Term-Centric Hierarchy Induction from Heterogeneous Corpora

Anonymous ACL submission

Abstract

Organizing knowledge from diverse text sources into interpretable hierarchies is crucial for tasks such as policy analysis, innovation monitoring, and exploratory domain mapping. Existing taxonomy induction methods typically rely on document-level representations that capture entire documents rather than the specific domain concepts relevant for knowledge organization, limiting their ability to generalize across heterogeneous sources. We propose a term-centric framework for inducing hierarchical taxonomies from heterogeneous corpora that scales to massive document collections. Our approach maps documents from diverse sources into a shared representation space using automatic term extraction, enabling robust cross-source alignment. Based on these representations, we construct interpretable hierarchies that integrate domain priors with data-driven clustering. Experiments on a novel English and German multi-source benchmark of over one million documents demonstrate that our method improves cross-source coherence and hierarchy quality over text- and summary-based baselines. A case study on German regional innovation analysis further demonstrates its practical utility for technology landscape mapping.

1 Introduction

Hierarchical representations of document collections provide a powerful tool for organizing large text corpora into interpretable, multi-level structures. Such hierarchies facilitate exploratory search, domain mapping, and trend analysis by enabling users to navigate from broad thematic areas to fine-grained topics. Recent work has demonstrated the potential of clustering- and LLM-based approaches for inducing taxonomies from scientific literature (Zhu et al., 2025; Katz et al., 2024; Oarga et al., 2024; Gao et al., 2025). In practice, however, meaningful analysis of of complex domains often requires integrating heterogeneous sources that reflect different contexts. Tasks such as policy foresight, regional innovation analysis, and domain-specific knowledge discovery require analysts to synthesize signals from multiple types of evidence. For example, identifying emerging technologies or monitoring strategic priorities often requires jointly analysing multiple data sources (Polchar, 2024; Hakiman and Stull-Lane, 2022).

Heterogeneous data sources pose two main challenges for hierarchy induction. First, they differ in style and structure. For example, scientific papers emphasize methods and findings, patents focus on technical claims, and funding records describe strategic objectives. When standard document-level embeddings are used, these stylistic conventions may cause clusters to reflect source boundaries rather than underlying thematic structure. Second, purely data-driven clustering reflects the empirical distribution of the corpus rather than the true structure of the domain. Since methods such as K-Means implicitly allocate more centroids to high-density or high-variance regions (Manning et al., 2008), sampling and coverage biases lead to finer partitions where documents are abundant while under-represented areas may be merged or fragmented (Ester et al., 1996; McInnes et al., 2017).

To address these challenges, we propose **TERMNET**, a term-centric framework for inducing hierarchical taxonomies from heterogeneous corpora that scales to massive document collections (see Figure 1). Unlike prior taxonomy induction methods that rely on raw–document or summary embeddings, our approach maps documents from diverse sources into a shared semantic space using automatic term extraction, reducing source-specific stylistic variation and improving cross-source alignment. Based on these representations, we construct hierarchies through a clustering process that integrates domain priors with data-driven signals, producing human-interpretable and

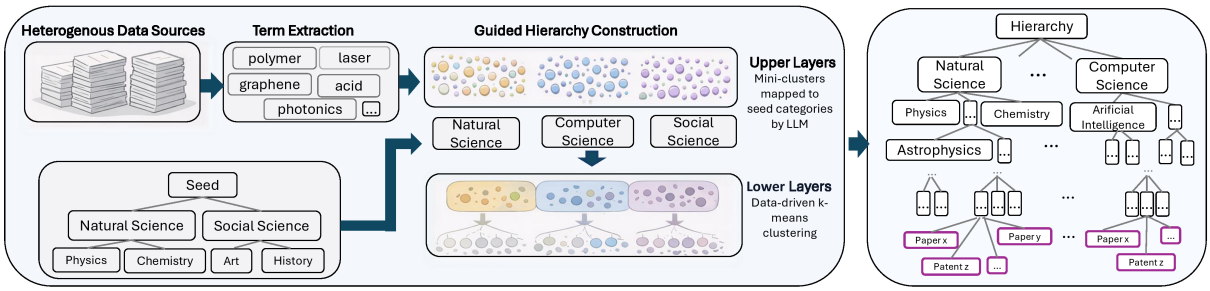

Figure 1: Overview of the TERMNET framework. Documents from heterogeneous sources are first mapped to a shared representation via automatic term extraction. The resulting embeddings are organized using a seed-guided hierarchical clustering procedure: predefined seed categories (representing broad scientific and technological domains) initialize the top levels of the hierarchy, which are then expanded in a data-driven manner.

domain-balanced taxonomy structures. We evaluate our method on our newly introduced large-scale multi-source benchmark containing over one million English and German documents. Experimental results show that TERMNET outperforms baselines based on raw text and summaries in clustering quality, cross-source integration, and human interpretability, as validated by automatic and human evaluation. A policy-oriented case study further demonstrates the practical utility of the induced hierarchies for technology monitoring and strategic analysis. Our main contributions are:

- We introduce **TERMNET**, a scalable term-centric framework for hierarchy induction from heterogeneous corpora.
- We propose an evaluation protocol for multi-source hierarchy induction, including source entropy and intruder detection, and perform large-scale automatic and human evaluations.
- We release a multi-source benchmark of over one million documents spanning publications, patents, and funding records to support research on heterogeneous knowledge organization.

2 Related Work

Research on hierarchy induction has traditionally focused on pattern-based hypernym extraction (e.g. Hearst, 1992; Shwartz et al., 2016; Panchenko et al., 2016) and distributional or clustering-based methods that organize semantically related concepts or documents into hierarchical structures (e.g. Wang et al., 2013; Liu et al., 2012; Mimno et al., 2007). More recently, LLMs have been applied either as standalone approaches or integrated into traditional pipelines (e.g. Gao et al., 2025; Zhu et al., 2025; Katz et al., 2024).

2.1 LLM-Enhanced Hierarchy Induction

We are not aware of prior work that induces unified hierarchies across heterogeneous data sources. Our approach builds on a broad literature in LLM-based hierarchy induction, which are typically applied to homogeneous, smaller-scale corpora.

The work most closely related to ours is SCYCHIC (Gao et al., 2025), which organizes collections of scientific literature (10k) into multi-level hierarchies. Their approach combines embedding-based K-Means clustering with selective LLM-based summarization to induce better interpretable hierarchies over scientific abstracts. A key insight of their work is that decomposing papers into specific contribution types, such as problem statements, yields more coherent structures than treating each paper as a single-topic entity. Oarga et al. (2024) leverage open-source LLMs for zero-shot ontology and knowledge graph generation from scientific literature by prompting models to extract vocabulary, infer hierarchical category structures, and extract relations in an end-to-end manner, demonstrating effectiveness in domain-specific settings such as chemistry. Other notable approaches include Zhu et al. (2025), who encode papers along multiple semantic aspects (e.g., methodology, data, evaluation) and cluster each summarized aspect with a probabilistic embedding-based model, followed by a dynamic search to ensure consistent cluster assignments when building a taxonomy. Katz et al. (2024) introduce Knowledge Navigator, an LLM-guided framework that organizes scientific query results (a few thousand papers) into two-level hierarchies. Their system first embeds and clusters retrieved papers using Gaussian Mixture Models,

then applies LLMs to name, describe, filter, and group clusters for exploratory browsing.

All of these methods operate on single-source scientific corpora and rely on document-level representations, either raw text, summaries, or aspect-based reformulations. However, summaries often preserve stylistic differences across sources and may introduce hallucinated content, particularly for short documents such as project titles. Therefore, we propose to use automatic terms extraction for the task of hierarchy induction.

2.2 Seed-Guided Hierarchy Induction

Another line of work studies hierarchy construction guided by seed structures, where a small initial hierarchy is expanded using information extracted from a corpus (Shen et al., 2025). Early approaches rely on embedding-based methods that recursively organize concepts or attach new ones to existing nodes (e.g. Zhang et al., 2018; Lee et al., 2022; Huang et al., 2020). More recent work leverages LLMs for seed-guided hierarchy construction. For instance, TAXOINSTRUCT (Shen et al., 2025) uses instruction-tuned LLMs to generate sibling entities and infer parent relations, and other approaches iteratively extend seed hierarchies via prompting strategies (Gao et al., 2025). Our work is related in that we also use seed hierarchies to guide taxonomy construction. However, our goal is cross-source alignment, we combine seed-guided category assignment for the upper layers with data-driven clustering for the lower layers. This allows us to approximate the conceptual structure of the domain rather than the empirical distribution of individual datasets.

3 Dataset Creation

To evaluate our approach, we construct a multi-source benchmark combining scientific publications, patents, and public research funding records. These sources capture complementary stages of the innovation pipeline: scientific knowledge production, technological protection, and publicly funded research activity. The dataset focuses primarily on German-affiliated organizations and contains documents in both German and English. The corpus integrates four major data sources. It contains 578,335 publication abstracts from OpenAlex (*openalex*) (Priem et al., 2022), 353,043 patent abstracts from the USPTO corpus (*uspto*) (Li et al., 2018), 12,979 project descriptions from EU framework programs Horizon 2020 (Publications Office of the European Union, 2015) and Horizon Europe (*horizon*) (Publications Office of the European Union, 2022), and 100,655 project descriptions from FöKAT (*foekat*), which catalogs research projects funded by the German Federal Government (Bundesministerium für Forschung, Technologie und Raumfahrt (BMFTR), 2026). Altogether, the resulting dataset comprises 1,044,977 documents.

The sources differ substantially in their linguistic characteristics and document structure. Publication and patent abstracts typically contain well-structured descriptions of research contributions and technological inventions, whereas funding records are shorter and often contain administrative or program-specific terminology. Combining these heterogeneous sources provides a challenging benchmark for methods aiming to capture technological and scientific topics across institutional contexts. Table 1 shows representative example documents and summary statistics for each source. Details on data retrieval, filtering criteria and licensing are provided in Appendix A.

4 Methods

4.1 Problem Formulation

Let $\mathcal{C} = \{d_1, \ldots, d_N\}$ denote a heterogeneous corpus consisting of documents from multiple sources. Each document may differ in structure, style, and length depending on its source.

Our goal is to induce a hierarchical taxonomy $\mathcal{H} = (\mathcal{V}, \mathcal{E})$ over the documents in $\mathcal{C}$, where $\mathcal{V}$ represents taxonomy nodes and $\mathcal{E}$ the parent–child relations between them. The hierarchy organizes documents into progressively more specific clusters, such that each document is associated with a unique path from the root to a leaf node. The resulting hierarchy should satisfy two objectives: (i) semantically coherent clusters, in which sibling nodes are thematically distinct and documents within a node share a common topic; and (ii) balanced domain coverage, such that the hierarchy reflects the breadth of the technological and scientific landscape rather than the frequency distribution of the corpus.

We further assume access to domain priors $\mathcal{H}_p$ that define coarse-grained categories in the upper layers of $\mathcal{H}$. These priors may guide the hierarchy construction but do not fully determine its structure.

Origin	raw text	Summary	Terms	Example
openalex	197.12	71.62	50.34	Marine Anthropogenic Litter; This book describes how ...
uspto	107.80	62.57	21.38	a canister vacuum cleaner comprising a main body ...
horizon	274.56	80.54	67.01	Redwave: the future of blood pressure measurement ...
foekat	16.30	59.77	6.20	Steigerung der Energieeffizienz beim Hochofenbetrieb ...
Average	177.82	67.54	36.51	

Table 1: Mean text, summary, and parsed response lengths and example sentences per origin.

4.2 TERMNET

We propose TERMNET a term-centric framework for inducing scalable and interpretable hierarchies from heterogeneous corpora, illustrated in Figure 1. The key ideas are i) to abstract away from source-specific linguistic conventions and document structures by representing documents through their contained technological terms; and ii) to ensure interpretability with a guided hierarchy construction procedure that balances domain priors with data-driven discovery.

Term-centric Representation Documents from different sources may follow distinct stylistic conventions, which can cause document representations to reflect source identity rather than thematic structure. We address this by representing documents through their domain-specific terms—words or phrases denoting defined concepts within a specialized domain. We use DiSTER, a fine-tuned model for cross-domain term extraction (Senger et al., 2025), to identify concepts, methods, materials, and technologies in each document. The extracted terms are concatenated and embedded to form the document representation. This maps heterogeneous texts into a shared representation space based on domain-specific terms. This encourages documents referring to similar concepts to align while reducing source-specific linguistic variation.

Hierarchy Guidance We construct the hierarchy using a hybrid strategy that combines domain priors with recursive clustering. The upper levels are initialized using predefined seed categories representing broad scientific domains $\mathcal{H}_p$. First, all documents are assigned to the root node v_0 (Alg. 1, line 2). We then apply K-Means over term-based document embeddings to obtain fine-grained clusters, where the number of clusters is determined as a multiple α of the number of seed categories under node v ($k = \alpha|\mathcal{C}_v|$). Each cluster is then assigned to one of the existing children of v or to a newly created child using zero-shot LLM classification (Alg. 1, lines 5–14). For this decision, the LLM receives representative keywords and documents retrieved using class-TF–IDF (Grootendorst, 2022). The model may select an existing category or propose a new subcategory, enabling the discovery of domains not covered by the seed hierarchy. New nodes are created only if the cluster exceeds a minimum relative size $s_{\min}$ and the proposed label is sufficiently dissimilar from existing category names (similarity threshold τ). This procedure is applied recursively for each level of the guidance hierarchy $\mathcal{H}_p$.

After reaching the seed depth P, the hierarchy is expanded in a fully data-driven manner using recursive top-down K-Means clustering (Alg. 1, lines 15–18). For each node, the number of children is chosen proportional to its document count while remaining within predefined bounds (B_m, B_M). This allows dense nodes to be partitioned more finely while maintaining comparable granularity across hierarchy levels. Overall, the hybrid strategy preserves interpretability at higher levels while enabling fine-grained organization at lower levels. The full TERMNET algorithm is shown in Algorithm 1.

4.3 Baselines

To evaluate the effectiveness of TERMNET, we compare against hierarchical clustering and recent taxonomy induction approaches. For completeness, we provide respective Algorithms in the Appendix B.

Recursive K-Means We implement recursive K-Means as a hierarchical clustering approach. Starting from a global partition, clusters are recursively subdivided in a breadth-first manner until the target granularity is reached. To ensure scalability, we apply Mini-Batch K-Means (Sculley, 2010) for clusters exceeding 10,000 documents and standard K-Means (MacQueen, 1967) otherwise. This approach produces a strictly nested hierarchy in which each document is assigned to a path defined

Algorithm 1: TERMNET algorithm

Input: Documents $\{d_1, \ldots, d_n\}$, hierarchy guidance $\mathcal{H}_p$ with depth P, total layers L, cluster multiplier α, size threshold $s_{\min}$, similarity threshold τ, branching bounds B_m, B_M

Initialization: Extract domain-specific terms for each d_i and compute embedding r_i
Add all documents d_i to the root node v_0
for $l = 0$ **to** $P - 1$ **do**
  for *each node* v *in* $\mathcal{V}_l$ **do**
    $k \leftarrow \alpha \cdot |\mathcal{C}_v|$
    Create k fine-grained clusters $\mathcal{M}$ using K-Means
    for *Fine-grained cluster* $m \in \mathcal{M}$ **do**
      Retrieve representative keywords and documents
      Zero-shot classify cluster m into children of v
      if *LLM proposes new category* ***and*** $|m| \geq s_{\min}|D_v|$ ***and*** *similarity* $< \tau$ **then**
        Create new child node d
        Assign documents of m to d
      else
        Assign documents of m to selected child of v
for $l = P$ **to** $L - 1$ **do**
  for *each node* v *in* $\mathcal{V}_l$ **do**
    Cluster document representations of v into $k \in \{B_m, B_m + 1 \cdots, B_M\}$ clusters
    Add every cluster as child of v
return *Hierarchical structure*

by successive centroid refinements. We apply this baseline to three document representations: raw text, document summaries, and extracted domain-specific terms.

SCYCHIC Due to compute feasibility, we adapt the SCYCHIC framework (Gao et al., 2025) to our large-scale corpus by using smaller embedding and summarization models (see Section 5.1). To address token constraints, in the bottom-up phase, we summarize only a representative subset of documents per cluster selected using Maximal Marginal Relevance (MMR), balancing similarity to the cluster centroid with semantic diversity.

5 Experimental Design

5.1 Setup

All methods use Qwen3-Embedding-0.6B (Zhang et al., 2025) to generate document embeddings. For document summarization, we employ Meta-Llama-3-8B-Instruct (Grattafiori et al., 2024), and for autmatic term extraction we use DiSTER with instruction-tuned Meta-Llama-3-8B-Instruct (Grattafiori et al., 2024) following Senger et al. (2025).

The hierarchy guidance $\mathcal{H}_p$ is fixed with two category layers curated by domain experts and inspired by the ASJC classification (Elsevier, 2024). Fine-grained clustering uses a multiplier $\alpha = 50$ that determines the number of clusters per node. The creation of new categories during guidance is controlled by a minimum relative cluster size $s_{\min} = 1\%$ and a semantic similarity threshold $\tau = 0.12$ to prevent redundant labels. For the data-driven hierarchy expansion phase, the branching factor is bounded by $B_m = 3$ and $B_M = 6$. The final hierarchy depth is fixed to $L = 4$, resulting in hierarchies with approximately $k = \{6, 40, 180, 680\}$ clusters from top to bottom. This configuration was selected in consultation with domain experts to support the downstream case study and analyses (Section 7). However, the method itself is not restricted to a fixed hierarchy depth or number of clusters.

5.2 Quantitative Analysis

5.3 Evaluation

We evaluate the induced hierarchies using a combination of unsupervised, supervised, and human-centered metrics to assess clustering quality, cross-source integration, structure and usability.

Unsupervised Metrics We measure lexical cluster distinctiveness at each hierarchy level using topic *diversity*. Topic diversity is measured using Inverted Rank-Biased Overlap (iRBO) (Webber et al., 2010; Terragni et al., 2021) computed over the top 10 representative terms of each cluster. To assess cross-source integration, we compute the *source entropy* for each cluster based on the distribution of document origins. Higher entropy indicates a more balanced source mixture, while lower entropy reflects source dominance. Since perfect balance is unattainable and some domains are inherently source-skewed, we comparatively evaluate reductions in source-driven fragmentation using entropy. Preliminary experiments showed that clustering raw documents produced source-separated clusters with low entropy; therefore, higher entropy indicates improved integration. The composition of clusters from different sources is further validated through human evaluation.

Supervised Metrics To evaluate high-level thematic alignment, we manually construct a gold standard for the top hierarchy layer using existing classification systems from the original data

Method	Precision	Diversity	Entropy	Intruder	Structure	Validity	Composition
K-Means (raw)	53.28	**99.79**	24.50	84.30	2.0	2.5	2.5
K-Means (summary)	47.56	99.41	23.98	91.25	3.5	2.0	4.0
K-Means (terms)	61.12	99.75	30.00	86.42	3.0	1.5	4.0
SCYCHIC	60.52	99.73	23.48	79.02	4.0	3.0	4.0
TERMNET wo. terms	66.39	99.57	33.11	69.94	4.0	2.5	5.0
TERMNET	**67.70**	99.22	**36.70**	**94.44**	**4.5**	**4.5**	**5.0**

Table 2: Automatic metrics and human evaluation scores across the different hierarchy induction approaches.

Cluster	Example Keywords	*openalex*	*uspto*	*horizon*	*foekat*	Method
Social Sciences And Humanities	[stakeholders, populist, mental]	148.668	4.309	1.806	3.993	K-Means (raw)
Materials Science And Advanced Technologies	[nanostructures, superconducting]	130.126	53.368	1.935	3.248	K-Means (raw)
Engineering And Applied Technologies	[reconfigurable, exoskeletons]	16.636	275.257	78	205	K-Means (raw)
Biomedical And Molecular Sciences	[fibrosis, protein, myelin]	140.317	11.017	1.810	3.526	K-Means (raw)
Life Sciences	[protein,biomass, genome]	103.866	5.595	1.558	2.586	K-Means (raw)
Information And Communication Technologies	[5g, innovativ, autonome]	38.722	3.462	5.792	87.097	K-Means (raw)
Social Sciences And Humanities	[kant, acculturation, unemployment]	55.764	192	671	6.140	TERMNET
Natural Sciences	[polymerization, phonon, chiral]	168.263	31.326	2.165	6.874	TERMNET
Engineering	[epitaxial, actuator, doped]	53.877	168.741	1.724	22.927	TERMNET
Health Sciences	[microglia, ecmo, hfpef]	139.286	16.937	1.696	7.793	TERMNET
Life Sciences	[isoforms, vesicles, rnas]	117.821	41.472	3.153	24.360	TERMNET
IT and Computing	[virtualized, subcarriers, ldpc]	43.324	94.340	3.570	32.561	TERMNET

Table 3: Top-level clusters and document counts per data source for the raw text K-Means baseline and TERMNET. Cluster names for the baseline were generated using an LLM. Keywords represent truncated examples of characteristic terms for each cluster. The distribution illustrates how TERMNET produces more balanced cross-source clusters compared to raw text clustering.

sources (e.g., OpenAlex fields and IPC codes for patents). For each experimental setting, these external categories are mapped to the induced top-level clusters. Documents are considered correctly assigned if their original categories correspond to the mapped cluster. For example, publications labeled with *General Immunology and Microbiology* and *Health Informatics* are expected to appear in the corresponding *Life Sciences* or *Health Sciences* clusters. Based on this mapping, we compute precision across all data sources, measuring the consistency between induced clusters and established domain classifications.

To evaluate the semantic coherence of the generated hierarchy, we design an intruder detection task inspired by Bhatia et al. (2018). Given a parent cluster and its child clusters, along with one semantically similar intruder cluster drawn from a different parent at the same hierarchy depth, the task is to identify the cluster that does not belong. To validate this evaluation procedure, we conducted a manual annotation study. We randomly sampled 60 parent clusters from the hierarchy, balanced across levels. For each instance, two human annotators independently identified the intruder cluster. Inter-annotator agreement reached Cohen's $\kappa = 0.73$, indicating substantial agreement. We further compared intruder predictions from an LLM (Llama-3.3-70B-Instruct (Grattafiori et al., 2024)) with the human annotations on the same subset. The mean Cohen's κ between the LLM and the annotators was 0.76, and the LLM achieved higher mean precision in detecting intruders than the individual annotators. We therefore use the LLM for large-scale evaluation and report macro accuracy across hierarchy layers, weighting each layer equally.

Human Evaluation Following Hu et al. (2025) and Zhu et al. (2025), we conduct expert-based evaluations of the generated hierarchies along three dimensions adapted to cross-data source hierarchy induction: *Structure*, *Validity* and *Composition*. *Structure* assesses whether the hierarchy exhibits a clear progression from broad research areas to specific sub-technologies with consistent granularity across levels. *Validity* evaluates whether the taxonomy aligns with experts' understanding of how concepts are typically grouped and related in the scientific and technological landscape. *Composition* examines whether clusters reflect a plausible mix of data sources given the nature of the technology (e.g., publication- vs. patent- or project-intensive domains). Each dimension is rated on a five-point Likert scale (see Appendix C for details). The mean inter-annotator agreement measured by Cohen's κ is $\kappa = 0.43$.

6 Results & Analysis

6.1 Quantitative Analysis

Table 2 summarizes automatic metrics and human evaluation scores. Overall, TERMNET achieves the strongest performance across most metrics, obtaining the highest precision (67.70), intruder accuracy (94.44), entropy (36.70), and human evaluation scores. Topic diversity remains very high for all methods (≈ 99), indicating that clusters are lexically distinct and largely non-overlapping regardless of the clustering strategy.

Clustering Quality Among the baselines, K-Means using term representations achieves the highest precision (61.12), outperforming clustering on raw text (53.28) and summaries (47.56). This suggests that term-centric representations reduce source-specific linguistic variation and better capture technological concepts. SCYCHIC reaches a comparable precision of 60.52. Seed-guided hierarchy construction further improves alignment. TERMNET without term extraction already increases precision to 66.39 and achieves the highest cross-source entropy among the baselines (33.11). Incorporating domain-specific terms yields the strongest results overall, with TERMNET reaching a precision of 67.70 and the highest intruder detection accuracy (94.44). The large improvement in intruder accuracy compared to baselines (e.g., 84.30 for K-Means raw and 79.02 for SCYCHIC) indicates substantially higher semantic coherence of the generated hierarchy.

Structure and Validity We assess whether hierarchies exhibit a logical progression from broad domains to specific technologies and consistent granularity (*Structure*) and whether cluster groupings match expert expectations (*Validity*). As shown in Table 2, TERMNET achieves the highest scores on both metrics, with Structure = 4.5 and Validity = 4.5. In comparison, the strongest baselines reach Structure scores of 4.0 (SCYCHIC and K-Means summary) and Validity scores of at most 3.0. Even without term extraction, TERMNET obtains strong structural ratings (Structure = 4.0), suggesting that guided hierarchy construction alone already align with the human expected structure and organization of domains.

Cross-Source Integration Domain-specific term representations substantially increase source entropy, indicating improved integration of publications, patents, and funding records within clusters. K-Means with terms increases entropy to 30.00 compared to 24.50 for raw text. TERMNET further improves integration, achieving the highest entropy overall (36.70). Human *Composition* scores support this observation, with TERMNET receiving the highest rating (5.0), indicating that clusters contain plausible mixtures of publications, patents, and projects rather than reflecting source-specific artifacts.

6.2 Qualitative Analysis

Analysis of intruder detection highlights subtle limitations and provides insights into hierarchy behaviour. For example, TERMNET without terms shows a low intruder score, especially at broader hierarchy levels, because some clusters include semantically distant subclusters. In one case, ‘Social and Community Development‘ was placed under ‘Life Sciences‘, which caused the model to select it instead of the true intruder, ‘Mathematics‘. Whereas the complete TERMNET approach with term extraction has no social science related subclusters in ’Life Sciences’ and the semantically most similar subcluster ’Automotive Engineering’ was identified correctly as the intruder. Such errors illustrate that while the seed-guided hierarchy ensures broad structural coherence, precision at fine-grained levels benefits from term-centric representations.

We hypothesize that the high *Composition* scores in human evaluation are largely driven by term extraction, while the elevated *Structure* and *Validity* scores are strongly influenced by hierarchy guidance. The upper layers of the hierarchy are initialized using the domain seed provided by experts, resulting the upper layers which roughly resemble established scientific hierarchies. This familiar structure helps experts navigate the taxonomy effectively. In contrast, purely data-driven approaches can deviate from established structures. For example, K-Means on summaries produces highly detailed clusters in engineering and technology at the top level but fails to allocate a clusters for social sciences and humanities. At deeper layers, subclusters like ‘Policy and Governance Studies‘ appear under ‘Life Sciences‘ or ‘Entrepreneurial Research and Innovation Transfer‘ under ‘Energy and Environmental Engineering‘, reflecting less coherent assignments in the absence of seed guidance.

Table 3 further illustrates differences in cluster composition. The raw text K-Means baseline

Method	Precision	Diversity	Entropy	Intruder %	Structure	Validity	Composition	Usefullness
K-Means (raw)	33.70	99.81	22.91	**98.01**	–	–	–	–
K-Means (summary)	37.71	99.13	18.07	95.63	–	–	–	–
K-Means (terms)	39.96	99.78	28.02	86.34	3.5	4	2	3.5
SCYCHIC	51.71	99.77	20.71	90.24	1	2	**3**	2
TERMNET	**64.77**	**99.88**	**35.63**	92.54	**4.5**	**4.5**	**3**	**4**

Table 4: Automatic metrics and human evaluation scores for the proprietary dataset.

Method	Entropy	Accuracy %
K-Means raw text	13.28	80
K-Means summary	11.75	78
K-Means terms	21.49	**86**

Table 5: Impact of document representations on source entropy and human-perceived accuracy.

produces clusters that are strongly dominated by individual data sources. For example, the cluster 'Engineering and Applied Technologies' contains 275k USPTO documents and 205k FÖKAT records, while other sources are sparsely represented. In contrast, TERMNET produces more balanced clusters across sources. For instance, the 'Engineering' and 'IT and Computing' clusters contain substantial contributions from both OpenAlex and USPTO documents, suggesting that documents from different sources referring to similar concepts are aligned more effectively.

These observations suggest that TERMNET effectively balances expert-informed hierarchy guidance with data-driven clustering, producing hierarchies that are interpretable at high levels while capturing fine-grained, semantically coherent subtopics.

7 Case Study: Structural Change

To illustrate TERMNET's utility for applied regional analysis, we provide hierarchies created using TERMNET and the baselines to regional geographers analyzing the technological landscape of structurally weak German regions. For this case study, we used in-house data, substituting publicly available OpenAlex abstracts with Scopus abstracts for higher-quality metadata and using PATSTAT patents instead of USPTO patents to focus on German affiliations. The *horizon* and *foekat* funding datasets are used as before. TERMNET achieves the highest scores on most supervised and unsupervised metrics on this proprietary dataset (Table 4), demonstrating robust performance. For human evaluation, we added a *Usefulness* dimension to capture the practical utility of the hierarchies for geographers. Expert annotations were limited to the most promising approaches due to resource constraints.

We further evaluate clustering quality and cross-source integration using K-Means with $k = 650$ on three input types: raw text, summaries, and terms. Only a single clustering layer was generated. From each clustering, 50 documents were randomly sampled for blinded manual assessment. A geography expert judged whether each document was thematically consistent with its cluster, using the raw document text, representative keywords, representative cluster documents, and an LLM-generated cluster label. Term-centric representations achieved higher source entropy and human-perceived accuracy, indicating improved cross-source mixing and more coherent clusters (see Table 4).

8 Conclusion

We introduced TERMNET, a term-centric framework for inducing hierarchical taxonomies from heterogeneous document collections. By abstracting documents into domain-specific terms and combining seed-guided hierarchy construction with data-driven clustering, TERMNET produces human interpretable hierarchies that align with established domain structures while integrating information from multiple data sources. Experiments on multi-source datasets show consistent improvements in thematic alignment, semantic coherence, and cross-source integration, as reflected in both automatic metrics and expert evaluations.

Limitations

While TERMNET demonstrates strong performance in cross-source hierarchy induction, several limitations remain that offer directions for future research. First, the interpretability and structure of the upper hierarchy levels depend on the quality of the expert-provided seed categories. While this guidance improves alignment with established domain structures, poorly specified seeds may bias the resulting taxonomy. Second, our evaluation focuses

on English and German corpora within scientific and technological domains. The generalizability of the approach to other languages, domains, or document genres remains an open question. Third, the current implementation constructs a static hierarchy from a fixed snapshot of documents. In real-world technology monitoring scenarios, corpora evolve continuously as new publications and patents appear. Supporting incremental updates and modelling the temporal evolution of hierarchical structures would therefore be an important extension for future research.

Ethical Considerations

This work uses publicly available textual data sources and does not involve personal or sensitive information. Nevertheless, hierarchical representations learned from large corpora may reflect biases present in the underlying data, which could influence how topics or domains are structured and interpreted. Users should therefore treat automatically induced hierarchies as exploratory tools rather than authoritative representations of knowledge.

Large language models were used to assist with grammar correction, spelling, and refinement of the manuscript text. They were not used to generate experimental results, analyses, or scientific claims.

References

Shraey Bhatia, Jey Han Lau, and Timothy Baldwin. 2018. Topic intrusion for automatic topic model evaluation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 844–849, Brussels, Belgium. Association for Computational Linguistics.

Bundesministerium für Forschung, Technologie und Raumfahrt (BMFTR). 2026. FöKAT – Federal Catalog of Research Projects in Germany. Database. Used and published with permission. The dataset aggregates publicly available information on research, development, and innovation projects, including contact details of German research institutions (see Bundesbericht Forschung und Innovation, BuFI, https://www.bmftr.bund.de), available research reports via Technische Informationsbibliothek (TIB, https://www.tib.eu), and information on project funding managed by administrative project carriers (https://www.ptnetz.de). Additional federal funding sources include BMFTR, BMWE, BMV, BMUKN, BMJV, BMLEH.

Elsevier. 2024. All science journal classification (asjc) codes. https://service.elsevier.com/app/answers/detail/a_id/15181/supporthub/scopus/. Accessed: 2026-03-13.

Martin Ester, Hans-Peter Kriegel, J"org Sander, and Xiaowei Xu. 1996. A density-based algorithm for discovering clusters in large spatial databases with noise. In *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD'96)*, pages 226–231.

Muhan Gao, Jash Shah, Weiqi Wang, Kuan-Hao Huang, and Daniel Khashabi. 2025. Science hierarchography: Hierarchical organization of science literature. *Preprint*, arXiv:2504.13834.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Maarten R. Grootendorst. 2022. Bertopic: Neural topic modeling with a class-based tf-idf procedure. *ArXiv*, abs/2203.05794.

Kamran Hakiman and Chloe Stull-Lane. 2022. Innovation in governance: Integrating technical and contextual perspectives to address fragility. Technical report, SPARC.

Marti A. Hearst. 1992. Automatic acquisition of hyponyms from large text corpora. In *COLING 1992 Volume 2: The 14th International Conference on Computational Linguistics*.

Yuntong Hu, Zhuofeng Li, Zheng Zhang, Chen Ling, Raasikh Kanjiani, Boxin Zhao, and Liang Zhao. 2025. Taxonomy tree generation from citation graph. *Preprint*, arXiv:2410.03761.

Jiaxin Huang, Yiqing Xie, Yu Meng, Yunyi Zhang, and Jiawei Han. 2020. Corel: Seed-guided topical taxonomy construction by concept learning and relation transferring. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery amp; Data Mining*, KDD '20, page 1928–1936. ACM.

Uri Katz, Mosh Levy, and Yoav Goldberg. 2024. Knowledge navigator: LLM-guided browsing framework for exploratory search in scientific literature. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 8838–8855, Miami, Florida, USA. Association for Computational Linguistics.

Dongha Lee, Jiaming Shen, Seongku Kang, Susik Yoon, Jiawei Han, and Hwanjo Yu. 2022. Taxocom: Topic taxonomy completion with hierarchical discovery of novel topic clusters. In *Proceedings of the ACM Web Conference 2022*, WWW '22, page 2819–2829, New York, NY, USA. Association for Computing Machinery.

Shaobo Li, Jie Hu, Yuxin Cui, and Jianjun Hu. 2018. Deeppatent: patent classification with convolutional neural networks and word embedding. *Scientometrics*, 117(2):721–744.

Xueqing Liu, Yangqiu Song, Shixia Liu, and Haixun Wang. 2012. Automatic taxonomy construction from keywords. In *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '12, page 1433–1441, New York, NY, USA. Association for Computing Machinery.

J. B. MacQueen. 1967. *Some methods for classification and analysis of multivariate observations*, pages :281–297. University of California Press.

Christopher D. Manning, Prabhakar Raghavan, and Hinrich Sch"utze. 2008. *Introduction to Information Retrieval*. Cambridge University Press.

Leland McInnes, John Healy, and Steve Astels. 2017. hdbscan: Hierarchical density based clustering. *The Journal of Open Source Software*, 2(11):205.

David Mimno, Wei Li, and Andrew McCallum. 2007. Mixtures of hierarchical topics with pachinko allocation. In *Proceedings of the 24th International Conference on Machine Learning*, ICML '07, page 633–640, New York, NY, USA. Association for Computing Machinery.

Alexandru Oarga, Matthew Hart, Andres M Bran, Magdalena Lederbauer, and Philippe Schwaller. 2024. Scientific knowledge graph and ontology generation using open large language models. In *AI for Accelerated Materials Design - NeurIPS 2024*.

Alexander Panchenko, Stefano Faralli, Eugen Ruppert, Steffen Remus, Hubert Naets, Cédrick Fairon, Simone Paolo Ponzetto, and Chris Biemann. 2016. TAXI at SemEval-2016 task 13: a taxonomy induction method based on lexico-syntactic patterns, substrings and focused crawling. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 1320–1327, San Diego, California. Association for Computational Linguistics.

Jan Polchar. 2024. Using foresight to anticipate emerging critical risks: Proposed methodology. Technical Report 79, OECD Publishing, Paris.

Jason Priem, Heather Piwowar, and Richard Orr. 2022. Openalex: A fully-open index of scholarly works, authors, venues, institutions, and concepts. *Preprint*, arXiv:2205.01833.

Publications Office of the European Union. 2015. CORDIS - EU Research Projects under Horizon 2020 (2014–2020). Accessed: 2026-03-13.

Publications Office of the European Union. 2022. CORDIS - EU Research Projects under HORIZON EUROPE (2021–2027). Accessed: 2026-03-13.

D. Sculley. 2010. Web-scale k-means clustering. In *Proceedings of the 19th International Conference on World Wide Web*, WWW '10, page 1177–1178, New York, NY, USA. Association for Computing Machinery.

Elena Senger, Yuri Campbell, Rob Van Der Goot, and Barbara Plank. 2025. Crossing domains without labels: Distant supervision for term extraction. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 1366–1378, Suzhou (China). Association for Computational Linguistics.

Yanzhen Shen, Yu Zhang, Yunyi Zhang, and Jiawei Han. 2025. A unified taxonomy-guided instruction tuning framework for entity set expansion and taxonomy expansion. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 3208–3220, Vienna, Austria. Association for Computational Linguistics.

Vered Shwartz, Yoav Goldberg, and Ido Dagan. 2016. Improving hypernymy detection with an integrated path-based and distributional method. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2389–2398, Berlin, Germany. Association for Computational Linguistics.

Silvia Terragni, Elisabetta Fersini, and Enza Messina. 2021. Word embedding-based topic similarity measures. In *International conference on applications of Natural Language to information systems*, pages 33–45. Springer.

Chi Wang, Marina Danilevsky, Nihit Desai, Yinan Zhang, Phuong Nguyen, Thrivikrama Taula, and Jiawei Han. 2013. A phrase mining framework for recursive construction of a topical hierarchy. In *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '13, page 437–445, New York, NY, USA. Association for Computing Machinery.

William Webber, Alistair Moffat, and Justin Zobel. 2010. A similarity measure for indefinite rankings. *ACM Transactions on Information Systems (TOIS)*, 28(4):1–38.

Chao Zhang, Fangbo Tao, Xiusi Chen, Jiaming Shen, Meng Jiang, Brian Sadler, Michelle Vanni, and Jiawei Han. 2018. Taxogen: Unsupervised topic taxonomy construction by adaptive term embedding and clustering. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD '18, page 2701–2709, New York, NY, USA. Association for Computing Machinery.

Yanzhao Zhang, Mingxin Li, Dingkun Long, Xin Zhang, Huan Lin, Baosong Yang, Pengjun Xie, An Yang, Dayiheng Liu, Junyang Lin, Fei Huang, and Jingren Zhou. 2025. Qwen3 embedding: Advancing text embedding and reranking through foundation models. *Preprint*, arXiv:2506.05176.

Kun Zhu, Lizi Liao, Yuxuan Gu, Lei Huang, Xiaocheng Feng, and Bing Qin. 2025. Context-aware hierarchical taxonomy generation for scientific papers via LLM-guided multi-aspect clustering. In *Proceedings of the 2025 Conference on Empirical Methods in*

Natural Language Processing, pages 15627–15645, Suzhou, China. Association for Computational Linguistics.

A Details on the Dataset Creation and Licensing

A.1 Dataset Creation

This appendix provides details on the data retrieval, filtering, and text construction steps applied to each source in the dataset.

- OpenAlex (Priem et al., 2022): We retrieved publications via the OpenAlex API with at least one German-affiliated author for the period 2015–2023. Only records with a DOI were retained. The text corpus was constructed by concatenating the `title` and `abstract` fields.
- USPTO (Li et al., 2018): We used patent records from the USPTO dataset for the years 2014–2015. The textual content consists of the `Abstract` field.
- CORDIS: We used project data from Horizon 2020 (funded 2014-2020) (Publications Office of the European Union, 2015) and Horizon Europe (funded 2021-2027) (Publications Office of the European Union, 2022). The text corpus combines `ProjectTitle` and `ProjectObjective`.
- FöKAT (Bundesministerium für Forschung, Technologie und Raumfahrt (BMFTR), 2026): We used funding records from 2014–2025, with the year determined by `Laufzeit von`. The text corpus consists of the `Thema` field. Compared to the other datasets, FOEKAT entries are substantially shorter and frequently contain funding-specific terminology or proper names of funding programs without descriptive context. To mitigate this, domain experts excluded selected categories from the FÖKAT taxonomy (*Leistungsplansystematik*) and applied additional keyword-based filtering to remove highly specific funding instruments (e.g., *Professorinnenprogramm*).

Across all sources, we additionally removed extremely short entries (less than five characters).

A.2 Licensing

The benchmark dataset aggregates information from United States Patent and Trademark Office (USPTO) patents (CC BY 4.0), OpenAlex publications (CC0), CORDIS project data (generally available under CC BY 4.0 / PSI reuse terms), and FöKAT records used with permission. To ensure license compatibility and proper attribution requirements across sources, the compiled benchmark dataset is released under the Creative Commons Attribution 4.0 (CC BY 4.0) license.

Algorithm 2: Recursive K-Means hierarchy construction

Input: Set of documents $\{d_i, \ldots, d_n\}$, number of layers L, target number of clusters $(k_1, \ldots, k_L)$
Initialization: Embed each d_i
for $l = 1$ **to** L **do**
  if $l = 1$ **then**
    Cluster all documents into k_1 clusters
  else
    for *each cluster* c *from layer* $l - 1$ **do**
      Cluster documents of c into k_l subclusters
return *Hierarchical structure*

Algorithm 3: Scychic algorithm

Input: Set of documents $\{d_i, \ldots, d_n\}$, number of layers L, target number of clusters $(k_1, \ldots, k_L)$
Initialization: Embed each d_i
for $l = 1$ **to** $\lfloor L/2 \rfloor$ `// Top-down phase`
do
  if $l = 1$ **then**
    Cluster all documents into k_1 clusters
  else
    for *each cluster* c *from layer* $l - 1$ **do**
      Cluster documents of c into k_l subclusters
for *each cluster* τ *at level* $\lfloor L/2 \rfloor$ `// Bottom-up phase`
do
  for $l = L$ **down to** $\lfloor L/2 \rfloor + 1$ **do**
    if $l = L$ **then**
      E = { embeddings of documents within τ }
    else
      E = { embeddings of summaries of cluster $l + 1$ }
    Cluster E into k_l subclusters
    Generate summary for each subcluster
return *Hierarchical structure*

B Algorithms

In Algorithm 2, we present the classical approach to hierarchical clustering using recursive K-Means. In Algorithm 3, we reproduce the SCYCHIC algorithm, introduced by Gao et al. (2025).

C Details on Human Evaluation

The four dimonsions for the Likert scale are:

- **Structure:** Does the organization follow a clear logical hierarchy, transitioning from broad research areas to specific sub-technologies, or do the clusters maintain a consistent level of granularity on each layer?

- **Validity:** Does the taxonomy align with experts' understanding of the scientific and technological landscape, including how concepts are commonly grouped, related, or distinguished?

- **Composition:** Does the clusters reflect an appropriate and plausible composition of data sources given the nature of the technology (e.g., publication-dominated scientific fields versus patent- or project-intensive applied domains)?

- **Usefulness (Case Study):** How useful is the hierarchy for policy analysts or foresight practitioners seeking to monitor technological developments, compare signals across data sources, or identify emerging or underexplored areas?

The 1-5 Likert Scale is defined as:

1. **Completely inaccurate**, with significant factual errors or misrepresentations of the domain.

2. **Mostly inaccurate**, capturing only a few correct facts but failing to represent the domain coherently.

3. **Moderately accurate**, containing some factual correctness but missing important concepts or relationships.

4. **Mostly accurate**, representing the domain well with minor factual inaccuracies or omissions.

5. **Highly accurate**, thoroughly reflecting the domain's factual structure with no noticeable errors.