the small number of clusters of evidence: for example, reducing pesticides in artificial terrestrial biomes (Figure 8), and designating conservation areas for Lepidoptera (Figure 9). The vast majority of other combinations of variables show no evidence (white squares), and a number show a small proportion of studies (light pink).

There are no obvious patterns related to geography, other than the preponderance of actions related to pesticide from studies in Europe and North America. No other clear discrepancy is immediately apparent.

The heat maps are best interpreted by experts as a baseline against which assumptions about expected volumes of evidence can be tested. On their own, they may highlight ‘sensible gaps’ due to a low relative importance of a topic, for example where taxa- or biome- specific actions are not relevant/possible.

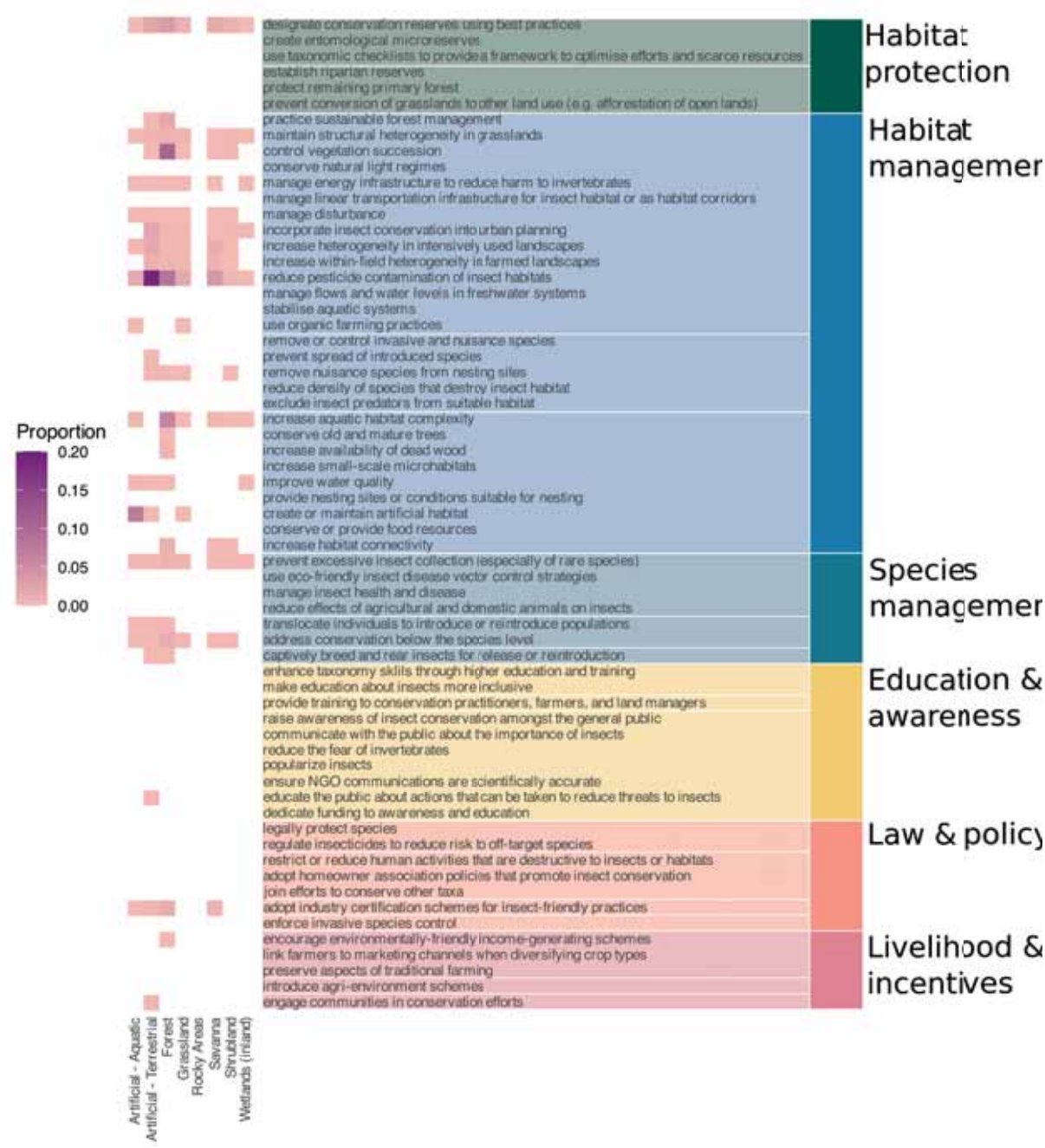


Figure 8. Heatmap showing the number of records tagged at the third level of specificity in the insect conservation actions ontology across biomes. Groupings of actions at the second level of the ontology are indicated with pale coloured boxes behind the specific actions, and the first level of the ontology is shown with labels and dark coloured boxes on the right of the figure.

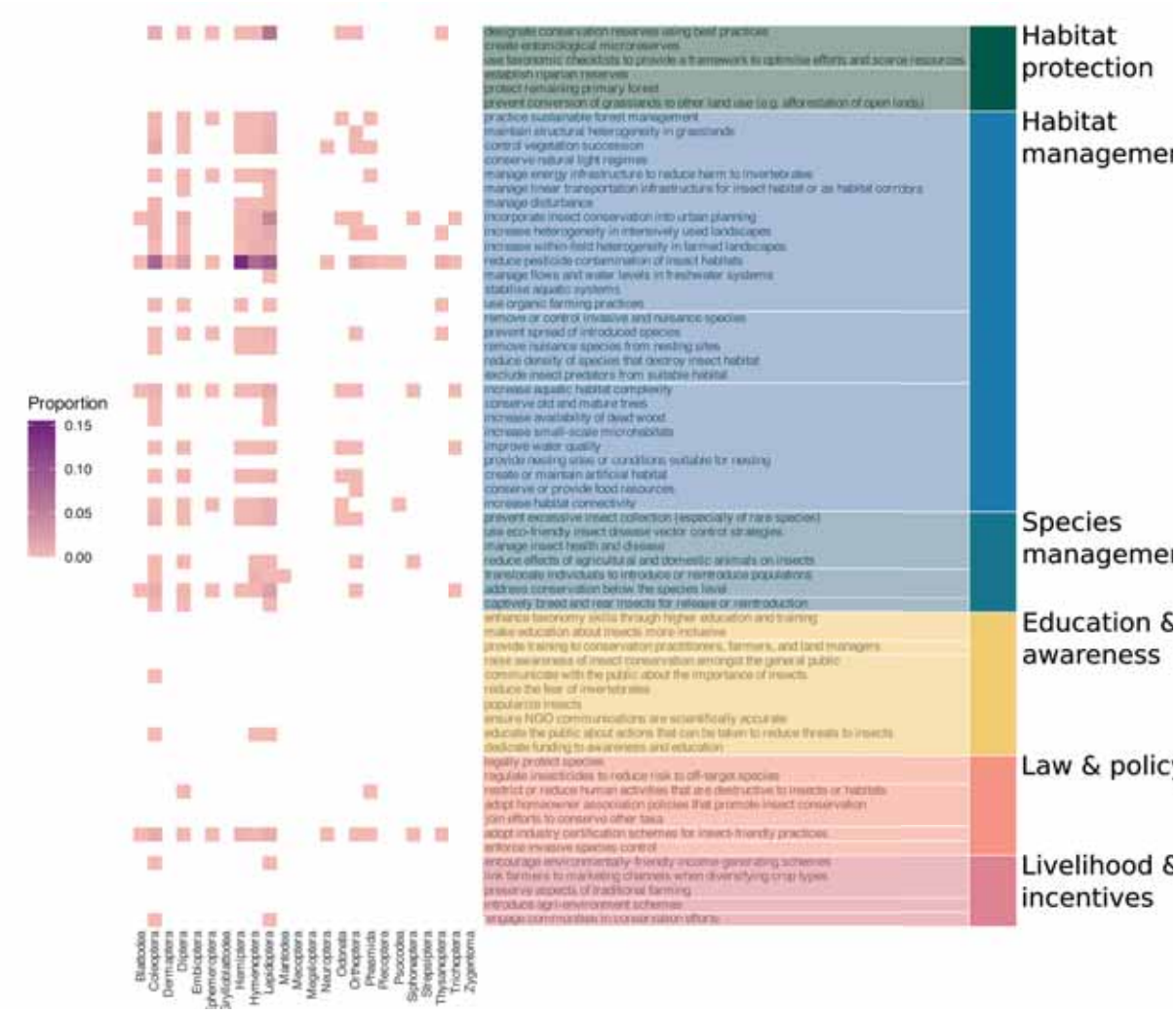


Figure 9. Heatmap showing the number of records tagged at the third level of specificity in the insect conservation actions ontology across insect taxonomic orders. Groupings of actions at the second level of the ontology are indicated with pale coloured boxes behind the specific actions, and the first level of the ontology is shown with labels and dark coloured boxes on the right of the figure.

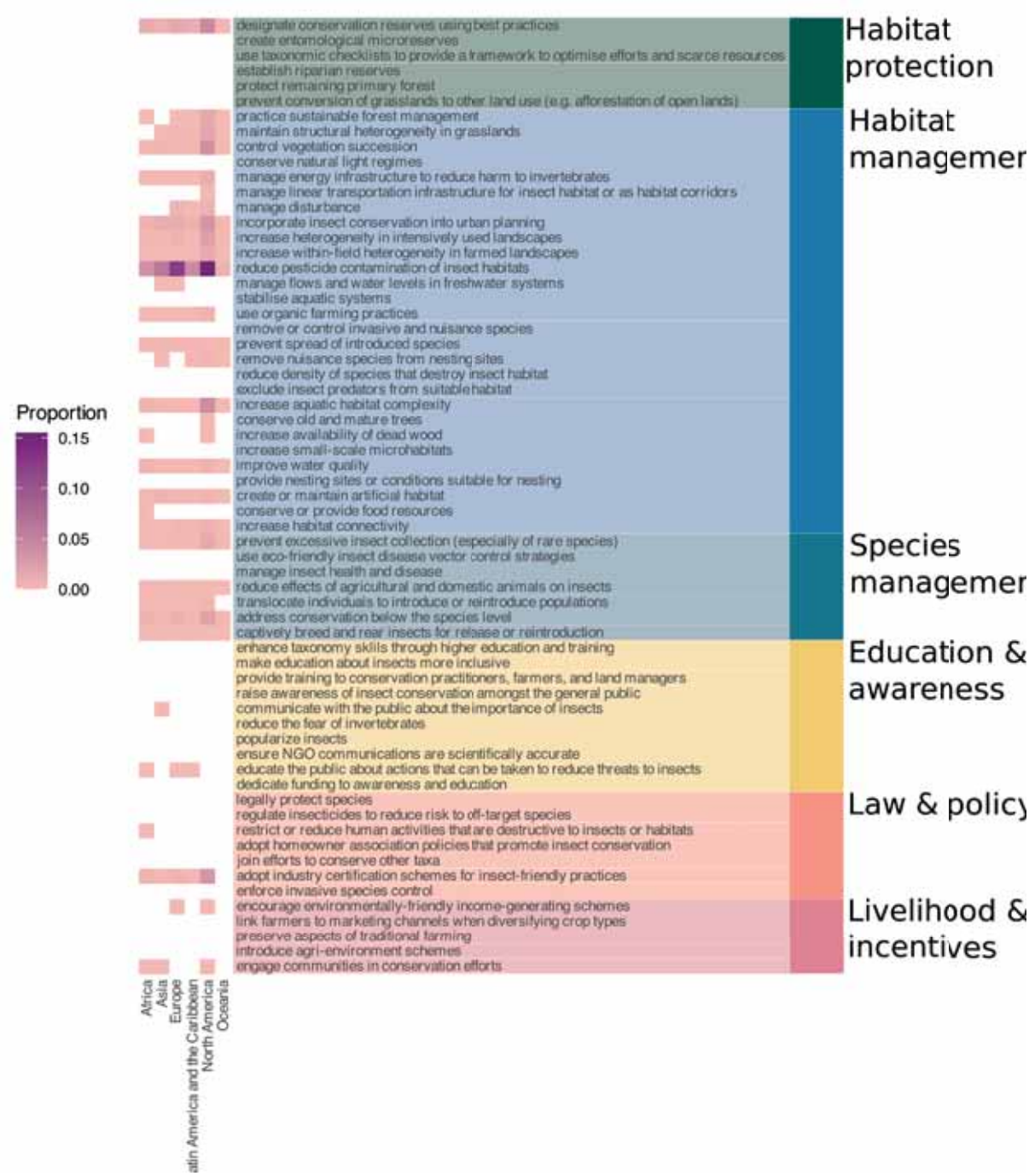


Figure 10. Heatmap showing the number of records tagged at the third level of specificity in the insect conservation actions ontology across continents. Groupings of actions at the second level of the ontology are indicated with pale coloured boxes behind the specific actions, and the first level of the ontology is shown with labels and dark coloured boxes on the right of the figure.

Discussion and conclusions

Although this map was completed semi-automatically and rapidly, it reveals patterns that correspond to what many subject experts suspected was true of the literature on insect conservation actions—in many cases there is substantial evidence of effectiveness for insect conservation actions, but limited research on the broader theories of change that involve institutional, political, and social contexts that affect the efficacy of insect conservation actions.

Insect populations globally are at risk from the combined effects of habitat loss and habitat degradation, invasive species, pollution and pesticide application, and climate change. These proximal causes of population declines and ways to mitigate or reverse their effects have received the most study, as revealed by the clusters of knowledge in our map on the topics of habitat management and species management, such as reducing pesticide contamination of insect habitats, controlling vegetation succession, and increasing habitat complexity. Entomologists, ecologists, and conservation biologists have been accumulating evidence on these stressors for decades. There is clearly no lack of knowledge about which on-the-ground actions can be taken to bend the curve and reverse insect population declines globally; as Forister et al. (2019) said: “we know enough to act now.”

What we do not know is how to implement insect conservation actions and gain broad social and political support for insect conservation. Our map revealed large gaps in all areas of the ontology related to human dimensions of insect conservation: habitat protection, education and awareness, law and policy, and livelihood, economic, and other incentives. Conservation biologists are not broadly trained in these areas of expertise, instead focusing primarily on management of habitats and species, as reflected in the clusters in our maps. This highlights the greater need for integration of social science in conservation research to test whether actions that work in theory can also work in the ‘real world’, embedded in complex human contexts.

Our map identifies a large number of highly specific topics that may represent gaps and warrant further funding. For example: habitat protection (creating insect reserves, using taxonomic checklists for resource allocation, establishing riparian reserves, protecting forests, preventing grassland conversion), and habitat management (controlling and preventing invasive and non-native species).

Our work highlights a significant gap related to interdisciplinary research teams spanning natural and social science that bring together entomologists, conservation biologists, and social scientists with experts in environmental law, business and industry certification boards, inclusive education, communications and public relations, and urban planning and community development. Because public education and awareness was one of the largest research gaps in the map, it merits consideration for additional primary research funding for social scientists that involves communication and education experts who can help answer research questions such as “How do we get the public to care about insects?” This is a key challenge for insect conservation efforts, because many people view insects as nuisances or pests, or are even frightened of them and find them ugly and threatening. Some entomologists are beginning to focus efforts in this area (New 2018; Basset & Lamarre 2019; Hall & Martins 2020; Howard & Dyer 2020), however, there is substantial work to be done to garner public support for insect conservation.

Updatability

As these research areas grow and new evidence emerges, our map can be automatically updated to incorporate new studies. New searches can be run periodically (e.g. every three months) to capture newly-published studies, which can be tagged using the dictionary approach or the models trained from the initial dataset. The models themselves can also be re-trained periodically (e.g. once a year) to keep pace with changing language and terminology used to study insect conservation actions. The ontology of insect conservation actions can be modified, updated, and expanded to include new conservation actions that were not identified through this project. We are making the ontology publicly accessible, so conservation researchers and practitioners can suggest modifications that we can incorporate into future versions of the ontology which will be deposited in publicly accessible repositories (e.g. Zenodo) where they can be archived and identifiable with a digital object identifier (DOI). Similarly, the other ontologies (i.e. biomes, insect taxa, and geographic regions) can be updated to include more synonyms for biomes, newly described insect species, or changing names of geopolitical regions.

Because the methods, R package, and ontologies from this project are being made publicly available and fully replicable, other research teams can adopt the tools to conduct similar rapid assessments of the state of knowledge on a topic. Although this approach is not recommended for full evidence syntheses such as those that would be used to inform policy, it could be used for other efforts to make recommendations for funding, or to rapidly identify research gaps that could merit further consideration. The materials can also be used for other purposes. For example, the insect conservation ontology can serve as a reference for conservation managers and practitioners seeking guidance on the types of actions that could be taken at a site to conserve a species. The R package developed for this project, *topictagger*, can be used to save time during full systematic reviews and meta-analyses by suggesting tags to researchers who manually verify them, or by prioritising articles for screening based on their similarity to studies that have already been included.

Limitations

Unlike a systematic map, this project relied heavily on text analysis methodologies to identify and classify documents based on abstract. As such, we have necessarily made a number of assumptions around the quantity and quality of information provided in this corpus. Table 4 lists key limitations that may affect the accuracy of the findings of the project, along with planned and possible mitigation strategies. Because we relied on the pre-existing EntoGEM database to populate our systematic map, our results may be biased towards on-the-ground actions that were identified as knowledge clusters in our map. The EntoGEM database is searching for studies that measure population trends over time; a study of law or education is unlikely to document long-term insect population trends. The searches used to compile the EntoGEM database were extremely broad, however, so this is unlikely to have biased the map so strongly that it could result in the observed patterns.

Table 4. Limitations and possible mitigation strategies for the project's main methods.

Limitation	Possible mitigation strategies
Region algorithm will suffer from false positives (e.g. '...the convention on biological diversity in Rio de Janeiro...')	Validate algorithm on a random sample of records and include exceptions when discovered, to prevent tagging non-study region information
Automated tagging of ontologies not (yet) validated	Validate automated tagging by manually assessing a random sample of tagged records
Primary focus on English language and USA/UK English terminology (due to Conservation Evidence being UK-based, and expert stakeholder bias towards UK and USA, underlying geographical researcher bias in the broader literature)	Perform targeted searches and make calls for conservation actions from other regions/languages/terminologies

Conclusions

Partially automated methods can help to identify research clusters and gaps that are consistent with patterns expected by subject experts in a way that is reproducible and avoids confirmation bias. In this project, we have developed a semi-automated method to rapidly map the knowledge in a field of research by combining expert opinions and text mining approaches, and applied it to the topic of insect conservation to identify which insect conservation actions have been studied and which are urgent priorities for future research. Entomologists and conservation biologists have arguably accumulated enough evidence on the proximal causes of insect population decline and actions that can be taken to mitigate losses and reverse population trends. What is lacking is research on how to effectively convert these recommendations into policies that have broad public support and clear pathways to implementation. Future research on insect conservation should focus on the human dimensions of solutions, such as community development, public education, and law and policy.

References

Agra, H.e., Carmel, Y., Smith, R. & Ne'eman, G. (2016) Forest conservation: Global evidence for the effects of interventions.

Basset, Y. & Lamarre, G.P.A. (2019) Toward a world that values insects. *Science,* **364,** 1230-1231.

Berthinussen, A., Richardson, O.C. & Altringham, J.D. (2020) *Bat conservation: global evidence for the effects of interventions*. Pelagic Publishing Ltd.

Blei, D.M. & Lafferty, J.D. (2009) Topic models. *Text mining: classification, clustering, and applications,* **10,** 34.

Dicks, L.V., Showler, D.A. & Sutherland, W.J. (2010) *Bee conservation: evidence for the effects of interventions*. Pelagic Publishing.

Forister, M.L., Pelton, E.M. & Black, S.H. (2019) Declines in insect abundance and diversity: We know enough to act now. *Conservation Science and Practice,* **1,** e80.

Friedman, J., Hastie, T. & Tibshirani, R. (2010) Regularization paths for generalized linear models via coordinate descent. *Journal of statistical software,* **33,** 1.

Grames, E. & Haddaway, N.R. (2020) elizagrames/topictagger: v0.0.0.9. Zenodo.

Haddaway, N.R., Grames, E.M., Boyes, D.H., Saunders, M.E. & Taylor, N.G. (2020) What evidence exists on conservation actions to conserve insects? A protocol for a systematic map of literature reviews. *Environmental Evidence,* **9,** 30.

Haddaway, N.R., Macura, B., Whaley, P. & Pullin, A.S. (2018) ROSES RepOrting standards for Systematic Evidence Syntheses: pro forma, flow-diagram and descriptive summary of the plan and conduct of environmental systematic reviews and systematic maps. *Environmental Evidence,* **7,** 7.

Haddaway, N.R. & Westgate, M.J. (2019) Predicting the time needed for environmental systematic reviews and systematic maps. *Conservation Biology,* **33,** 434-443.

Hall, D.M. & Martins, D.J. (2020) Human dimensions of insect pollinator conservation. *Current Opinion in Insect Science,* **38,** 107-114.

Howard, S.R. & Dyer, A.G. (2020) How to engage public support to protect overlooked species. *Animal Sentience,* **4,** 25.

James, K.L., Randall, N.P. & Haddaway, N.R. (2016) A methodology for systematic mapping in environmental sciences. *Environmental Evidence,* **5,** 7.

Marshall, I.J. & Wallace, B.C. (2019) Toward systematic review automation: a practical guide to using machine learning tools in research synthesis. *Systematic Reviews,* **8,** 163.

New, T.R. (2018) Promoting and developing insect conservation in Australia's urban environments. *Austral Entomology,* **57,** 182-193.

Smith, R.K. & Sutherland, W.J. (2014) *Amphibian conservation: global evidence for the effects of interventions*. Pelagic Publishing Ltd.

Stansfield, C., Thomas, J. & Kavanagh, J. (2013) 'Clustering'documents automatically to support scoping reviews of research: a case study. *Research synthesis methods,* **4,** 230-241.

The Collaboration for Environmental Evidence (2018) *Guidelines and Standards for Evidence synthesis in Environmental Management. Version 5.0.*

Westgate, M.J., Haddaway, N.R., Cheng, S.H., McIntosh, E.J., Marshall, C. & Lindenmayer, D.B. (2018) Software support for environmental evidence synthesis. *Nature Ecology & Evolution,* **2,** 588-590.

Wright, H., Ashpole, J., Dicks, L., Hutchison, J. & Sutherland, W. (2013) ENHANCING NATURAL PEST CONTROL AS AN ECOSYSTEM SERVICE.

Stockholm Environment Institute
Linnégatan 87D, Box 24218

104 51 Stockholm, Sweden

Tel: +46 8 30 80 44

Contact:
neal.haddaway@sei.org
eliza.grames@uconn.edu

visit us: sei.org
@SEIresearch
@nealhaddaway
@elizagrames