Methods in Ecology and Evolution

British Ecological Society

Methods in Ecology and Evolution 2016, **7,** 1262–1272 doi: 10.1111/2041-210X.12602

REVIEW

Automated content analysis: addressing the big literature challenge in ecology and evolution

Gabriela C. Nunez-Mir[1], Basil V. Iannone III[1], Bryan C. Pijanowski[1], Ningning Kong[2] and Songlin Fei[1]*

[1]*Department of Forestry and Natural Resources, Purdue University, West Lafayette, IN 47907, USA; and* [2]*Purdue University Libraries, Purdue University, West Lafayette, IN 47907, USA*

Summary

1. The exponential growth of scientific literature – which we call the 'big literature' phenomenon – has created great challenges in literature comprehension and synthesis. The traditional manual literature synthesis processes are often unable to take advantage of big literature due to human limitations in time and cognition, creating the need for new literature synthesis methods to address this challenge.

2. In this paper, we discuss a highly useful literature synthesis approach, automated content analysis (ACA), which has not yet been widely adopted in the fields of ecology and evolutionary biology. ACA is a suite of machine learning tools for the qualitative and quantitative synthesis of big literature commonly used in the social sciences and in medical research.

3. Our goal is to introduce ecologists and evolutionary biologists to ACA and illustrate its capacity to synthesize overwhelming volumes of literature. First, we provide a brief history of the ACA method and summarize the fundamental process of ACA. Next, we present two ACA studies to illustrate the utility and versatility of ACA in synthesizing ecological and evolutionary literature. Finally, we discuss how to maximize the utility and contributions of ACA, as well as potential research directions that may help to advance the use of ACA in future ecological and evolutionary research.

4. Unlike manual methods of literature synthesis, ACA is able to process high volumes of literature at substantially shorter time spans, while helping to mitigate human biases. The overall efficiency and versatility of this method allow for a broad range of applications for literature review and synthesis, including both exploratory reviews and systematic reviews aiming to address more targeted research questions. By allowing for more extensive and comprehensive reviews of big literature, ACA has the potential to fill an important methodological gap and therefore contribute to the advancement of ecological and evolutionary research.

Key-words: concept map, literature review, machine learning, quantitative review, research synthesis, text mining, topic modelling

Introduction

In this age of digitized information and highly developed communication systems, the explosive velocity with which information is being generated and made readily available has greatly challenged our ability to comprehend it, resulting in new hurdles for the advancement of science – a challenge akin to big data that we call 'big literature'. In parallel to the generation of big data (Laney 2001), the generation of big literature is characterized by the high velocity, volume and variability of the literature being generated. Simply illustrated, a basic search for publications of scientific content (i.e. research articles, proceedings papers, letters, notes and reviews) under the topic category of 'ecology' on Web of Science (2015) from 1950 to 2014 generated 125 000 research articles, more than half of which occurred in the last 7 years (Fig. 1).

Because this growth in available literature is not paralleled by an increase in available time or cognition, literature synthesis becomes progressively more difficult, hindering the advancement of science (Stockwell *et al.* 2009). Literature synthesis is crucial for the advancement of science as it provides the basis and theoretical groundwork on which conceptual frameworks are created and general theories are developed (Arnqvist & Wooster 1995). This process involves the sifting, classifying and simplifying of published research findings, methods, theories or applications with the goal of integrating past literature, critically analysing the existing literature or identifying issues that are central to a field (Cooper, Hedges & Valentine 2009). As publications continue to accumulate over time at rates difficult to embrace (Fig. 1), the proportion of available literature covered in each review study is substantially reduced. Such reductions result in smaller representations of the entire literature,

*Correspondence author. E-mail: sfei@purdue.edu

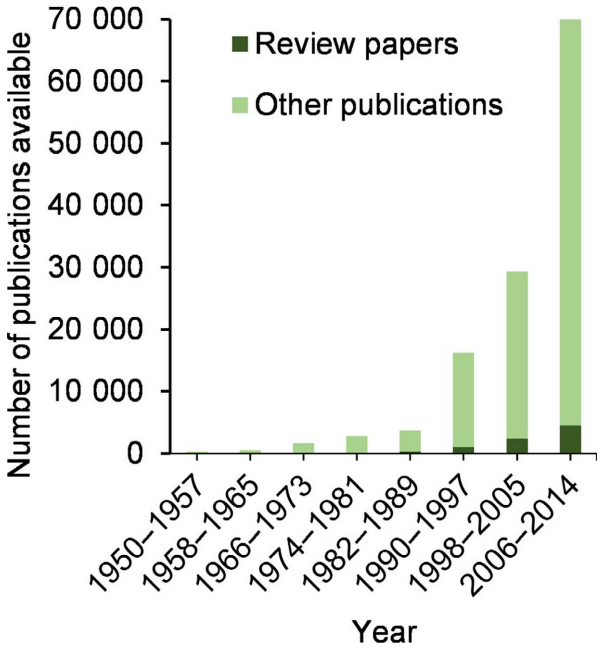

Fig. 1. Number of publications classified under the topic of ecology by Web of Science available each year from 1950 to 2014. Stacked bars indicate the total number of articles (review articles and other publications) published in 7-year intervals. A cumulative total of 125 000 publications were available by 2014.

potentially hindering comprehensive in-depth reviews and resulting in biased syntheses of findings.

Literature synthesis, here distinguished from meta-analyses of results from tables/figures in the form of effect sizes, slopes, etc., can be either quantitative (e.g. 'vote counting') or qualitative (e.g. narrative reviews, systematic summaries). However, the vast majority of review articles in ecology and evolutionary biology focus on the qualitative synthesis of literature in the form of narrative reviews (in *sensu* Koricheva, Gurevitch & Mengersen 2013). Narrative reviews, along with the other literature synthesis techniques, have certain limitations. First, these literature synthesis techniques are unable to handle large numbers of studies, resulting in under-sampling and the potential limitations associated with it (e.g. biased estimates, incorrect conclusions). Secondly, even when an explicit methodology to identify, select and analyse articles has been set *a priori*, a certain degree of subjectivity with which papers are chosen still persists, at times driving researchers to inadvertently showcase exemplary studies in their reviews (Pullin & Stewart 2006). There is a growing need to find new methods that can assist the synthesis of the growing corpora of big literature efficiently and objectively.

In this paper, we introduce automated content analysis (ACA), a method for qualitative and quantitative literature synthesis not yet adopted in the ecology and evolutionary biology field. ACA is a text-mining tool that uses text-parsing and machine learning, a subfield of computer science that focuses on pattern recognition and making predictions from data, to identify and define concepts/topics (hereafter referred to simply as concepts) within a body of literature. ACA utilizes a suite of statistical algorithms capable of discovering and describing concepts in a large body of text (Blei 2012a). For this reason, ACA is comparable to feature- or object-based classification in remote sensing (Walter 2004; Blaschke 2010). The concepts that ACA discovers are defined as groups of words that are strongly correlated in the literature and therefore are likely to represent a common theme or idea (Alexa & Zuell 2000; Smith & Humphreys 2006; Krippendorff 2013). These concepts are then used as categories by which to classify surveyed literature. The use of concepts to classify text distinguishes ACA from simple word-frequency counts (e.g. word clouds), as it is able to account for semantic and linguistic complexities, such as synonyms, co-occurrence frequencies and sentence construction (Roberts 2000). As evidenced by various studies in other fields (Travaglia *et al.* 2011; Zhao, Zou & Chen 2014), ACA offers considerable advantages to scholars who desire a thorough and comprehensive assessment of the existing literature on any topic with extensive research. Here, we introduce readers in the fields of ecology and evolution to ACA by providing an overview of the general conceptual framework and process of ACA, and by illustrating the utility of ACA through two examples of different applications of the method. We then introduce a variety of ACA tools and finalize with a discussion of the considerations that must be made to maximize the method's unique utility. Our goal is to encourage ecologists and evolutionary biologists to use and expand upon ACA, as have scientists from many other fields, to address the challenges of synthesizing big literature.

Overview of ACA

Automated content analysis refers to a suite of algorithms that use probabilistic models, called 'topic models' or 'concept mapping' models (Blei 2012a), to discover the hidden thematic composition of a body of literature. We use the term *thematic composition* to refer to the overarching themes in a body of literature, the frequency at which they appear and the relationships among them. The goal of these algorithms is to identify these themes and to categorize literature according to the presence of these themes.

The history of ACA can be traced back to the late 1990s. One of the first models to exploit the statistical properties of text corpus, Latent Semantic Indexing (LSI; Papadimitriou *et al.* 1998), was proposed as an information retrieval method able to capture the underlying semantics of a body of literature via linear algebra techniques. This model served as a theoretical basis for the field of information retrieval, topic modelling and concept mapping. Yet, perhaps the most influential and commonly used model is Latent Dirichlet Allocation (LDA; Blei, Ng & Jordan 2003), a three-level hierarchical Bayesian model designed to iteratively infer the concepts present in a body of text and the proportion of the literature in which they occur. LDA represented a step forward from previous models by discarding the 'bag-of-words' assumption – that the order of words in a document is unimportant – and by accounting for the exchangeability/interchangeability of particular words (Blei, Ng & Jordan 2003). Models beyond LDA have been

designed to identify higher levels of complexity of these hidden thematic structures, such as syntax, concept hierarchies, document networks and temporal trends in themes, furthering our ability to visualize and explore the literature (Blei 2012b).

The ACA process

Although ACA encompasses a group of algorithms and models, they are all based on the same fundamental approach. The fundamental process of ACA can be divided into three stages: identification, definition and text classification (Fig. 2). At the first stage, concept identification, the concepts by which the literature will be classified are determined. Some ACA tools do this through the use of concept seeds, which are single words that occur frequently in the literature, and that are therefore most likely to represent important concepts. These concept seeds serve to guide the identification and definition of concepts from the literature, and depending on the purpose of the analysis and the capacity of the ACA tool, can either be extracted from the literature through unsupervised seeding or provided by the researcher through supervised seeding.

The next stage of ACA is concept definition. During this stage, a thesaurus (i.e. the group of words that forms a concept) is compiled for each concept. Thesaurus building is accomplished by the topic model or concept mapping algorithm used by the ACA tool (e.g. LDA, LSI, non-negative matrix factorization, Leximancer). The output of the definition stage is a set of predominant concepts, each defined by their own thesaurus. The primary objective of the words in the thesaurus is to determine whether a concept occurs in a given portion of surveyed text (i.e. text segment). For this purpose, some ACA tools assign weights to each word in a concept's thesaurus based on their relevance to the concept. A given concept is determined to be present in a text segment when the cumulative weights of its thesaurus words reach a pre-chosen threshold (described further below).

In the third and final stage of ACA, text classification, the literature is classified by the concepts identified and defined in

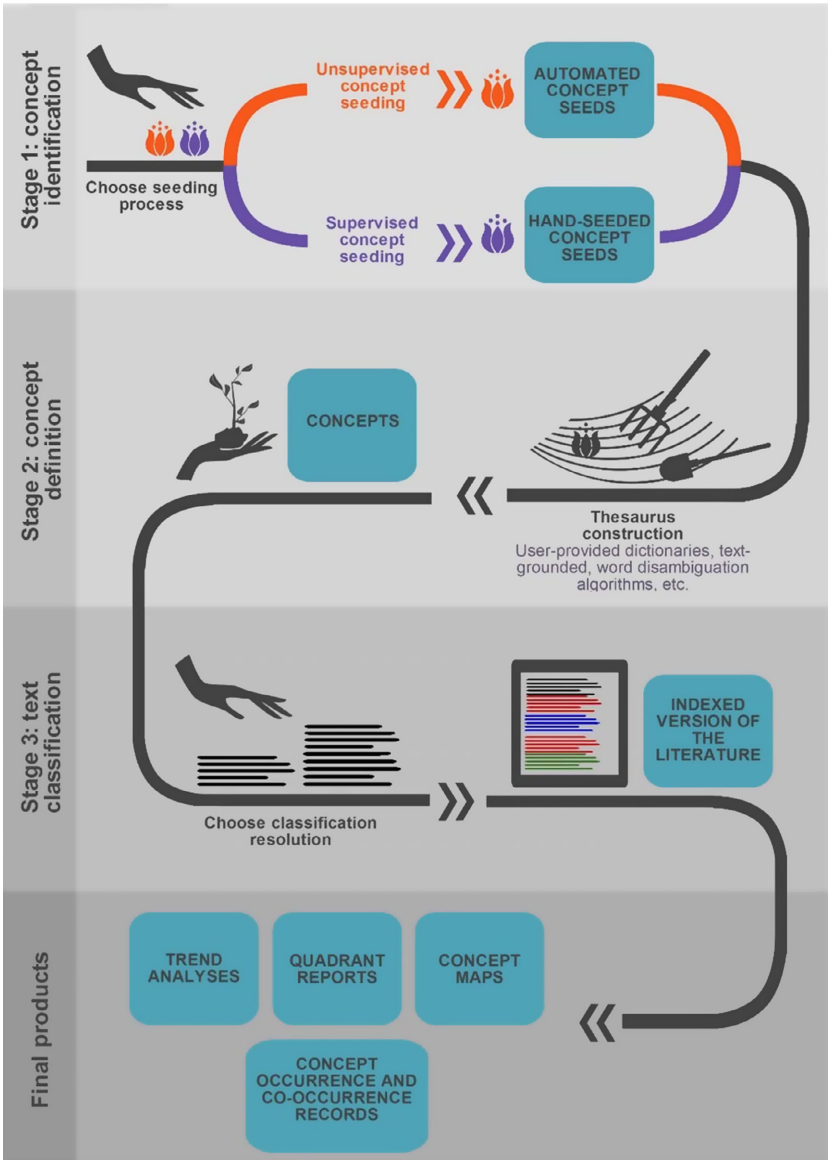

Fig. 2. Conceptual map of the stages of automated content analysis. Blue boxes indicate outputs from each stage.

the two previous stages. The categorization of the literature is generally performed at high resolutions (e.g. by sentence or by line segments, hereafter referred to as 'text segments'), but the coarseness of the classification, usually customizable, depends on the ACA system used and the purpose of the analysis. During text classification, a text segment is analysed for evidence of the occurrence of each concept.

After the ACA process is complete, the researcher gains access to a number of final outputs that allow further exploration of the results. For instance, certain ACA tools are able to generate graphic summaries of results (e.g. conceptual maps, quadrant reports or social network maps) that allow the researcher to visually comprehend the literature analysed, the concepts it contains and the associations among these concepts. In addition, ACA tools are able to provide statistical and concept co-occurrence data, such as records of concept frequency and co-occurrence frequency between two concepts, offering a quantitative illustration of the dominance of certain concepts and the strength of associations among concepts. These outputs are also suitable for a wide range of follow-up analyses, including comparisons of different bodies of literature and analyses of trends in the literature. In some systems, text segments classified under each concept may be visualized or retrieved as well.

Validation of ACA tools

As the field of ACA has progressed, so has the development of methods to evaluate the reliability (i.e. the consistency of results across measuring processes) and validity (i.e. the extent to which inferences reflect reality; Krippendorff 2013) of ACA tools (Wallach *et al*. 2009). Furthermore, several studies have evaluated these systems by comparing the results of ACA to those obtained by traditional manual methods. Smith & Humphreys (2006) undertook an extensive validation study of ACA in which they confirmed the validity, reproducibility and stability of Leximancer, an ACA software. An analysis of a large number of accident reports classified manually and by ACA found almost identical percentages between the manual and automated analyses, confirming the reliability of ACA and its utility in the analysis of these reports (Grech, Horberry & Smith 2002). When using ACA to analyse data on the variability in human experience and thinking, Penn-Edwards (2010) found ACA to be superior to manual methods in dealing with large amounts of data without bias and in identifying a broader span of syntactic properties. In a large-scale study of digital libraries, Newman *et al*. (2010) asked people to evaluate and score 500 individual concepts for semantic coherence across a wide variety of genres and domains and found that the ACA model scores were highly correlated to the human scores, even exceeding human inter-rater correlation.

Applications of ACA

The efficiency and versatility of ACA allows for a broad range of applications, including exploring the development of a scientific discipline by comparing conceptual trends between temporal periods of published research (Cretchley, Rooney & Gallois 2010; Nunez-Mir *et al*. 2015), analysing changing forest social values and their implications for ecosystem management (Bengston & Xu 2006) and giving unbiased estimates of categorical proportions in political science studies (Hopkins & King 2010). ACA can be applied to identify both important topics in the literature and research gaps. Furthermore, it can be used in preliminary studies to guide more focused, intensive synthesis efforts or to become familiarized with a new domain/subject without having to blindly sift through massive volumes of literature (Stockwell *et al*. 2009).

The qualitative and quantitative data obtained from ACA are particularly useful for both targeted systematic and broad exploratory reviews. With regard to targeted systematic reviews, supervised seeding can be used to define and classify predetermined concepts that are relevant to a specific *a priori* question or topic, even if these concepts are rare in the literature or possess more abstract definitions. These targeted reviews are able to provide insight about specific trends or aspects of the surveyed literature. For example, from a targeted ACA of 29,766 abstracts published over a 35-year period, Nunez-Mir *et al*. (2015) were able to ask whether the prominence of the concept restoration ecology in forestry science is increasing over time and if the focus of restoration-based forestry research has fluctuated over time. They did so by quantifying the proportion of the reviewed literature containing the concept *restoration ecology* and determining which other concepts were most strongly associated to *restoration ecology* in each decade. An exploratory ACA review, on the other hand, is akin to the broad, comprehensive coverage of the literature performed in narrative reviews. Using unsupervised seeding, ACA provides the researcher with an overview of the major concepts in the literature, along with a classified version of the literature reviewed, condensing large volumes of text into manageable subsections that can be further analysed or quantified.

The following sections illustrate these applications of ACA (i.e. targeted and exploratory reviews) through two reviews of published literature. We used Leximancer (V4; Leximancer Pty Ltd; Brisbane, Australia) to exemplify the ACA approach, because it is among the most advanced tools currently available, possessing a variety of unique features for analysis and visual representations. Furthermore, we chose Leximancer because there is a considerable body of literature available describing the method, validating it and exploring its variety of uses in other fields (Smith 2003; Smith & Humphreys 2006; Cretchley, Rooney & Gallois 2010; Penn-Edwards 2010).

Like all other ACA tools, Leximancer performs its analyses following the three stages of the ACA process. First, the Leximancer algorithm (Smith 2003) identifies concepts using concept seeds that can either be extracted from the literature through unsupervised seeding or provided by the researcher through supervised seeding. In the second stage of ACA, Leximancer defines each concept identified through a concept thesaurus built by an iterative, bootstrapping, machine learning algorithm derived from a word disambiguation technique (Yarowsky 1995). This algorithm first creates a

co-occurrence matrix with the co-occurrence frequencies of all concept seeds. From this co-occurrence matrix, it then finds words that co-occur frequently with a given seed word and infrequently elsewhere by identifying the nearest cells with peak values (i.e. nearest local maxima) for each concept seed (Smith 2003; Smith & Humphreys 2006; Stockwell *et al.* 2009). Each word in the concept thesaurus is then weighted for relevancy to the concept using a naïve Bayesian co-occurrence metric (Salton 1989) that creates a tighter binding of relevant terms to concepts by taking into consideration how often words co-occur, as well as how often they occur apart (Smith & Humphreys 2006). In the last stage of ACA, text classification, a text segment is considered to have enough evidence for a given concept if the summed weights of all of the concept's thesaurus words present in the text segment surpass a user-customizable classification threshold (ranging from 0·1 to 4·9, default 2·4). Increasing this threshold from the default will increase the amount of evidence required for a concept to be identified in a text segment. Increasing the classification threshold is particularly useful when classifying concepts that are more abstract (e.g. interdisciplinary, ecosystem health), which are typically more difficult to identify due to the inherent broadness or ambiguity of their definition. If a text segment is found to have enough evidence, it is classified under the given concept. The result is an indexed version of the literature, in which text has been finely classified. This information is then presented through tables of concept frequency and co-occurrence, and a wide variety of graphical and visual summaries to further explore the results of the analysis.

ACA FOR TARGETED SYSTEMATIC REVIEWS

To illustrate the utility of ACA in systematic reviews with a specific *a priori* question, we conducted a targeted ACA using the same body of literature used in Nunez-Mir *et al.* (2015; 29 766 article abstracts published from 1980 to 2014 in 14 leading forestry journals, for more information on how these articles were obtained, see Table S1, Supporting information). The anteceding paper used ACA on these abstracts to understand the knowledge gaps and the evolution of restoration ecology research in the context of forestry. Here, we use the same body of literature to demonstrate the utility of ACA. We illustrate how ACA can be used to determine the proportion of the analysed text discussing a specific phenomenon, in this case, forest fragmentation (Bhagwat, Kettle & Koh 2014) and to determine the concepts that are most strongly associated with this phenomenon. We hand-seeded the concept *fragmentation* (concept seeds: fragmentation, fragmented, fragmenting, fragmentation's) and performed ACAs on the literature from four distinct time periods (1980s, 1990s, 2000s and early 2010s) to detect temporal trends in forest fragmentation research published in forestry journals. Once the literature was classified, we generated a concept map limited to concepts that were associated to *fragmentation* in all decades. We also generated ranked lists of the concepts most strongly associated to *fragmentation* within each decade (i.e. concepts ranked by their likelihood or probability of co-occurring with fragmentation in a given text segment).

Reporting the results of this targeted ACA on fragmentation in depth is beyond the scope of this paper. Nevertheless, our findings clearly illustrate the utility of ACA for addressing a predetermined, specific question. For instance, our results revealed that although research on forest fragmentation represents a small proportion of forestry research, its prominence has increased by 500% from the 1990s to the early 2010s (Fig. 3a). Therefore, our ACA reveals fragmentation to be a concept of increasing importance to forestry, likely reflecting increasing changes in land use and the negative impacts that these changes can have on forests. Our results also reveal a shift in what topics are most often associated with forest fragmentation in forestry research. For instance, the concept map of the literature containing the concept forest fragmentation revealed three major themes: edges, habitat and land (Fig. 3b). The ranked lists of concepts most strongly associated to fragmentation generated for each time period indicate shifts in research focus. For example, the concepts matrix and roads detected in 2000–2009 were not detected in the prior two decades (Fig. 3c).

The results of our targeted ACA give us the opportunity to quantify and describe the existing literature on a topic in its entirety. The themes and concepts identified describe the major foci of the literature and are therefore able to give an overarching understanding of its content and act as a guide for more focused reviews. For example, our targeted ACA allowed us to delineate the trajectory of research on forest fragmentation throughout the decades, going from being a technical, stand-level issue (as suggested by the concepts prominent in early decades, such as *oak-pine* and *manager*), to a landscape-level, socioecological issue (as suggested by concepts prominent later on, such as *landscapes*, *agriculture*, and *roads*; Fig. 3c). For those interested in pursuing research in the topic of forest fragmentation, the concepts prominent in the most recent years can indicate where the attention of the field is at the moment and by doing so help to guide future research directions by revealing important research gaps. In addition, the table in Fig. 3c gives an indication of the strengths at which these concepts are associated with *fragmentation* by displaying the number of text segments containing both the listed concept and *fragmentation* (Count), and by providing a likelihood measure, which estimates the probability that the listed concept will co-occur with the concept of *fragmentation* given its total number of occurrences. These two measures complement each other, as they give indications of both directions of conditional probability (i.e. how likely it is for *fragmentation* to co-occur with a given concept given the total number of occurrences for *fragmentation*, and how likely it is for the concept to co-occur with *fragmentation* given the total number of co-occurrences for the concept). In this case, because the literature classified under *fragmentation* comprise such a small proportion of the total literature analysed, the likelihood of ranked concepts appears to be low, as these concepts are probably co-occurring with other concepts elsewhere in the analysed text.

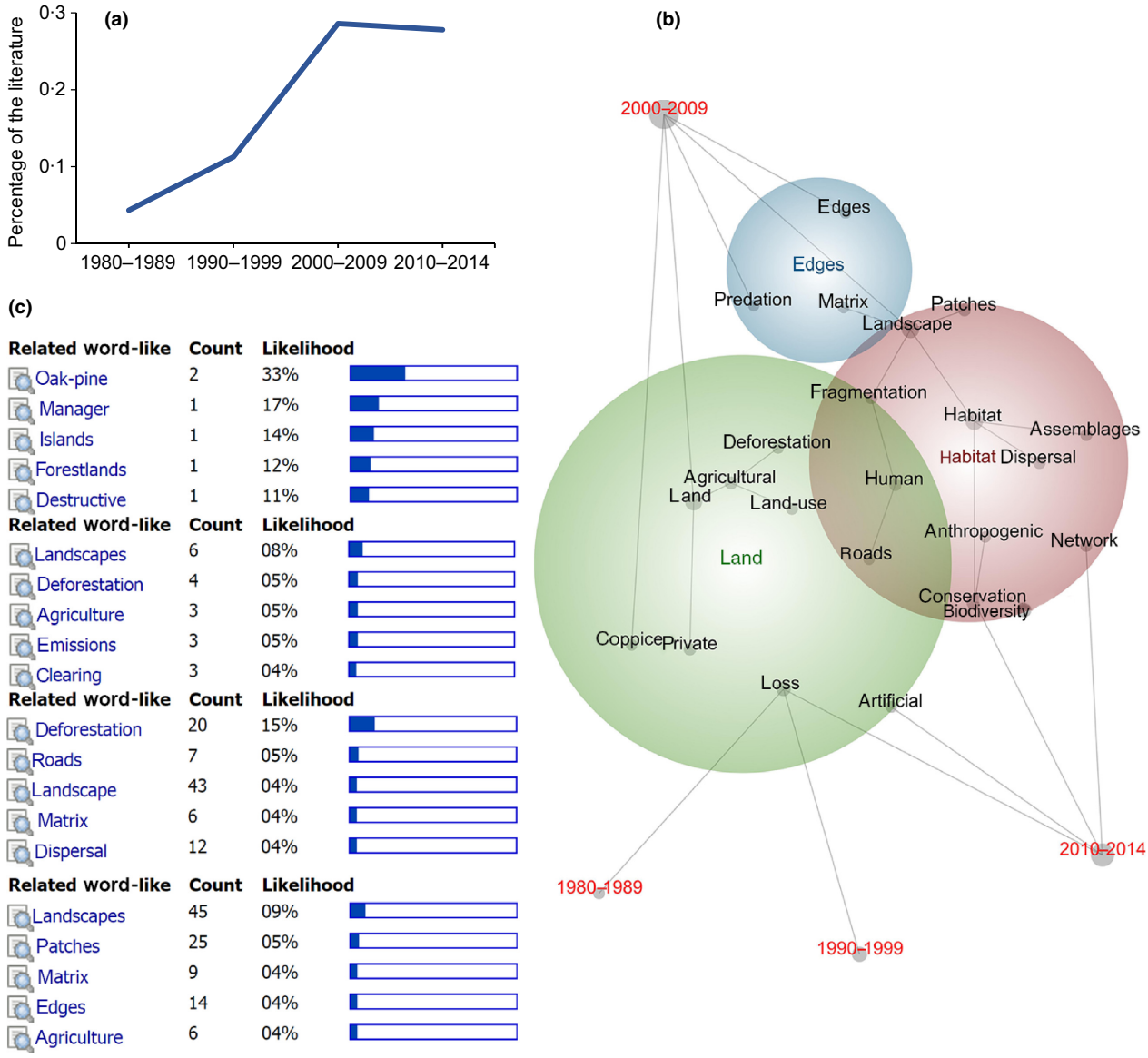

Fig. 3. Results of a targeted systematic automated content analysis (ACA) of all article abstracts published in fifteen forestry journals since 1980 (n = 29 766) to explore the extent and focus of research on forest fragmentation in the literature. (a) Proportion of the literature reviewed using ACA classified under the ACA concept *fragmentation*. (b) Concept map representing the major themes in the overall literature associated with *fragmentation* (coloured bubbles), the major concepts associated with the concept *fragmentation* (circles), and the complex relationships among these concepts (connecting lines). (c) Ranked lists of the concepts most highly associated with the concept *fragmentation* each decade, where 'counts' represent the number of text segments in which each concept co-occurs with the concept *fragmentation* and 'likelihoods' represent the probability of each concept co-occurring with the concept *fragmentation* (i.e. no. of text segments in which the concept co-occurs with *fragmentation*/no. of text segments it occurs overall in the literature).

ACA FOR EXPLORATORY REVIEWS

To demonstrate the application of ACA in exploratory reviews, we performed an ACA on 51 empirical studies investigating the effects of land use on exotic plant invasions (Appendix S1). These studies were previously reviewed manually by Vilà & Ibáñez (2011). We then compared the findings of our ACA to Vilà & Ibáñez (2011)'s manual review. Vilà & Ibáñez (2011) classified each study by which land-use driver or landscape attribute was studied (see supplementary materials in Vilà & Ibáñez 2011). In our review, we used Leximancer to identify the top 200 predominant concepts. We then re-classified both Vilà & Ibáñez's (2011) landscape categories and the ACA concepts that were relevant (i.e. concepts associated with land-use effects on invasion) into six new categories: human practices, fragmentation/edge effects, dispersal opportunities, biotic resistance, disturbance and metapopulation dynamics (Tables S2 and S3). These new categories, which represent different 'pathways' by which land-use affects invasion, were intended to standardize the classification systems used by Vilà & Ibáñez (2011) and by us (i.e. land-use drivers/landscape attributes vs. ACA concepts, respectively), thus facilitating our comparison. We then classified land-use drivers/landscape attributes and ACA concepts according to the pathway (or

pathways) with which they were associated. We compared the proportion of literature classified under each of the new categories as determined by ACA to the proportions determined manually by Vilà & Ibáñez (2011).

Our ACA review revealed the categories most discussed in the literature, as well as potential research gaps (Fig. 4). The most discussed categories were *human practices*, followed by *fragmentation/edge effects*. Concepts classified under these two categories appeared in 91 and 74% of all surveyed text segments, respectively. The least discussed category was *metapopulation dynamics*, found in <10% of surveyed text segments.

The results of our ACA review strongly evidence the ability of this method to provide results comparable to those of manual reviews, but in a fraction of the time. For instance, although the metrics used to record the proportion of the literature in each category differed between the manual and ACA review (proportion of articles vs. proportion of text segments, respectively), both methods produced similar trends (Fig. 4). Furthermore, ACA detected a considerably larger proportion of text segments for *disturbance* and *biotic resistance*, suggesting that the presence of these themes in the literature might be stronger than suggested from manual review. The differing trends detected by ACA emphasize the technique's potential utility in objectively identifying trends that may be disregarded or overlooked during manual analysis.

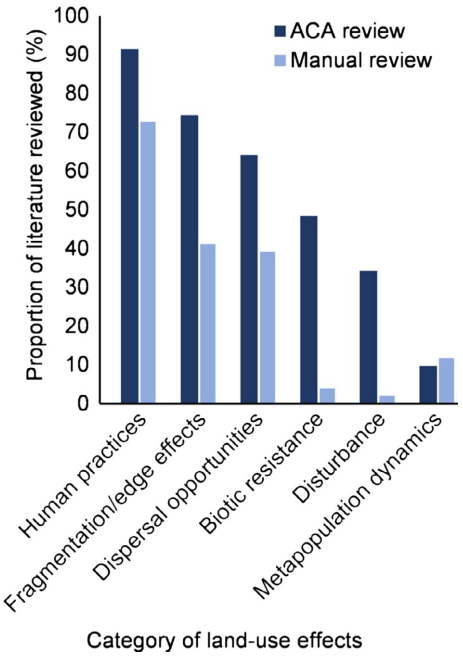

Fig. 4. Comparison between the results of an exploratory automated content analysis (ACA) review and a previously published manual review of 51 empirical studies on the effects of land use on biological invasions. Bars represent the proportion of literature exploring each category of land-use effects as determined manually and by ACA. Manual review used individual studies as a unit, while ACA used text segments.

ACA tools available

Many tools have been developed to facilitate the use of the existing topic model and concept mapping algorithms for ACA. Here we briefly introduce eight of these frequently used tools, including R packages and python libraries, and compare their features and capabilities (detailed in Table 1). We made these comparisons using the information provided in each tool's manual, website or vignette (see source of reference in Table 1). We chose to highlight features that would be of interest to researchers new to ACA and that would therefore inform and facilitate the process of choosing an ACA tool that best fits the researcher's needs.

We first distinguished tools that require coding (Mallet, Stanford TMT, topicmodels, lda, stm, Gensim and sklearn) from those that feature user-friendly graphic user interfaces (GUI; Leximancer and Google TMT). Tools that have a GUI make ACA uncomplicated and accessible and may therefore be better options for individuals having little or no coding experience. On the other hand, tools that require coding provide users with a deeper understanding of the methods and the algorithms used by the tool, as the user is responsible for running each step. Furthermore, coding gives the user more control over the parameters used in the analyses, allowing the user more freedom to adjust parameters to fit his or her needs.

Next, we wanted to determine which tools required extra steps to pre-process documents. Mallet, topicmodels, Gensim, lda and sklearn all require the user to run extra lines of code to convert the source document into the required format. Pre-processing entails various procedures such as cleaning the text data (e.g. eliminating tags and non-alpha numeric characters), deleting stopwords (e.g. and, about, towards, others) and even vectorizing the text data. These processes can take from minutes to hours to complete depending on the size of the text corpus and the ACA tool used.

We also identified whether or not tools are able to identify and measure interrelationships among concepts. Leximancer and stm were the only tools found to allow the user to go beyond identifying the concepts discussed in the literature and the frequency in which they are discussed by also measuring the associations between concepts. This feature adds a layer of complexity that is useful when the researcher is interested in a particular concept, as opposed to understanding the literature as a whole. Furthermore, exploring concept interrelationships give a better idea of how and in what context a concept of interest is being discussed.

In addition to analysing interrelationships, we wanted to identify the tools that allowed for hand-seeding (user-provided concept seeds), a feature that is particularly useful when performing targeted reviews. This feature gives the user the opportunity to analyse concepts that might be difficult to identify and define automatically (i.e. through unsupervised seeding) due to their being rare and/or having abstract definitions (e.g. ecosystem services, interdisciplinary). From the tools surveyed, Leximancer appears to be the only programme that allows hand/supervised seeding.

We also evaluated what ACA tools provide with respect to analytical outputs. Most of the tools compared (Table 1) give the user access to raw data results in the form of co-occurrence

Table 1. Comparison of the capabilities and features of nine commonly used automated content analysis (ACA) tools. Table compares ACA software, R packages and python libraries

		Features user-friendly GUI	Requires user coding	Does not require pre-processing of documents	Analyses inter-relationship among concepts	Allows for extraction of raw results	Allows for hand-seeding user-defined concepts	Generates graphic/visual outputs	Open source
Software	Mallet* (McCallum 2002) http://mallet.cs.umass.edu		✓			✓		✓	✓
	Leximancer (Smith 2003) http://info.leximancer.com/	✓		✓	✓	✓	✓	✓	
	Stanford TMT (Ramage & Rosen 2009) http://nlp.stanford.edu/software/tmt/tmt-0.4/		✓	✓		✓			✓
	Google TMT https://code.google.com/p/topic-modelling-tool	✓		✓		✓			✓
R	Topicmodels (Hornik & Grün 2011)		✓						✓
	lda (Chang 2012)		✓			✓			✓
	stm (Roberts, Stewart & Tingley 2016) http://www.structuraltopicmodel.com		✓	✓	✓	✓		✓	✓
Python	Gensim (Řehůřek & Sojka 2010) https://radimrehurek.com/gensim/index.html		✓						✓
	Sklearn (Pedregosa *et al.* 2011) http://scikit-learn.org/stable/auto_examples/applications/topics_extraction_with_nmf_lda.html		✓						✓

*The 'mallet' R package provides an interface in R for the JAVA software.

matrices, concept counts and classified text segments. We were particularly interested in highlighting tools that were able to generate attractive and highly useful graphical visualizations of results. These graphic summaries showcase the power and utility of ACA by providing a straightforward understanding of a large body of literature in an attractive, yet efficient way. We found that only Leximancer, stm and Mallet were able to generate graphic summaries or other visualizations. These capabilities are integrated into Leximancer and stm, while they are provided for Mallet in the form of an available add-on. Nevertheless, the development of visualization tools for ACA is a rapidly advancing area. For instance, the R Shiny package 'LDAvis' (Sievert & Shirley 2014) features a web-based, interactive topic model visualization system capable of providing global and in-depth views of the literature, as well as describing concept interrelationships through a variety of features, including concept maps and barcharts. Although LDAvis does not fit the model on its own, it complements ACA R packages, such as topicmodels and lda, by generating shareable, interactive visual outputs for fitted topic models. This tool can also be used to provide visual outputs for ACA analyses conducted using Gensim or Mallet.

Finally, we compared the accessibility of these tools by distinguishing open source from commercial ACA software. All tools, excluding Leximancer, are open sourced and freely available, providing an opportunity to the ecological and evolutionary communities to contribute to their further development and advancement.

It should be noted that there are other tools for text analysis and text mining. These tools (e.g. NVivo and DeepDive), however, exhibit important differences from those intended for ACA. The first important difference is that these tools do not form or use concepts, as defined in this paper, but instead rely mostly on single words to analyse large volumes of text. An exception is the tool NVivo, which in its latest version (NVivo 11 Plus; QSR International Pty Ltd.; Melbourne, Australia), offers an experimental feature called 'automatic coding using existing coding patterns'. This feature compares each text segment to text segments previously classified by the user under categories called 'nodes'. If the word content in the text segment is similar to the content of a text segment already classified under a node, then the text segment is classified under that node. This feature is similar to ACA in that it uses groups of words, rather than individual words, to identify categories. However, as opposed to ACA tools which can be completely automatic, this experimental aspect of NVivo requires much more human intervention (the user must designate nodes and manually train NVivo for classification), potentially increasing the likelihood of human bias. The second important difference between ACA tools and other text-mining tools is that some of these tools are designed for other literature synthesis purposes. For instance, DeepDive is a powerful tool that also uses machine learning and statistical inference, but is designed for knowledge-base construction (i.e. the process of populating a knowledge base with information extracted from text) and not for literature classification (Niu *et al.* 2012).

Maximizing the utility of ACA

Automated content analysis possesses two key characteristics that contribute to its utility for the review and synthesis of big literature. First, unlike the manual methods of literature synthesis primarily used in ecology, ACA is able to process large amounts of literature much more quickly. Illustrating this capability is an early implementation of LDA which was able to process 1000 concepts from 100 000 documents in 40 h (Zeng, Liu & Cao 2012). Similarly, the python-based ACA tool, Gensim, has been used to process all articles on English Wikipedia (8 GB) as a demo of this tool's processing capabilities (Rehurek 2015). The tool was found capable of processing 16 000 documents per minute. Theoretically, there is no limit to the amount of text that can be analysed by ACA tools, other than the amount of literature available within the focus of the investigation (Smith & Humphreys 2006). The ability to rapidly process large amounts of text allows for the analysis, re-analysis and synthesis of much larger samples of the literature – if not an exhaustive analysis of the entire population.

The second property of ACA contributing to its utility pertains to unintentional human bias. Manual classification by humans is subject to multiple influences (e.g. fatigue, personal bias and perception), many of which classifiers are unaware of, and therefore, unable to report (Nisbett & Wilson 1977; Downe-Wamboldt 1992; Smith & Humphreys 2006). ACA is able to mitigate these influences, potentially limiting subjective human bias. Unlike manual analysis, which derives its categorizing schema from previous knowledge, domain expertise and personal experiences, ACA objectively develops its concept categories from the text data using strategies based on 'grounded theory' – the reciprocal informing and shaping of data collection, and data analysis through an emergent iterative process (Smith & Humphreys 2006; Charmaz 2011). Nevertheless, ACA is not completely void of subjectivity, as it requires human inputs (e.g. hand-seeding concepts). For this reason, it is important to clearly document these inputs, enabling methodological transparency. Furthermore, not unlike traditional reviews, ACA is not able to address publication bias or bias due to paper accessibility. It has been suggested that to minimize the problem of publication bias, both published and unpublished data should be included in reviews (Pullin & Stewart 2006). However, doing so would only benefit ACA if the unpublished data is textual.

Certain key considerations will help to maximize the utility of ACA. For instance, human interpretation of ACA results is still necessary to place the synthesized literature in an ecological context (Blei 2012b). This consideration is of particular importance in cases where a concept's definition is context-dependent, as in concepts that vary in definition, scope or connotation across fields or through time (e.g. the concept of 'scale' can be inferred differently in a geographical versus as statistical context). For this reason, the advantages and utility of ACA are likely to be maximized with increased domain expertise and understanding of the topic being reviewed. Moreover, the utility of a given ACA study to readers will be

maximized by authors providing clear definitions and context for the concepts that they are investigating and/or detecting.

As with any other analytical method, ACA's utility will be maximized with replicability, that is the degree to which the same results for a given body of text can be reproduced by different analysts (Krippendorff 2013). Although ACA tools go through rigorous testing of their replicability (Smith & Humphreys 2006), the nature of the tool's interface (i.e. GUI vs. code-based) will determine the level of care required by the user to achieve replicability. To explain, utilizing the same code for an analysis in a code-based ACA tool should provide the same results as long as settings and parameters (e.g. maximum number of concepts to detect, size of text segments, list of stopwords, concept seeds used, thesaurus settings, concept generality threshold) remain constant. In contrast, with GUI-based tools, despite being more user-friendly, key choices within an analytical algorithm may be less transparent. For this reason, it is important for the researcher to document all chosen GUI settings. Justification of all settings should also be documented to further ensure replicability regardless of whether or not the researcher relies on code-based or GUI-based ACA tools. Of particular importance is the size of the text segments used to perform the analysis, as this choice may affect the identification of both concepts and interrelationships among concepts (Smith & Humphreys 2006).

Given this utility, ACA has great potential to fill an important methodological gap in literature review in the fields of ecology and evolutionary biology. This technique provides the tools necessary for the synthesis of big literature as it efficiently and reliably processes large volumes of text in substantially shorter periods and with much less effort than is possible with manual methods of literature review. ACA is able to produce statistical descriptions of the literature by identifying and quantifying the frequency of major concepts and the relationships among concepts. Despite its own limitations, ACA is able to surpass most of the limitations inherent to the manual literature synthesis methods that currently dominate in ecology and evolutionary biology. Such capabilities allow for a broad range of applications and provide the foundations needed for conceptual synthesis of the rapidly growing body of ecological and evolutionary literature. ACA's potential contributions to literature synthesis, as well as its favourable reception and effective use in many other fields, highlight how ecology and evolution can benefit by adopting this new methodology. Analogous to the tools being developed to analyse big data, ACA can help researchers to address the grand challenges in the environmental sciences (e.g. climate change, biodiversity loss, land-use change) by helping to harness the wealth of information contained within big literature.

Acknowledgements

We thank the Associate Editor, Chris Grieves, Dr. Scott Chamberlain and two anonymous reviewers for constructive comments on an earlier version of the manuscript. This research was supported by the NSF Macrosystems Biology Program grant #1241932.

Data accessibility

Data used in the targeted systematic review are currently archived in the online abstract and citation data base, Scopus. The details of the methodology used to obtain these data from Scopus are detailed in Table S1. References for the data used in the exploratory review are listed in Appendix S1.

References

Alexa, M. & Zuell, C. (2000) Text analysis software: commonalities, differences and limitations: the results of a review. *Quality and Quantity*, **34**, 299–321.

Arnqvist, G. & Wooster, D. (1995) Meta-analysis: synthesizing research findings in ecology and evolution. *Trends in Ecology & Evolution*, **10**, 236–240.

Bengston, D.N. & Xu, Z. (2006) Changing National Forest Values: a content analysis. Research Paper NC-323. US Department of Agriculture, Forest Service, North Central Forest Experiment Station, St. Paul, Minnesota, USA.

Bhagwat, S., Kettle, C.J. & Koh, L.P. (2014) The history of deforestation and forest fragmentation: a global perspective. *Global Forest Fragmentation* (eds C.J. Kettle & L.P. Koh), pp. 5–19. CABI, Oxfordshire, UK.

Blaschke, T. (2010) Object based image analysis for remote sensing. *ISPRS Journal of Photogrammetry and Remote Sensing*, **65**, 2–16.

Blei, D. (2012a) Probabilistic topic models. *Communications of the ACM*, **55**, 77–84.

Blei, D. (2012b) Topic modeling and digital humanities. *Journal of Digital Humanities*, **2**, 8–11.

Blei, D.M., Ng, A.Y. & Jordan, M.I. (2003) Latent dirichlet allocation. *Journal of Machine Learning Research*, **3**, 993–1022.

Chang, J. (2012) lda: collapsed Gibbs sampling methods for topic models. R package version 1.3. 2.

Charmaz, K. (2011) Grounded theory methods in social justice research. *The SAGE Handbook of Qualitative Research* (eds N.K. Denzin & Y.S. Lincoln), pp. 359–380. SAGE Publications, Thousand Oaks, California, USA.

Cooper, H., Hedges, L.V. & Valentine, J.C. (2009) *The Handbook of Research Synthesis and Meta-Analysis*. Russell Sage Foundation, New York, New York, USA.

Cretchley, J., Rooney, D. & Gallois, C. (2010) Mapping a 40-year history with Leximancer: themes and concepts in the Journal of Cross-Cultural Psychology. *Journal of Cross-Cultural Psychology*, **41**, 318–328.

Downe-Wamboldt, B. (1992) Content analysis: method, applications, and issues. *Health Care for Women International*, **13**, 313–321.

Grech, M.R., Horberry, T. & Smith, A. (2002) Human error in maritime operations: analyses of accident reports using the Leximancer tool. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, pp. 1718–1721. SAGE Publications, Thousand Oaks, California, USA.

Hopkins, D.J. & King, G. (2010) A method of automated nonparametric content analysis for social science. *American Journal of Political Science*, **54**, 229–247.

Hornik, K. & Grün, B. (2011) Topicmodels: an R package for fitting topic models. *Journal of Statistical Software*, **40**, 1–30.

Koricheva, J., Gurevitch, J. & Mengersen, K. (2013) *Handbook of Meta-Analysis in Ecology and Evolution*. Princeton University Press, Princeton, New Jersey.

Krippendorff, K. (2013) *Content Analysis: An Introduction to Its Methodology*, 3rd edn. SAGE Publications, Thousand Oaks, California, USA.

Laney, D. (2001) 3D data management: controlling data volume, velocity and variety. *META Group Research Note*, **6**, 70.

McCallum, A.K. (2002) Mallet: a machine learning for language toolkit. Available at: http://mallet.cs.umass.edu.

Newman, D., Noh, Y., Talley, E., Karimi, S. & Baldwin, T. (2010) Evaluating topic models for digital libraries. *Proceedings of the 10th Annual Joint Conference on Digital Libraries*, pp. 215–224. ACM, Gold Coast, Australia.

Nisbett, R.E. & Wilson, T.D. (1977) Telling more than we can know: verbal reports on mental processes. *Psychological Review*, **84**, 231.

Niu, F., Zhang, C., Ré, C. & Shavlik, J.W. (2012) DeepDive: Web-scale Knowledge-base Construction using Statistical Learning and Inference. *VLDS*, **12**, 25–28.

Nunez-Mir, G., Iannone, B. III, Curtis, K. & Fei, S. (2015) Evaluating the evolution of forest restoration research in a changing world: a "big literature" review. *New Forests*, **46**, 1–14.

Papadimitriou, C.H., Tamaki, H., Raghavan, P. & Vempala, S. (1998) Latent semantic indexing: a probabilistic analysis. *Proceedings of the Seventeenth ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems*, pp. 159–168. Seattle, Washington.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., *et al.* (2011) Scikit-learn: machine learning in python. *Journal of Machine Learning Research*, **12**, 2825–2830.

Penn-Edwards, S. (2010) Computer aided phenomenography: the role of Leximancer computer software in phenomenographic investigation. *The Qualitative Report*, **15**, 252–267.

Pullin, A.S. & Stewart, G.B. (2006) Guidelines for systematic review in conservation and environmental management. *Conservation Biology*, **20**, 1647–1656.

Ramage, D. & Rosen, E. (2009) Stanford topic modeling toolbox (stmt). Available at: http://nlp.stanford.edu/software/tmt/tmt-0.2/.

Řehůřek, R. & Sojka, P. (2010) Software framework for topic modelling with large corpora. *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*. ELRA, Valletta, Malta.

Rehurek, R. (2015) Experiments on the English Wikipedia. *Gensim: Topic Modelling for Humans*. Available at: https://radimrehurek.com/gensim/wiki.html.

Roberts, C.W. (2000) A conceptual framework for quantitative text analysis. *Quality and Quantity*, **34**, 259–274.

Roberts, M.E., Stewart, B.M. & Tingley, D. (2016) Package 'stm'. Available at: http://structuraltopicmodel.com

Salton, G. (1989) *Automatic Text Processing: The Transformation, Analysis, and Retrieval of Information by Computer*. Addison-Wesley Longman Publishing Co., Inc., Boston, Massachusetts, USA.

Sievert, C. & Shirley, K.E. (2014) LDAvis: A method for visualizing and interpreting topics. *Proceedings of the Workshop on Interactive Language Learning, Visualization, and Interfaces*, pp. 63–70. Association for Computational Linguistics, Baltimore, Maryland, USA.

Smith, A.E. (2003) Automatic extraction of semantic networks from text using leximancer. *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology: Demonstrations – Volume 4*, pp. 23–24. Association for Computational Linguistics, Edmonton, Canada.

Smith, A.E. & Humphreys, M.S. (2006) Evaluation of unsupervised semantic mapping of natural language with Leximancer concept mapping. *Behavior Research Methods*, **38**, 262–279.

Stockwell, P., Colomb, R.M., Smith, A.E. & Wiles, J. (2009) Use of an automatic content analysis tool: a technique for seeing both local and global scope. *International Journal of Human–Computer Studies*, **67**, 424–436.

Travaglia, J.F., Debono, D., Spigelman, A.D. & Braithwaite, J. (2011) Clinical governance: a review of key concepts in the literature. *Clinical Governance: An International Journal*, **16**, 62–77.

Vilà, M. & Ibáñez, I. (2011) Plant invasions in the landscape. *Landscape Ecology*, **26**, 461–472.

Wallach, H.M., Murray, I., Salakhutdinov, R. & Mimno, D. (2009) Evaluation methods for topic models. *Proceedings of the 26 th International Conference on Machine Learning*, pp. 1105–1112. ACM, Montreal, Canada.

Walter, V. (2004) Object-based classification of remote sensing data for change detection. *ISPRS Journal of Photogrammetry and Remote Sensing*, **58**, 225–238.

Web of Science (2015) Thomson Reuters. Available at: https://webofknowledge.com/ (accessed on August 2015).

Yarowsky, D. (1995) Unsupervised word sense disambiguation rivaling supervised methods. *Proceedings of the 33rd Annual Meeting on Association for Computational Linguistics*, pp. 189–196. Association for Computational Linguistics, Cambridge, Massachusetts.

Zeng, J., Liu, Z.Q. & Cao, X.Q. (2012) A new approach to speeding up topic modeling. arXiv preprint arXiv:1204.0170.

Zhao, W., Zou, W. & Chen, J.J. (2014) Topic modeling for cluster analysis of large biological and medical datasets. *BMC Bioinformatics*, **15**, S11.

Received 26 January 2016; accepted 23 May 2016
Handling Editor: Richard Fitzjohn

Supporting Information

Additional Supporting Information may be found online in the supporting information tab for this article:

Table S1. The methodology used to obtain the 29 766 abstracts used in the targeted systematic ACA review presented in this article.

Table S2. Predominant concepts automatically identified by ACA classified into six pathways of influence of land-use on invasion.

Table S3. Land-use drivers/landscape attributes identified by Vilà & Ibáñez (2011) in a manual review of the literature classified into six pathways of influence of land-use on invasion.

Appendix S1. Supplementary references for the 51 articles used in the exploratory ACA review presented in this article.