Preprint of article for 12th International Conference on Information and Communication Technologies in Agriculture, Food & Environment (HAICTA 2026)

proceedings

MDPI

Proceeding Paper

From Scientific Literature to Decision Support: A Digital Workflow for Ecosystem Restoration Analysis †

Yannis Marketakis [1,2,*], Michalis Mountantonakis [1,2], Pavlos Fafalios [2,3], Valantis Zervos [2], Anastasia Analyti [1], Arne Langlet [4], Anton Ellenbroek [5], Kim J. Friedman [5,6], Marcel Mallow [5] and Yannis Tzitzikas [1,2]

[1] Institute of Computer Science, Foundation for Research and Technology-Hellas, Heraklion, Greece; marketak@ics.forth.gr (Y.M.); mountant@ics.forth.gr (Mi.M.); fafalios@ics.forth.gr (P.F.); analyti@ics.forth.gr (A.A.); tzitzik@ics.forth.gr (Y.T.)
[2] Computer Science Department, University of Crete, Heraklion, Greece; vzervos@csd.uoc.gr (V.Z.)
[3] School of Production Engineering & Management, Technical University of Crete, Chania, Greece
[4] Department of Political Science, University of Vienna, Vienna, Austria; arne.langlet@univie.ac.at (A.L.)
[5] Food and Agriculture Organization of the United Nations, Rome, Italy; anton.ellenbroek@fao.org (A.E.); Kim.Friedman@fao.org (K.F.); marcel.mallow@fao.org (Ma.M.)
[6] Oceans Institute, The University of Western Australia, Crawley, Australia
* Correspondence: marketak@ics.forth.gr (Y.M.)
† Presented at the 12th International Conference on Information and Communication Technologies in Agriculture, Food & Environment (HAICTA|2026), Kos Island, Greece, 30 September – 3 October, 2026.

Abstract

Effective ecosystem restoration depends on timely access to scientific evidence. However, the rapid growth of environmental literature makes large-scale synthesis increasingly impractical, limiting the ability of policymakers and researchers to identify trends, gaps, and biases. This paper presents a scalable digital workflow that transforms unstructured scientific text into structured knowledge and interactive analytics to support evidence-based decision-making in ecosystem restoration. The approach integrates automated text extraction, vocabulary-based entity recognition, and large language model (LLM)-based semantic classification to characterize publications across ecosystem realms, biomes, and habitat contexts. Applied to a large corpus of ecosystem restoration literature, the workflow demonstrates both feasibility and scalability. Results show that LLM-based classification achieves high accuracy (up to 96.9%) using abstracts alone, enabling efficient large-scale analysis with low computational cost. The resulting outputs support the generation of indicators and visualizations that reveal research focus, biases, and knowledge gaps. Overall, this study illustrates how ICT-driven knowledge extraction can bridge scientific literature and decision-support systems, supporting more informed restoration planning.

Keywords: literature analysis; ecosystem restoration; entity extraction; large language models; knowledge extraction

Academic Editor: Firstname Lastname

Published: date

Copyright: © 2026 by the authors. Submitted for possible open access publication under the terms and conditions of the Creative Commons Attribution (CC BY) license.

1. Introduction

The analysis of large bodies of scientific literature provides insight into research trajectories, emerging themes, and persistent gaps in knowledge. Bibliometric approaches have been widely applied to characterize publication patterns and research focus across scientific domains, yet the rapid growth of published research has made large-scale literature analysis increasingly arduous, time-consuming, and difficult to reproduce. These

limitations apply also to the context of ecosystem restoration, where activities span diverse ecosystems, biomes, and taxa. Understanding how research attention is distributed across these dimensions is essential for assessing current investments and identifying future priorities.

The FAO-UNEP UN Decade on Ecosystem Restoration[1], and the Kunming-Montreal Global Biodiversity Framework, notable Target 2 [1], call for large-scale restoration of degraded ecosystems, requiring systematic and transparent synthesis of scientific evidence. Effective restoration requires understanding current trends, thematic focus, and knowledge gaps in scientific research. One important source of knowledge for this purpose is scientific literature.

The objective of this paper is to present a digital workflow implemented to support a large-scale case study on ecosystem restoration literature. The emphasis is on identifying the distribution of publications across thematic dimensions, such as ecosystem realms or biomes. The workflow combines automated text extraction, vocabulary-based entity recognition, and large language model (LLM)-based semantic classification to characterize publications according to ecosystem realms, biomes, and habitat context by relying on corresponding taxonomic resources. The workflow produces structured representations that are used to derive aggregated indicators. These can be used to build interactive visualizations that reveal research focus, biases, and knowledge gaps. In this work, we focus on research papers and reports published primarily after the year 2000 from several publication databases, reflecting the period during which ecosystem restoration has gained increasing scientific and policy attention.

This paper contributes: (i) a scalable workflow for transforming ecosystem restoration literature into structured knowledge, (ii) a hybrid information extraction approach combining controlled vocabularies and LLM-based classification, (iii) a large-scale evaluation demonstrating high classification accuracy (up to 96.9%) using abstracts, and (iv) an interactive analytics framework supporting evidence-based decision-making.

2. Related Work

Bibliometric analysis has been widely applied to explore research trajectories, thematic evolution, and publication patterns across scientific domains [2]. These approaches are commonly divided into metadata-based and content-based analyses. While metadata-based methods provide useful high-level indicators, they often fail to capture the semantic depth of scientific publications.

Within biodiversity and ecosystem research, several studies have employed bibliometrics and content analysis techniques to investigate trends in ecosystem services, biodiversity management, and the use of community science data [3][4][5]. Large-scale bibliometric analyses of ecosystem restoration literature have examined publication growth, thematic evolution, and research hotspots using thousands of articles [6][7]. These studies consistently report a strong emphasis on terrestrial ecosystems, particularly forests, grasslands, and wetlands, with biodiversity, ecosystem services, and climate change as dominant themes. More specialized analyses have focused on specific ecosystems types and regions, such as salt marsh restoration [8], mangrove ecosystem services [9], marine and estuarine rehabilitation in Europe [10], and broader and thematic trends in restoration research [11].

Beyond bibliometric reviews, meta-analyses have synthesized empirical evidence on restoration effectiveness across habitats and interventions [12], but these efforts typically rely on manual or semi-manual synthesis and are hard to scale. Recent work has explored machine learning methods for automated analysis of marine data and literature [13]; however, such approaches rarely rely on standardized ecological vocabularies or typologies.

[1] https://www.decadeonrestoration.org/

In contrast, this work integrates controlled ecological vocabularies with LLM-based semantic classification in a unified, scalable workflow for ecosystem restoration literature analysis.

3. Digital Workflow and Implementation

The digital workflow implemented to support the large-scale study on ecosystem restoration literature is designed as a modular processing pipeline that transforms scientific documents into structured semantically rich representations. As illustrated in Figure 1, the workflow consists of four main phases: (a) knowledge base construction, where domain-specific vocabularies are assembled; (b) text extraction from scientific publications; (c) information extraction and entity identification using the knowledge base and large language models; and (d) results aggregation and interactive analytics. The following subsections describe each stage in detail.

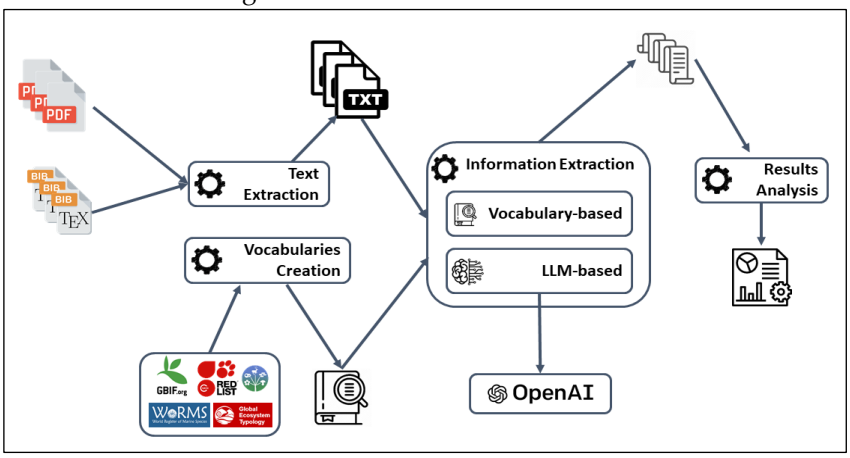


Figure 1. Overview of the digital workflow for large-scale content-based ecosystem restoration literature analysis

3.1. Knowledge Base Construction

A core component of the workflow is the construction of domain-specific knowledge resources used to support entity identification and semantic classification. Two primary vocabularies were assembled. First an ecosystem vocabulary aligned with the IUCN Global Ecosystem Typology[2], providing a hierarchical structure of ecosystem realms and biomes. This typology enables consistent classification of publications across terrestrial, freshwater, marine and transitional systems.

Second, a species vocabulary using authoritative taxonomic resources, including the Global Biodiversity Information Facility (GBIF)[3], the Catalogue of Life[4], the IUCN Red List of threatened species[5], the Knowledge Base of Aquatic Resources about fisheries [14], and others. Information from these sources was collected programmatically using their APIs. Scientific names were programmatically resolved against taxonomic APIs that provide structured habitat and environmental metadata (e.g., marine, freshwater, terrestrial). These metadata fields were extracted and aligned through rule-based mappings to the hierarchical ecosystem vocabulary. This alignment enabled the automated inference of ecosystem context when publications referred primarily to species rather than explicitly to ecosystem types.

[2] https://global-ecosystems.org/

[3] https://www.gbif.org/

[4] https://www.catalogueoflife.org/

[5] https://www.iucnredlist.org/

The resulting ecosystem vocabulary consists of approximately 200 hierarchically organized terms, while the species vocabulary comprises several hundred thousand terms. These sizes reflect the broad coverage and support large-scale fine-grained entity identification. The vocabulary is limited by the information available in the source datasets and, particularly for lesser studied species, may lead to erroneous classifications or underrepresentation. To mitigate this limitation, the workflow combines complementary taxonomic resources and uses the LLM-based mode to infer ecosystem context from publications, partially compensating for incomplete species metadata. Nevertheless, classification quality ultimately depends on the completeness and accuracy of the datasets.

3.2. Text Extraction

The text extraction component prepares the textual input used in subsequent analysis phases. For publications available as PDF documents, the workflow applies automated parsing[6] to extract text at different levels of granularity, including title, abstract, and full text. This flexibility allows analysis to be performed using either concise summaries or complete document contents, depending on the requirements of the task. In addition, the workflow supports structured bibliographic inputs in BibTeX format; for these records, metadata fields such as title and abstract are directly parsed and extracted without the need for further processing. The text extraction step is a preparatory step that homogenizes literature contents so that they can be subsequently analyzed.

3.3. Information Extraction and Entity Identification

The information extraction component supports two complementary modes. The first one is a vocabulary-based mode that relies solely on the constructed domain vocabularies. In this mode, the workflow performs literal string matching against the constructed knowledge base. Textual inputs are scanned for exact occurrences of ecosystem and species terms, with matched entities recorded alongside their frequencies. To ensure full transparency and reproducibility, this mode intentionally bypasses stemming, lemmatization, or morphological normalization. This ensures that every classification is explicitly grounded in the source text, although it inherently excludes implicitly expressed contexts, a limitation addressed by the complementary LLM-based mode.

The second mode combines vocabularies with LLMs [15]. Here the vocabularies define the target entities, while LLMs identify and classify relevant entities within the text. This enables the detection of both explicitly mentioned entities and implicitly expressed ecosystem contexts that may not correspond to direct term matches. Both modes can operate on inputs of different granularity, including titles, abstracts, and full text, allowing the workflow to balance accuracy, computational cost, and scalability. The LLM-based mode was implemented using OpenAI's GPT-5.1 model through the OpenAI API[7]. To improve reproducibility, a fixed prompt and temperature value of 0 were used throughout the experiments, promoting consistent classifications across runs.

3.4. Results Aggregation and Interactive Analytics

The outputs produced by the information extraction phase are consolidated and aggregated into structured summaries across ecosystem realms, biomes, species, and habitat contexts. The results are stored in machine-readable CSV and JSON files, enabling their further analysis and reuse. These datasets are also loaded into Looker Studio[8], where

[6] Using Apache PDFBox (https://pdfbox.apache.org/)

[7] https://openai.com/api/

[8] https://lookerstudio.google.com/

through interactive dashboards, users can filter, compare, and drill down into specific ecosystem categories and view publications associated with selected criteria. Figure 2 shows a screenshot of this interactive dashboard, which facilitates transparent inspection of results and supports the identification of patterns, biases, and knowledge gaps.

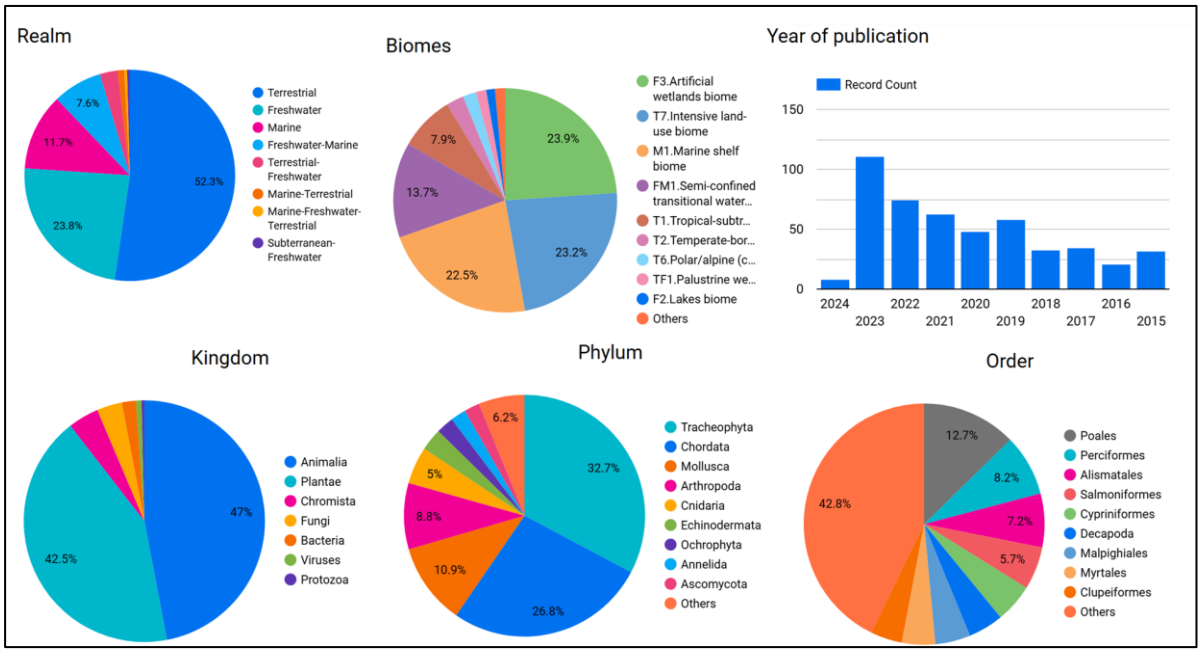


Figure 2. Interactive analytics (from Looker Studio) demonstrating the ecosystem terms (upper) and species taxonomy distributions (lower).

4. Results and Implications

The performance of the workflow was evaluated by comparing automated classification results against a manually annotated subset of publications produced by domain experts in ecosystem restoration and biodiversity science. Accuracy was computed at the document level by comparing predicted and expert-assigned ecosystem categories, and was assessed for both vocabulary-based and LLM-based modes using different contextual textual inputs, including titles, abstracts, and full text.

The evaluation, summarized in Table 1, was conducted over two complementary corpora: a full-text collection of 639 PDF documents, and an abstract-only collection of 8,776 BibTeX records. The results show that the LLM-based mode consistently outperforms the vocabulary-based approach. Notably, LLM-based classification achieves high accuracy (96.9%) even when relying solely on abstracts, due to its ability to infer implicit entities. This provides an effective balance between accuracy and computational efficiency for large-scale analysis. Moreover, this level of accuracy enables reliable large-scale screening without full-text processing, reducing computational cost and enabling near real-time updates. Conversely, the low accuracy of the vocabulary-based mode when using titles (5.6%) and abstracts (24.7%), highlights the limitations of literal matching and demonstrates the added value of LLMs for detecting implicitly expressed ecosystem contexts.

The workflow outputs reveal clear patterns in ecosystem restoration research. Publications targeting terrestrial ecosystems constitute the largest share of the literature, with freshwater and marine ecosystems receiving comparatively less attention. These patterns are consistent across vocabulary-based and LLM-based analysis, although the LLM-based mode identifies a slightly larger number of publications with terrestrial context. The LLM's sensitivity to implicit context effectively recovered terrestrial classifications where explicit keywords were absent.

Table 1. Document-level classification accuracy of the information extraction modes of the workflow using different modes

		mode	
		Vocabulary-based	**LLM-based**
input	**title**	5.6 %	86.7 %
	abstract	24.7 %	96.9 %
	full text	80.1 %	99.9 %

The results demonstrate how ICT-driven literature analysis can support evidence-based ecosystem restoration planning. By revealing the research focus as well as under-represented ecosystems, the workflow enables policymakers and funding bodies to identify potential gaps and rebalance research and restoration efforts. Through the interactive analytics, stakeholders can explore literature by ecosystem type and habitat context, supporting more informed prioritization of restoration actions. Beyond exploratory analysis, the workflow can be integrated into decision-support systems, enabling data-driven prioritization of restoration actions based on research gaps, ecosystem coverage, and thematic trends.

5. Conclusion

This paper presented and applied a digital workflow for large-scale analysis of ecosystem restoration literature, integrating automated text extraction, domain-specific vocabularies, and LLM-based classification. The results demonstrate that accurate and scalable characterization of publications can be achieved using abstracts only, enabling efficient processing of large corpora. The analysis reveals a strong dominance of terrestrial ecosystems in restoration research, with freshwater and marine ecosystems comparatively under-represented, highlighting important knowledge gaps. Overall, the proposed approach contributes to the digital transformation of environmental knowledge management and can be extended to related domains such as agri-food systems and biodiversity monitoring.

Future work will focus on extending vocabularies, constructing a knowledge graph from the extracted information in order to provide question answering services to support evidence-based ecosystem restoration planning, and applying the workflow to related datasets from FAO, such as Aquatic Sciences and Fisheries Abstracts (ASFA)[9].

Author Contributions: Conceptualization, Y.M., Mi.M., P.F., A.A., A.L., A.E., K.F. and Y.T.; methodology, Y.M., A.A., A.E., K.F. and Y.T.; software, Y.M., Mi.M., A.L. and V.Z.; validation, A.E., Ma.M. and K.F.; formal analysis, A.A., Ma.M., K.F. and Y.T.; investigation, Y.M., A.A., A.L., A.E., K.F., M.Ma. and Y.T.; resources, A.L., A.E. and Ma.M.; data curation, A.L., A.E. and Ma.M.; writing—original draft preparation, Y.M., Mi.M., P.F., A.A., A.L., A.E. and Y.T.; writing—review and editing, Y.M., A.A., A.E., K.F., Ma.M. and K.F.; visualization, Y.M.; supervision, K.F. and Y.T.; project administration, K.F. and Y.T.; funding acquisition, K.F. All authors have read and agreed to the published version of the manuscript.

Funding: This research was supported by the United Nations Decade on Ecosystem Restoration supported by the International Climate Initiative of Germany and the Government of Denmark, and the AIM4NatuRe Initiative under AIM4Forests Programme, supported by the United Kingdom of Great Britain and Northern Ireland. Funding was also provided by the FAO Regular Programme, and the Japanese Government ('Biodiversity mainstreaming for sustainable fisheries', part of GCP/GLO/173/JPN).

[9] https://www.fao.org/asfa/en

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data are available on reasonable request from the corresponding author.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

API	Application Programming Interface
GBIF	Global Biodiversity Information Facility
IUCN	International Union for Conservation of Nature
LLM	Large Language Model
PDF	Portable Document Format

References

1. Stephens, T., 2023. The Kunming–Montreal global biodiversity framework. International Legal Materials, 62(5), pp.868-887.
2. Donthu, N., Kumar, S., Mukherjee, D., Pandey, N. and Lim, W.M., 2021. How to conduct a bibliometric analysis: An overview and guidelines. Journal of business research, 133, pp.285-296.
3. Abdullah, K.H., Sofyan, D., Roslan, M.F., Shukri, N.N.M. and Hammood, W.A., 2022. Biodiversity management: a bibliometric analysis of Scopus database publications. Asian Journal of Fundamental and Applied Sciences, 3(3), pp.41-53.
4. Blanco-Zaitegi, G., Etxeberria, I.Á. and Moneva, J.M., 2022. Biodiversity accounting and reporting: A systematic literature review and bibliometric analysis. Journal of Cleaner Production, 371, p.133677.
5. Binley, A.D., Vincent, J.G., Rytwinski, T., Proctor, C.A., Urness, E.S., Davis, S.A., Soroye, P. and Bennett, J.R., 2023. Patterns of community science data use in peer-reviewed research on biodiversity. Biological Conservation, 280, p.109985.
6. Wei, X., Song, W., Shao, Y. and Cai, X., 2022. Progress of ecological restoration research based on bibliometric analysis. International Journal of Environmental Research and Public Health, 20(1), p.520.
7. Shen, Z., Tian, Y., Yao, Y., Jiang, W., Dong, J., Huang, X., Wu, X., Farooq, T.H. and Yan, W., 2023. Ecological restoration research progress and prospects: A bibliometric analysis. Ecological Indicators, 155, p.110968.
8. Gonçalves, C., Fernandes, J., Neto, J.M., Veríssimo, H., Caçador, I. and Verdelhos, T., 2024. Evolution and effectiveness of salt marsh restoration: a bibliometric analysis. Water, 16(8), p.1175.
9. Tasneem, S. and Ahsan, M.N., 2024. A bibliometric analysis on mangrove ecosystem services: Past trends and emerging interests. Ocean & Coastal Management, 256, p.107276.
10. Oliveira, V.H., Díez, S., Dolbeth, M. and Coelho, J.P., 2024. Restoration of degraded estuarine and marine ecosystems: a systematic review of rehabilitation methods in Europe. Journal of hazardous materials, 469, p.133863.
11. Bao, H., Nikolaeva, A., Xia, J. and Ma, F., 2025. Evolution Trends and Future Prospects in Artificial Marine Reef Research: A 28-Year Bibliometric Analysis. Sustainability, 17(1), p.184.
12. Danovaro, R., Aronson, J., Bianchelli, S., Boström, C., Chen, W., Cimino, R., Corinaldesi, C., Cortina-Segarra, J., D'Ambrosio, P., Gambi, C. and Garrabou, J., 2025. Assessing the success of marine ecosystem restoration using meta-analysis. Nature Communications, 16(1), p.3062.
13. Malde, K., Handegard, N.O., Eikvil, L. and Salberg, A.B., 2020. Machine intelligence and the data-driven future of marine science. ICES Journal of Marine Science, 77(4), pp.1274-1285.
14. Marketakis Y, Tzitzikas Y, Gentile A, Ellenbroek A, Taconet M. Building a Global Aquatic Resource Knowledge Base for Fisheries (HAICTA 2024). Proceedings. 2025; 117(1):4. https://doi.org/10.3390/proceedings2025117004
15. Naveed, H., Khan, A.U., Qiu, S., Saqib, M., Anwar, S., Usman, M., Akhtar, N., Barnes, N. and Mian, A., 2025. A comprehensive overview of large language models. ACM Transactions on Intelligent Systems and Technology, 16(5), pp.1-72.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.