Tracing Ecological Metaphors in Discourses on Open Science using LLMs and Knowledge Graphs

Nil Yagmur Ilba, Simon Dumas Primbault

Tracing Ecological Metaphors in Discourses on Open Science using LLMs and Knowledge Graphs

Nil Yagmur Ilba[1,2], and Simon Dumas Primbault[1,2]

[1] OpenEdition Lab, 22 rue John Maynard Keynes, 13013 Marseille, France
[2] Laboratory for the History of Science and Technology (LHST), Swiss Federal Institute of Technology (EPFL), CH-1015 Lausanne, Switzerland

Abstract

The term ”ecosystem” is frequently used to describe various concepts, not only in open science but also in broader discussions of research and innovation. Despite its widespread use, it is rarely explicitly defined, often functioning as a boundary object that facilitates communication across diverse communities. Systematically documenting its varied, context-dependent meanings presents a significant challenge. This work in progress explores the term ”ecosystem” within the discourse on open science, offering a systematic approach to mapping its varied meanings and uses. We pose a twofold research question: from a social scientific perspective, how can the diverse uses of ”ecosystem” be systematically documented? And from a methodological standpoint, how can computational techniques be leveraged to trace such a boundary object? Drawing on a curated corpus of 211 scholarly articles and exploratory ontological work, we use LLMs to construct a detailed knowledge graph, yielding 1,067 semantic relations. This graph is then integrated with a citation network to create a multilayer model for analyzing the term’s dissemination. Our preliminary results identify seven distinct, data-driven thematic communities. Although the application of knowledge graphs is now an emerging practice, our pipeline offers a novel application for revealing the term’s underlying meanings. By mapping its surrounding ontology, this ongoing work suggests how such a term allows knowledge to circulate between different scholarly communities, providing deeper insight into the conceptual landscape shaping the digital transition of research.

Keywords: Open Science, Knowledge Graphs, Large Language Models, Boundary Objects, Co-occurrence Networks, Prompt Engineering

1 Introduction

For some years now, a number of researchers have used the term ‘ecosystem’ to refer, in very different ways, to the open environment in which their practice occurs. However, the numerous uses of the semantic field of ecology to understand the digital transition of research environments are very diverse and very polarised. This study aims to analyze the contextual usage of “ecosystem” within a corpus of academic texts related to open access, open science, and open data at different levels.

To achieve this, we employ a multi-stage methodology that combines natural language processing (NLP) with advanced network analysis. We begin by constructing co-occurrence networks to identify the primary thematic clusters associated with “ecosystem.” Building on this, we use a Large Language Model (LLM) to extract entities and relations, forming a detailed knowledge graph that formalizes the term’s semantic roles. Finally, we integrate this semantic data with the

Nil Yagmur Ilba, and Simon Dumas Primbault. “Tracing Ecological Metaphors in Discourses on Open Science using LLMs and Knowledge Graphs.” In: *Computational Humanities Research 2025*, ed. by Taylor Arnold, Margherita Fantoli, and Ruben Ros. Vol. 3. Anthology of Computers and the Humanities. 2025, 1372–1389. https://doi.org/10.63744/ImYNsefqcE1n.

corpus's citation network to create a multilayer model. This approach allows us to offer novel insights into the evolving semantic landscape of "ecosystem," tracing its interdisciplinary adoption and conceptual crystallization over time.

2 Literature Review and Research Question

A central theoretical lens for this study is the concept of "boundary object", first introduced in the context of museum studies to describe artifacts that connect different stakeholder communities. According to Star and Griesemer [14], boundary objects are items or concepts that are robust enough to maintain a common identity across different social worlds but are flexible enough to be adapted to the local needs and work of the groups using them. They serve as points of translation and coordination, enabling collaboration and knowledge sharing even when participants have different perspectives or goals.

We propose that the term "ecosystem" itself functions as a boundary object within the discourse on open science. Its malleability allows diverse actors—including researchers, policymakers, publishers, and technologists—to discuss the complex scholarly landscape from their unique viewpoints. For example, a software developer might refer to a "software ecosystem" as a set of interdependent and modular software components linked through APIs, while a publisher might discuss the "book ecosystem" as the entire assemblage of actors and processes involved in publication, from editorial staff to distribution norms like DOIs.

This flexibility also reveals a deep polarization in its use: for some, the metaphor implies a world of competition for scarce resources and the 'survival of the fittest', while for others, it is a way to promote diversity, collaboration, and mutual assistance. Analyzing how the term is defined and used can therefore reveal these underlying dynamics and alignments within the field [2].

Ecological metaphors first emerged to describe research environments in the 1970s-1980s at the intersection of cybernetics and symbolic interactionism, and later in the 1990s-2000s within the sociology of digital infrastructures and business management. Since 2010, the concept has expanded in platform studies and political discourse, addressing issues like resilience, modular design, bibliodiversity, virality, and life cycles [11].

Large literature reviews of "ecologies" have recently been led in the fields of Human-Computer Interaction (HCI) and Computer-Supported Collaborative Work (CSCW). For example, two common traits were identified by Lyla et. al (2020): "(1) an interest in relationships" and "(2) an attempt to take a holistic perspective" (p. 4), and three scales of analysis: macro *i.e.* "organisations and activity system(s)", meso *i.e.* "people and practices", and micro *i.e.* "individuals, artifacts, and tasks" (p. 7) [7; 8].

More specifically, the term "ecosystem" has gained prominence in discussions about open science, emphasizing the need for not only open research outputs but also transparent research processes. This vision encourages more rigorous, collaborative approaches to research [15]. In this context, Marton (2021) traces the genealogy of digital ecosystems, revealing their deep connection to biological concepts since the 1990s, structured around Gregory Bateson's ecological epistemology [10]. Similarly, Jaime (2021) systematically examines OA Research, proposing that the Ecosystem consists of three components: Actors, Technology, and Regulatory Framework [5].

Traditional computational methods for textual analysis, such as topic modeling or frequency analysis, are valuable for tracing the usage of terms over time. However, these approaches can sometimes lack the necessary semantic depth, especially when terminologies are nuanced and difficult to differentiate through surface-level reading alone. When a term like "ecosystem" is used in subtly different but significant ways, a more powerful approach is needed to capture these distinctions accurately.

This study therefore poses a twofold research question centered on the use of the term "ecosystem" as a boundary object to understand "open science". From a social scientific perspective, we

ask: ***how can its diverse uses within this scholarly literature be systematically documented?*** From a methodological standpoint, we then ask: ***how can modern computational techniques be leveraged to trace its evolution and application?***

To achieve this deeper level of analysis, our work draws on state-of-the-art techniques that combine structured Knowledge Graphs (KGs) with the generative power of LLMs. This approach leverages their complementary strengths: while LLMs excel at parsing the complexities of unstructured text, their outputs can lack factual grounding and are prone to hallucination. KGs, conversely, provide a structured, verifiable, and explicit representation of information that can serve as a ”ground truth” to mitigate these weaknesses. The emerging methodology, therefore, often uses LLMs for the initial, large-scale extraction, with the output then structured into a KG to create a robust and auditable knowledge base [1]. The core inspiration for our pipeline is drawn from Graph RAG, a technique where an LLM is fed domain-specific knowledge from a KG to produce more accurate and contextually-aware outputs [3]. We adapt this principle by first using an LLM to build the KG of term usages and then using that structured knowledge to inform a classification of those same terms [6; 9].

While the integration of LLMs and knowledge graphs enables a more nuanced semantic analysis, it also raises fundamental questions about reproducibility and interpretation in computational humanities. Because LLMs are inherently stochastic, producing slightly different outputs under identical conditions, their use challenges the notion of reproducibility that characterizes traditional computational paradigms. In the humanities, however, interpretation is itself an integral part of inquiry. Rather than seeking to reproduce exactly identical results, the objective is to structure an exploratory process in which interpretation can be systematically examined and made explicit. In this sense, our approach seeks to make the interpretive dimension traceable through upstream structuring by means of designed prompts, taxonomies, and evaluation processes. The outputs produced are not conceived as final results but as instruments for thought, supporting reflection and hypothesis formation on how computational methods can assist rather than determine analysis.

3 Dataset

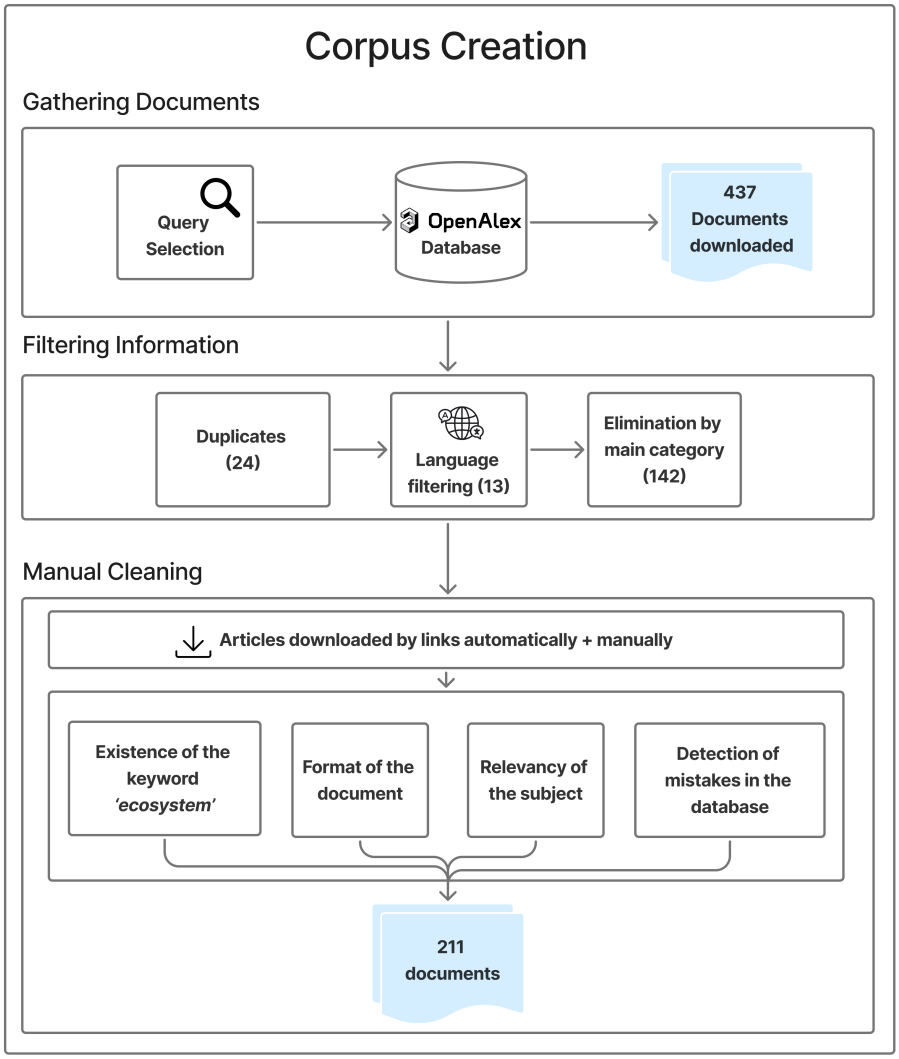

Figure 1: An overview of the data collection and cleaning pipeline.

The initial phase of this research involved the creation of a specialized corpus. The data collection and cleaning process, illustrated in Figure 1, was designed to systematically identify and refine a set of relevant scholarly articles.

3.1 Data Retrieval and Initial Filtering

The primary data source was the OpenAlex database [13]. Using a detailed query composed of several related phrases (Table 1), we retrieved an initial set of 437 documents, of which 346 were open access. This dataset was then refined through a multi-stage automated filtering process that removed 24 duplicates, 13 non-English documents, and 142 articles with irrelevant subject areas. To validate our retrieval method, the resulting corpus was also checked against a manually assembled collection of key literature to ensure it had captured known, relevant publications.

Query Component
`"open science ecosystem"` or `"open access ecosystem"` or `"ecosystem of open science"` `"ecosystem of open access"` or `"open science of the ecosystem"` or `"open access of the ecosystem"` or `"ecosystem in open science"` or `"ecosystem in open access"`

Table 1: Search Query Components Used for Corpus Retrieval.

3.2 Manual Cleaning

Following the automated filtering, the remaining articles underwent a rigorous manual cleaning phase to ensure the final corpus's quality and relevance. This crucial step balanced the efficiency of automation with the nuance required for semantic analysis. Each document was carefully read to assess the contextual and thematic relevance of the term 'ecosystem', verify its format as a research article, and correct for any database metadata errors. This meticulous process of assessment and categorization resulted in the final, curated corpus of 211 highly relevant documents.

3.3 Corpus Characteristics

The final corpus exhibits a diverse disciplinary and temporal distribution. As shown in Figure 2, the number of relevant publications has seen a significant increase since approximately 2018. The distribution spans multiple fields, with notable concentrations in Computer Science, Social Sciences, and Decision Sciences, highlighting the interdisciplinary nature of the topic.

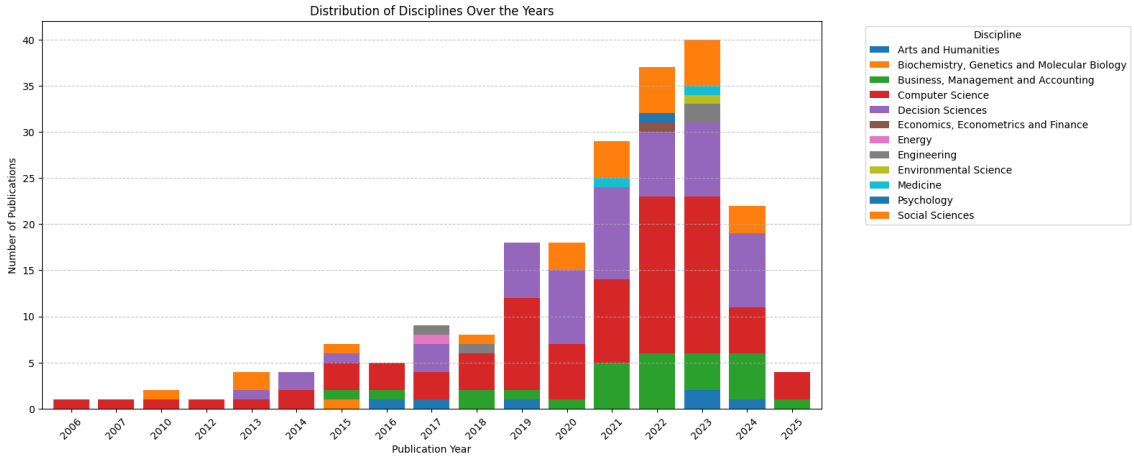

Figure 2: Distribution of publications in the final corpus by discipline and publication year.

4 Methodology

Our study employs a multi-stage methodology as visualized in the pipeline in Figure 3. The process consists of three core stages: (1) a large-scale extraction of entities and relations from the corpus using a Large Language Model; (2) an entity refinement phase involving filtering and semi-automated taxonomic labeling; and (3) a final network construction phase to build both a semantic knowledge graph and an integrated multilayer network from the refined data and external citation information. The following sections detail each of these stages.

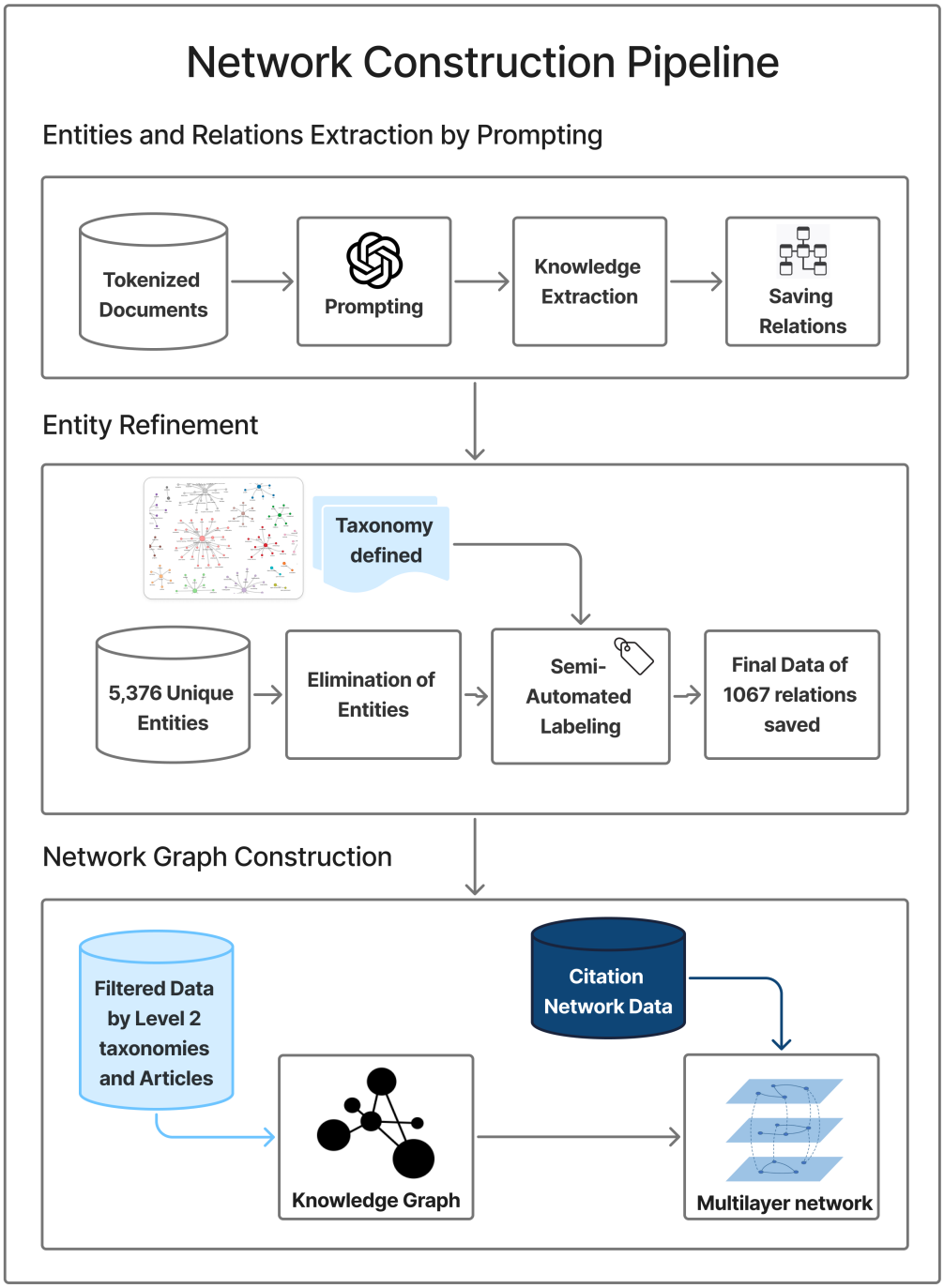

Figure 3: Pipeline of methodological process

4.1 Exploratory Work: Co-occurrence Networks

Our current methodology is informed by preliminary research that established a quantitative foundation for mapping the diverse applications of the term 'ecosystem'. This initial work aimed to create a data-driven overview of the term's semantic field before undertaking a more profound qualitative and ontological analysis. The central finding of this exploration was that 'ecosystem' is frequently used to denote the new dynamics of knowledge, data, and innovation emerging from the growth of digital networks in research, standing in contrast to an earlier, more mechanical un-

derstanding centered on stable infrastructures and large, vertically integrated organizations such as publishing houses.

The core of this preliminary study was a word co-occurrence network built from textual chunks extracted from 39 articles that explicitly used ‘ecosystem’ in their titles. Statistical analysis of the network revealed frequent co-occurring terms (e.g., “actors,” “data,” “innovation”) and significant bigrams (e.g., “innovation ecosystem”), highlighting the term’s strong association with both social and technical dimensions. The Louvain clustering algorithm was then applied to this network to group related terms and uncover its underlying thematic structures.

Cluster ID	Description
0	Technology and Research Tools
1	Data and Information
2	Ecosystem Research (Broad)
3	Publishing in Open Access (covers clusters 3 & 8)
4	Actors and Stakeholders
5	Structure/Dynamics of Ecosystems
6	Innovation and Business
7	Software Systems (covers clusters 7, 11, & 12)
8	Open Science (Broad)

Table 2: Clusters of co-occurrences identified from the preliminary analysis.

The thematic clusters identified in this preliminary analysis are summarized in Table 2. This initial quantitative grouping provided a foundational map of the discourse, which confirmed the term’s association with the interdependent nature of digital research environments and directly informed the detailed ontological framework developed in this study.

4.2 Entities and Relations Extraction by Prompting

Building on the preliminary analysis, we moved to a deeper extraction of semantic relations from the entire 211-article corpus to construct a knowledge graph. Using the OpenAI API (GPT-3.5 Turbo), each document was first tokenized and segmented into smaller, manageable chunks. This strategy preserved the local context around each mention of “ecosystem” while adhering to the model’s input limits, which is crucial for accurate relation extraction.

A systematic process of prompt engineering was then undertaken. We designed and tested several prompt variations to optimize for precision and recall, instructing the model to act as a domain expert. The final prompt mandated that the model identify entities and their relationships as (subject, relation, object) triples in a structured format, focusing specifically on text surrounding the term “ecosystem.” The full prompt is detailed in Appendix A.

The outputs from the API calls, corresponding to each processed text chunk, were programmatically parsed, validated, and aggregated. This robust process yielded a substantial knowledge base containing a total of **7,198 relations** across the 211 articles in the corpus. The resulting graph is extensive, comprising **3,509 unique subjects** and **5,376 unique objects**, which demonstrates a wide diversity of extracted concepts. On average, approximately **34 relations** were extracted per article. However, this distribution was skewed by a small number of articles with a high volume of relations; after an outlier analysis, a more representative mean of **25.7 relations** per article was calculated.

4.3 Entity Refinement and Taxonomy Construction

Following the extraction of raw entities and relations, the next step was to refine this knowledge base and structure it into a coherent taxonomy for analysis. To reduce noise and focus on the most salient concepts, we first filtered the entity list by removing terms that appeared in only a single article. We then crafted a hierarchical, two-level taxonomy to impose a clear semantic structure on the data. The first level consists of high-level domains (e.g., “Open Science”, “Research Processes / Practices”), while the second captures more specific sub-themes (e.g., “Sharing” and “Collaboration”). The creation of this taxonomy followed a hybrid methodology, beginning with a keyword-based pass to assign preliminary labels, which simplified subsequent steps. We then combined automated semantic clustering of the entity embeddings with manual curation based on expert domain knowledge. The overall structure was also informed by the thematic groups identified in our preliminary co-occurrence network analysis, ensuring consistency across the different stages of our research.

Main Category	**Number of Mentions**
Economy	89
Ecosystem	108
Events	44
Fields and Disciplines	84
Frameworks	201
Institutional Action	89
Open Access	348
Open Data	52
Open Government	8
Open Innovation	13
Open Science	461
Policies	192
Research Outputs / Resources	474
Research Processes / Practices	879
Research Values / Virtues	287
Science and Society	56
Sociotechnical Devices	595
Stakeholders/Actors	791
Grand Total	**4771**

Table 3: Frequency of Main Categories in the Developed Taxonomy.

The resulting ontology, visualized as a network of distinct categories and their constituent entities in Table 4, provides a structured framework for categorizing the key concepts present in the ‘ecosystem’ discourse.

Main Category	Sub-categories (in lowercase)
Economy	apc, bpc, business, cost, growth, ip, labor, market, innovation
Ecosystem	ecosystem
Events	activities, community building, conference, engagement, initiatives, training
Fields and Disciplines	archaeology, digital humanities, disciplines, engineering, humanities, metascience, ror, social sciences, studies, science
Frameworks	fair, framework, license, metric, norms, principle, project, rights, standards, knowledge commons, decentralized science
Institutional Action	evaluating, funding, monitoring, program, standardization, supporting
Open Access	open access
Open Data	open data
Open Government	open government
Open Innovation	open innovation
Open Science	open science
Policies	advocacy, agreements, governance, plan, policies, strategy
Research Outputs / Resources	article, book, citation, data, indigenous knowledge, journal, metadata, model, ontology, preprint, publications, research output, resource, software, findings
Research Processes / Practices	analyzing, assessment, citation, co-creation, collaboration, communication, control, cooperation, curation, data management, data practices, knowledge transfer, learning, licensing, machine learning, management, peer review, practice, preservation, publishing, research, research process, reuse, sharing, trials, visualizing, workflow
Research Values / Virtues	accessibility, bias, data security, diversity, efficiency, equity, ethics, fair, impact, integrity, interoperability, multilingualism, quality, reproducibility, responsibility, solidarity, transparency, trust, value
Science and Society	citizen science, participation, social justice, society, public goods, digital transformation
Sociotechnical Devices	ai, api, cloud, infrastructure, pid, platform, publications, repository, services, software, technology, tools
Stakeholders/Actors	actor, author, community, european commission, funders, incubators, indigenous people, information professionals, institutions, intermediaries, learned societies, librarians, libraries, organisations, people, policy makers, publishers, researcher, stakeholder, students, team, universities, university presses, users

Table 4: The Two-Level Taxonomy of Thematic Categories and their Constituent Sub-categories.

4.4 Knowledge Graph

We built a directed knowledge graph from the **1,067 valid relations** identified in our corpus on second-level manually labeled data with **146** unique entities. Each relation triple was enriched with its subject's and object's Level 1 and Level 2 category labels. The structure of this graph was then analyzed using the `NetworkX` library to identify thematic sub-structures [4]. We applied the Louvain community detection algorithm to the network, which grouped the entities into distinct communities based on their connectivity patterns.

For visualization, we used the `Pyvis` library to generate an interactive HTML file [12]. In this network, node size corresponds to its in-degree, and node color represents its main category, providing an immediate visual summary of an entity's role and classification. Edges are weighted based on the frequency of relations between two nodes. The HTML output includes a custom-built user interface with a dropdown menu, allowing for the interactive filtering of the graph by the algorithmically detected communities.

4.5 Multilayer Network of Citations and Semantics

Our analysis incorporates a large-scale citation network from a prior study, visualized in Figure 4. This network was constructed starting from our primary corpus of 211 articles and was then extended by including their more than 5,600 referenced publications. That research used this network to map the scholarly communities discussing the "open science ecosystem," finding that the "ecosystem" metaphor emerged independently across several distinct communities of practice, such as those focused on innovation, publishing platforms, and data management standards.

Building on this foundation, our current work integrates that structural citation data with our new semantic knowledge graph.

Figure 4: The extended citation network of over 5,600 publications, originating from the 211-article primary corpus. Colors indicate distinct scholarly communities.

5 Results

Our analysis of the constructed knowledge graph and the integrated multilayer network reveals key structural patterns in the discourse surrounding ‘ecosystem’ in the open science literature. These visualizations allow us to move from a static classification of terms to a dynamic map of their interactions.

5.1 The Semantic Knowledge Graph

The complete knowledge graph of extracted concepts, shown in Figure 5[1], illustrates the dense web of relationships centered around the core concepts of open science. The visualization highlights a core-periphery structure, where central terms like ‘ecosystem‘, ‘open science‘, ‘policies‘, and ‘researchers‘ are highly interconnected. In contrast, more specific or technical concepts are situated on the periphery, often connecting to only one or two central themes. This density demonstrates the complexity and the high degree of conceptual overlap within the corpus.

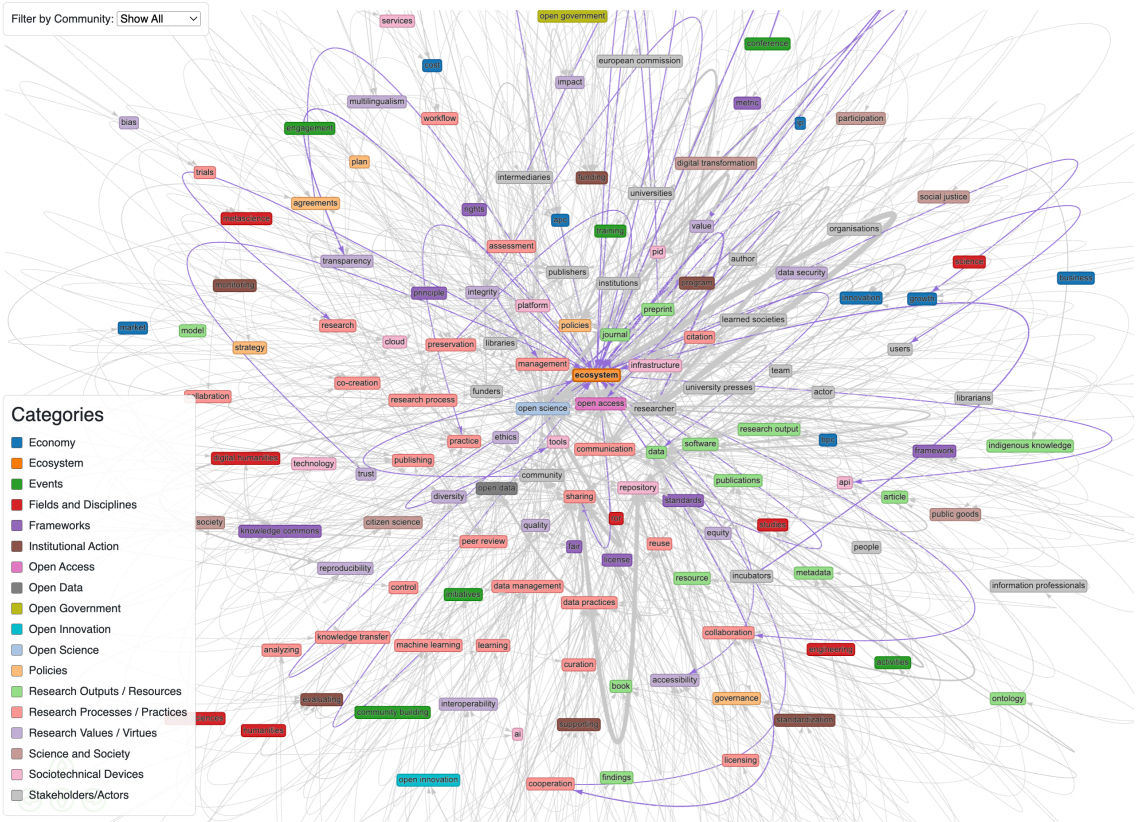

Figure 5: The full knowledge graph of extracted entities. The dense central cluster reveals the strong interconnection between core concepts, with node colors indicating their classification according to the taxonomy.

Particularly noteworthy is `ecosystem` node. While not having the highest degree, its position within the dense core and its high betweenness centrality score (Table 5) underscore its role as a crucial conceptual bridge. It frequently connects disparate topics such as technology, policy, and research practices, acting as a key mediator in the network.

[1] An interactive version of this graph is available at: `https://knowledge-graph-ecosytem-os.tiiny.site/`

The knowledge graph consists of **146 unique entities** (nodes) and **706 relations** (edges), derived from 1,067 valid semantic relations. To identify the most influential concepts, we analyzed the in-degree, out-degree, and betweenness centrality of the nodes (Table 5).

As expected given the corpus, `open science` and `open access` dominate the centrality rankings, acting as both major subjects and objects of discussion. A notable distinction appears in the roles of related concepts: while the node `sharing` exhibits a high in-degree (making it a frequently referenced outcome), `researcher` shows a high out-degree, positioning it as a primary agent. This suggests a common narrative within the literature that frames researchers as the principal actors who perform the action of sharing. This dynamic is further specified by the high centrality of `data` and `repository`, indicating that the discourse places a strong emphasis on the specific practice of sharing data via repository-based infrastructures.

In-Degree		**Out-Degree**		**Betweenness Centrality**	
Node	**Score**	**Node**	**Score**	**Node**	**Score**
open access	0.2621	open science	0.4138	open science	0.1429
sharing	0.2069	open access	0.2000	open access	0.1203
open science	0.1724	researcher	0.2000	sharing	0.1050
data	0.1724	repository	0.1862	data	0.0941
repository	0.1448	data	0.1724	ecosystem	0.0671

Table 5: Top 5 Nodes by Centrality Measures

To uncover the thematic substructures within the graph, the Louvain algorithm detected seven distinct communities. The most central nodes within each community reveal specialized discourses, as detailed in table 6.

Community 1 (31 members)		**Community 2 (26 members)**		**Community 3 (22 members)**	
Node	**Local Centrality**	**Node**	**Local Centrality**	**Node**	**Local Centrality**
research output	0.1857	open access	0.1922	data	0.2740
infrastructure	0.1816	peer review	0.1139	standards	0.2131
sharing	0.1791	integrity	0.0467	data practices	0.1347
repository	0.1600	communication	0.0406	quality	0.0903
tools	0.0601	apc	0.0400	curation	0.0889

Table 6: Top Nodes for Communities 1, 2, and 3.

Community 4 (21)		**Community 5 (18)**		**Community 6 (15)**		**Community 7 (13)**	
Node	**Centrality**	**Node**	**Centrality**	**Node**	**Centrality**	**Node**	**Centrality**
policies	0.0360	ecosystem	0.1765	funding	0.0769	community	0.0833
open science	0.0263	machine learning	0.0625	universities	0.0440	practice	0.0530
institutions	0.0105	digital transformation	0.0478	multilingualism	0.0055	metascience	0.0303
governance	0.0079	technology	0.0221	stakeholder	0.0000	bias	0.0303
trust	0.0075	business	0.0000	impact	0.0000	trials	0.0152

Table 7: Top Nodes for Communities 4, 5, 6, and 7.

Community 1 is centered on the material outputs and infrastructure of research, with top nodes like `research output`, `infrastructure`, `sharing`, and `repository`. Similarly, Community 3 focuses specifically on the practices surrounding research `data`, such as `standards`, `data practices`, and `curation`.

Community 4 highlights the intersection of policy and governance. Here, `open science` is discussed in the context of `policies`, `institutions`, and `governance`, demonstrating a cluster focused on the high-level, structural implementation of open principles.

Notably, **Community 5** is where the `ecosystem` node itself is most central. As shown in Figure 6, it is strongly associated with concepts of `machine learning`, `digital transformation`, `technology`, and `business`. This suggests a distinct discourse where ”ecosystem” is used as a techno-economic framework for innovation.

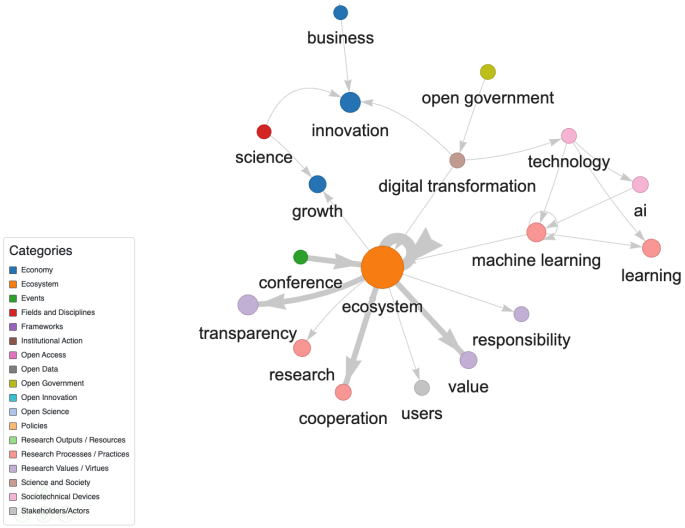

Figure 6: A focused view of Community 5, where the ‘ecosystem‘ node has the highest community centrality.

Finally, Community 6 relates to the social and financial aspects of research, led by terms like `funding` and `stakeholders`, while Community 7 groups together terms related to research `community` and `practice` on a meta-scientific level.

5.2 Multilayer Network of Concepts and Citations

To understand how these semantic themes are embedded within the scholarly conversation, we constructed a multilayer network, depicted in Figure 7. This network integrates two distinct types of relationships: semantic mentions and citation links. In this model, the colored nodes represent the main taxonomic categories, while the grey nodes represent individual articles.

A key insight from this integrated view comes from analyzing the bridging nodes—those with high betweenness centrality that connect different parts of the graph (Table 8). The most influential bridges in the network are not individual articles, but the thematic concepts themselves. The fact that core concepts from the semantic knowledge graph, such as `open science`, `open access`, and `sharing`, remain the most central nodes even after integrating the citation layer underscores their foundational role.

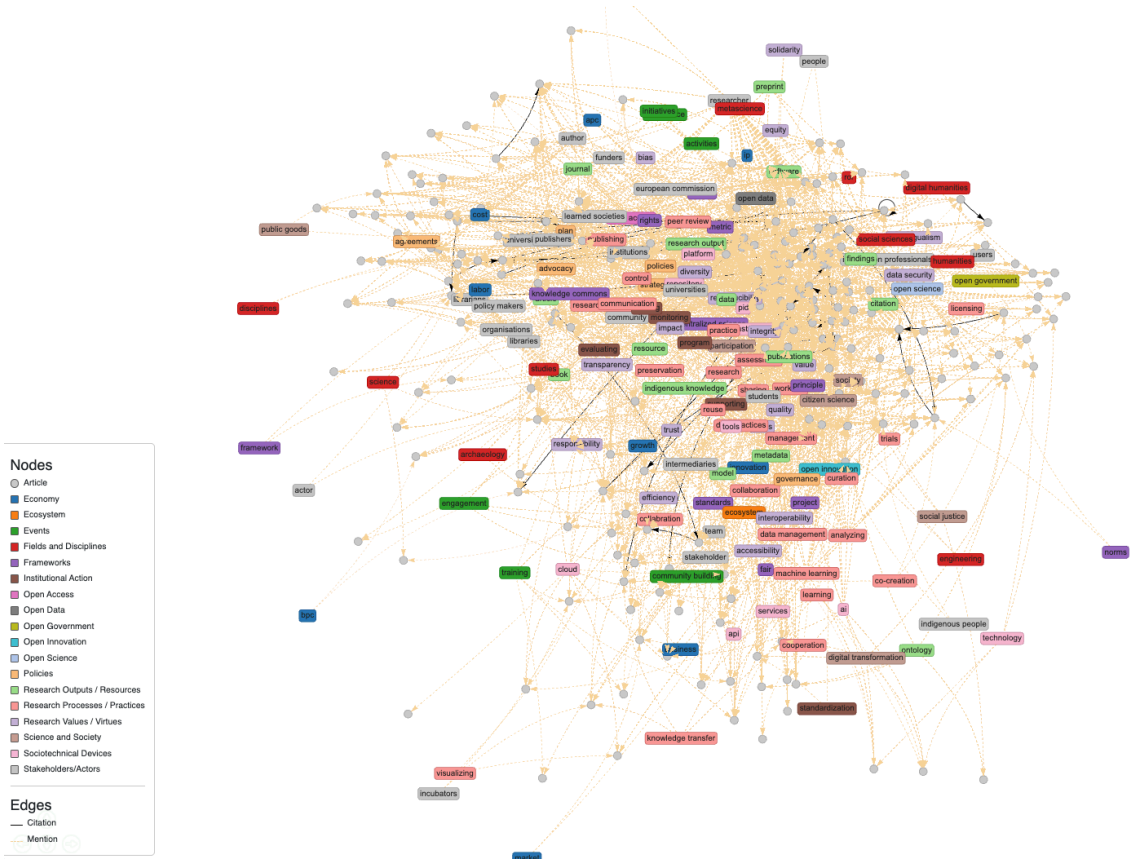

Figure 7: The multilayer network integrating semantic and citation data.

Rank	Node	Betweenness Score
1	open science	0.0674
2	open access	0.0561
3	sharing	0.0473
4	data	0.0290
5	repository	0.0250
6	ecosystem	0.0241
7	researcher	0.0224
8	community	0.0188
9	infrastructure	0.0176
10	practice	0.0144

Table 8: Top 10 Most Influential Nodes by Betweenness Centrality in the Multilayer Network

6 Discussion

Our analysis, which combines knowledge graph construction for semantic analysis with network science, supported by structured qualitative work on the ontology of the term ‘ecosystem,’ reveals that its usage in open science literature is not singular, but rather diverse and context-dependent. We observed that these different use cases are well-distributed across the articles in our corpus, indicating that this conceptual diversity is a widespread feature of the discourse. Our preliminary results show a core-periphery structure in the underlying knowledge graph, where general terms like `open science` provide overall coherence, while more specific concepts define the periphery.

These findings, particularly the role of the `ecosystem` node as a primary 'bridge' in the network, strongly support our central thesis that the term functions as a **boundary object**. The distinct character of the seven identified communities provides clear evidence for this: the 'policy' community (Community 4) and the 'techno-economic' community (Community 5), for example, use the same term to frame very different priorities—one of governance, the other of innovation. In this way, the term acts as a flexible tool for disseminating knowledge and coordinating actions between different actors and fields, even when their interpretations vary.

Methodologically, this study provides a novel workflow for analyzing boundary objects through computational methods. Our work in progress shows the power of combining LLMs with network science to analyze complex conceptual evolution. While traditional methods can trace term frequency and map semantics based on a term's primary meaning, our approach maps the *relational structure* of concepts and provides a pipeline for understanding boundary objects through computational analysis.

The workflow of using an LLM for large-scale knowledge extraction, followed by community detection and a multilayer network analysis, helps to capture both the underlying semantic meaning of terms, offering a significant advance in depth over simpler textual analyses.

This ongoing study directly addresses our twofold research question. From the social scientific perspective, we systematically document the diverse uses of 'ecosystem' as a boundary object by mapping its semantic roles into data-driven thematic communities. From the methodological standpoint, we preliminary results show how modern computational techniques can trace such an object; our pipeline leverages LLMs to reveal the deep relational structure of a concept's usage, a significant advance over traditional textual analysis.

This study is, however, subject to certain limitations. Our corpus, while systematically compiled, is limited to a specific set of keywords and the OpenAlex database and may not capture all relevant literature. Furthermore, the LLM-based extraction, despite our validation and cleaning steps, is subject to the model's inherent biases and potential for misinterpretation, which we sought to mitigate through our semi-automated taxonomy creation.

7 Future Work

Further research will extend this study in two main directions. First, we are extending the analysis of the multilayer network. This involves a deeper investigation into the intersection of the semantic communities and the citation network to precisely map how different conceptualizations are cited and propagated by distinct scholarly communities.

Second, while the current analysis provides a static snapshot of the discourse, the created knowledge graph allows for the development of dynamic tools to engage with it. The knowledge graph and its detected communities can serve as a foundation for more advanced, domain-specific applications using a Graph RAG approach. For instance, we plan to create a system that can provide automated, query-focused summaries for each thematic community. This framework can be further extended into a sophisticated question-answering system. Such a tool would allow researchers not only to query the conceptual content of the existing corpus but also to analyze new articles, effectively allowing our static map to interpret emerging discourse.

References

[1] Abolhasani, Mohammad Sadeq and Pan, Rong. “Leveraging LLM for Automated Ontology Extraction and Knowledge Graph Generation”. In: *arXiv preprint arXiv:2412.00608* (2024). DOI: 10.48550/arXiv.2412.00608.

[2] Dumas Primbault, Simon. “OpenEdition as a *governed milieu*: Towards an ecological understanding of open digital knowledge infrastructures”. In: *Politics of Open Infrastructures: Exploring open digital knowledge infrastructures and socio-political dynamics*, ed. by Astrid Mayer Katja ; Mager. Open Books Publishers, 2026. Forthcoming.

[3] Edge, Darren, Trinh, Ha, Cheng, Newman, Bradley, Joshua, Chao, Alex, Mody, Apurva, Truitt, Steven, Metropolitansky, Dasha, Ness, Robert Osazuwa, and Larson, Jonathan. “From local to global: A graph rag approach to query-focused summarization”. In: *arXiv preprint arXiv:2404.16130* (2024). DOI: 10.48550/arXiv.2404.16130.

[4] Hagberg, Aric, Swart, Pieter J, and Schult, Daniel A. “Exploring network structure, dynamics, and function using NetworkX”. Tech. rep. Los Alamos National Laboratory (LANL), Los Alamos, NM (United States), 2008. DOI: 10.25080/TCWV9851.

[5] Jaime, Astrid, Osorio-Sanabria, Mariutsi Alexandra, Alcantara-Concepcion, Tamara, and Barreto, Piedad Lucia. “Mapping the open access ecosystem”. In: *The Journal of Academic Librarianship 47*, no. 5 (2021), p. 102436. DOI: 10.1016/j.acalib.2021.102436.

[6] Kokash, Natallia, Romanello, Matteo, Suyver, Ernest, and Colavizza, Giovanni. “From books to knowledge graphs”. In: *Journal of Data Mining & Digital Humanities 2023* (2023). DOI: 10.46298/jdmdh.9380.

[7] Kuehn, Evan F. “The information ecosystem concept in information literacy: A theoretical approach and definition”. In: *Journal of the Association for Information Science and Technology 74*, no. 4 (2023), pp. 434–443. DOI: 10.1002/asi.24733.

[8] Lyle, Peter, Korsgaard, Henrik, and Bødker, Susanne. “What’s in an Ecology? A Review of Artifact, Communicative, Device and Information Ecologies”. In: *Proceedings of the 11th Nordic Conference on Human-Computer Interaction: Shaping Experiences, Shaping Society*. NordiCHI ’20. Tallinn, Estonia: Association for Computing Machinery, 2020. DOI: 10.1145/3419249.3420185.

[9] Manghi, Paolo. “Challenges in building scholarly knowledge graphs for research assessment in open science”. In: *Quantitative Science Studies 5*, no. 4 (2024), pp. 991–1021. DOI: 10.1162/qss_a_00322.

[10] Márton, Attila. “Steps toward a digital ecology: ecological principles for the study of digital ecosystems”. In: *Journal of Information Technology 37*, no. 3 (2022), pp. 250–265. DOI: 10.1177/02683962211043222.

[11] Mounier, Pierre and Dumas Primbault, Simon. “Sustaining Knowledge and Governing its Infrastructure in the Digital Age: An Integrated View”. Oct. 2023. DOI: 10.5281/zenodo.10036402.

[12] Perrone, Giancarlo, Unpingco, Jose, and Lu, Haw-minn. “Network visualizations with Pyvis and VisJS”. In: *arXiv preprint arXiv:2006.04951* (2020). DOI: 10.48550/arXiv.2006.04951.

[13] Priem, Jason, Piwowar, Heather, and Orr, Richard. “OpenAlex: A fully-open index of scholarly works, authors, venues, institutions, and concepts”. 2022. DOI: 10.48550/arXiv.2205.01833. arXiv: 2205.01833 [cs.DL].

[14] Star, Susan Leigh and Griesemer, James R. “Institutional ecology,translations’ and boundary objects: Amateurs and professionals in Berkeley’s Museum of Vertebrate Zoology, 1907-39”. In: *Social studies of science 19*, no. 3 (1989), pp. 387–420. DOI: 10.1177/030631289019003001.

[15] Thibault, Robert T, Amaral, Olavo B, Argolo, Felipe, Bandrowski, Anita E, Davidson, Alexandra R, and Drude, Natascha I. “Open Science 2.0: Towards a truly collaborative research ecosystem”. In: *PLoS Biology 21*, no. 10 (2023), e3002362. DOI: 10.1371/journal.pbio.3002362.

A LLM Prompt for Knowledge Extraction

The following is the full prompt provided to the GPT-3.5 Turbo model for the extraction of entities and relations from the text chunks.

```
You are an expert in knowledge extraction.
Your task:
- Focus ONLY on text where the word "ecosystem" appears.
- From those parts, extract meaningful, well-formed entities (as nodes)
and their relationships (as triples).

Guidelines:
- Include any terms that directly contain the word "ecosystem"
(e.g., "Open Science Ecosystem", "Blockchain Ecosystem") as part
of the concept list.
- Only include entities that are concrete or conceptual - no vague
or generic entries.
- Limit to a maximum of 20 concepts and maximum of 20 relations.
- It is perfectly fine if there are fewer than 20 concepts or
relations, depending on the richness of the information.
- Do NOT invent or hallucinate content that is not explicitly or
implicitly supported by the text.
- Prioritize clarity, importance, and relevance.
- Avoid very long phrases (keep concept labels under 6-7 words).
- For relations, avoid generic verbs like "is", "are", or
conjugated forms like "includes". Use meaningful, specific verbs
(e.g., "participate in", "enable", "comprise").

Output Format (strictly):
Return a valid JSON object with exactly these two keys:
- "concepts": a list of important concept strings
- "relations": a list of [subject, relation, object] triples

Example:
{
"concepts": ["Open Science Ecosystem", "Repositories",
"Researchers", "Knowledge Sharing"],
"relations": [
["Researchers", "participate in", "Open Science Ecosystem"],
["Repositories", "enable", "Knowledge Sharing"]
]
}

Strict instructions:
- Only return the JSON object - no explanations, commentary, or
extra text.
- Do not repeat terms unnecessarily.
- Ensure that the JSON is valid (no missing commas, brackets, or
quotation marks).
```