Are Scientific Annotations Consistently Represented across Science Knowledge Graphs?

Jenifer Tabita Ciuciu-Kiss[1], Daniel Garijo[1]

[1]Universidad Politécnica de Madrid, Boadilla del Monte, Madrid, Spain

Abstract

Scientific Knowledge Graphs (SKGs) are increasingly used to annotate and interlink research outputs. However, little is known about how consistently they annotate the same publication. This paper presents a comparative analysis of category annotations across four major SKGs (ORKG, OpenAlex, OpenAIRE, and Papers with Code) using a manually curated gold-standard dataset of 70 AI-related papers. We examine differences in annotation coverage, granularity, and semantic alignment, highlighting frequent inconsistencies such as label mismatches, overly generic terms, and coverage gaps. Our analysis reveals that manual curation offers high-quality but sparse annotations, while automated systems achieve broader coverage at the cost of precision. This work contributes insights into the reliability of SKG metadata and outlines pathways for improving interoperability and annotation practices.

Keywords

Science Knowledge Graphs, Comparative Analysis, Metadata Quality

1. Introduction

In recent years, Scientific Knowledge Graphs (SKGs) [1, 2] have become essential infrastructures for representing scholarly information [3] in a machine-readable format [4]. By linking research entities [5] such as publications, datasets, software, authors, and their associated annotations, SKGs enable advanced services for scientific discovery [6, 7], evaluation, and reuse. While these infrastructures do not fully realize all aspects of the FAIR principles [8, 9, 10], they fall short particularly in interoperability and reusability. Annotation vocabularies are rarely harmonized across platforms, documentation of classification pipelines is often incomplete, and provenance metadata is inconsistently recorded. Nonetheless, they contribute toward findability and partial interoperability through metadata enrichment [11], standardized identifiers, and the homogenization [12] of scholarly records. Examples include OpenAlex [13], OpenAIRE [14, 15, 16, 17], the Open Research Knowledge Graph (ORKG) [18], AI-KG [19], Crossref [20], and Papers with Code (PwC)[1] among others, each offering its own approach to structuring and classifying research outputs.

A core functionality of these graphs is the annotation of research publications [21, 22], typically through labels such as subjects [16], concepts [13], or tasks [18][1] with or without a hierarchy. These annotations are key for enabling semantic search [23], recommendation systems [24], benchmarking platforms [25], and large-scale meta-analyses [26]. However, SKGs differ substantially in how they generate and apply such labels, ranging from manual curation [18][1] to automated topic modeling [13, 18], resulting in inconsistent representations of the same scientific work among various sources.

While prior studies have explored overlaps between SKGs through quantitative methods [27], such as measuring lexical similarity between annotations, there remains limited understanding of how SKG category annotations differ in practice when describing the same publication across multiple SKGs. In this paper we explore these differences, which may stem from divergent modeling assumptions, annotation pipelines, and classification goals. We present a comparative analysis of 70 AI-research

5th International Workshop on Scientific Knowledge: Representation, Discovery, and Assessment, Nov 2024, Nara, Japan
jenifer.ciuciu-kiss@alumnos.upm.es (J. T. Ciuciu-Kiss); daniel.garijo@upm.es (D. Garijo)
https://jeniferciuciukiss.com/ (J. T. Ciuciu-Kiss); https://dgarijo.com/ (D. Garijo)
0000-0002-3170-6730 (J. T. Ciuciu-Kiss); 0000-0003-0454-7145 (D. Garijo)

[1]https://paperswithcode.com/

papers from recent years, annotated across four major SKGs: ORKG, OpenAlex, OpenAIRE, and PwC. We examine the types of inconsistencies that emerge when annotating the same publications, such as mismatches in granularity, terminology, coverage, and discuss their implications for interoperability, metadata quality, and downstream applications.

To better understand SKG-based annotations, we formulate the following research questions, each paired with the main contribution that addresses it:

- **RQ1:** How do category annotations differ across SKGs?
  We construct and release a manually curated dataset of 70 AI-related publications, each annotated across four major SKGs: ORKG, OpenAlex, OpenAIRE, and PwC. These annotations include tasks, methods, subjects, and other topical labels. We use this dataset to compare how SKGs differ in annotation scope, specificity, and structural conventions. Throughout this paper, we refer to such labels as (category) annotations.
- **RQ2:** How accurate are these annotations compared to a manually curated standard?
  We manually reviewed the title and abstract of each paper to determine whether the SKG-assigned annotations accurately reflected the paper's content. Based on this expert validation, we constructed a gold-standard dataset and evaluated each SKG's annotations in terms of precision, recall, and F1-score [28].
- **RQ3:** What types of annotation inconsistencies occur most frequently?
  We conduct a comparative evaluation across the four SKGs, identifying frequent inconsistencies such as mismatches in granularity, label ambiguity, incomplete coverage, and semantic misalignment. We further analyze these issues through representative examples and quantitative summaries.

The remainder of this paper is structured as follows. Section 2 reviews related work on SKG-based classification and metadata annotation. Section 3 outlines our methodology, including dataset construction, annotation guidelines, and evaluation metrics. Section 4 describes the initial and gold-standard datasets used for the comparative analysis. Section 5 presents empirical results on annotation coverage, accuracy, and overlap across SKGs. Section 6 discusses the findings in light of key annotation challenges and provides representative examples. Finally, Section 7 concludes with a summary of insights and recommendations for improving annotation practices in scientific knowledge graphs.

1.1. Background: Scientific Knowledge Graphs (SKGs)

SKGs are structured representations of scholarly knowledge [29] that encode entities (e.g. publications and concepts) and their semantic relationships in a graph-based format. Their primary aim is to support advanced search, integration, and analysis of scientific information by making research outputs machine-interpretable and interlinked. Depending on their design goals, SKGs differ in scope, domain coverage, and update mechanisms, ranging from large-scale, automatically constructed graphs to smaller, community-curated platforms. In the following, we discuss the key characteristics of some of the most widely used SKGs that serve as the foundation for our analysis.

1.1.1. OpenAlex

OpenAlex [13] is an open catalog of scholarly entities that emerged as the successor of the Microsoft Academic Graph [30]. It compiles metadata on publications, authors, institutions, venues, concepts, and more. OpenAlex applies machine learning (ML) models trained on titles, abstracts, and citation contexts to assign fine-grained topic annotations from a curated ontology of over 60,000 concepts. These topic concepts are assigned probabilistically, with each publication receiving a primary concept and possibly several secondary ones, each associated with confidence scores. The classification pipeline is documented and accessible through the OpenAlex API[2]. As one of the largest open scholarly KGs, OpenAlex

[2] https://docs.openalex.org/api-entities/topics

prioritizes breadth and scalability but, due to its automated nature, it may introduce inconsistencies in category granularity and semantic relevance, particularly across disciplines.

1.1.2. OpenAIRE - Open Access Infrastructure for Research in Europe

Open Access Infrastructure for Research in Europe (OpenAIRE) [14, 15, 16, 17] is a major European Open Science infrastructure designed to foster open scholarship and improve the accessibility and reusability of scientific knowledge. The OpenAIRE Knowledge Graph aggregates metadata from a broad spectrum of sources, including publications, datasets, projects, and research organizations across Europe, thus providing an integrated view of the research landscape. The OpenAIRE APIs[3] offer access to this aggregated metadata, enabling structured queries over scholarly content. In this work, we used the Search API[4][5] to retrieve category-level annotations for individual publications. OpenAIRE's metadata model is enriched with subject classifications based on taxonomies such as the OECD Fields of Science (FOS), SCINOBO[6] and others integrated across its pipeline. As a component of the European Open Science Cloud, OpenAIRE benefits from frequent updates and ongoing standardization efforts. It supports the discoverability and interoperability of scientific content through harmonized metadata ingestion from compliant repositories and data providers.

1.1.3. ORKG - The Open Research Knowledge Graph

ORKG [18][7] offers a semantic infrastructure for representing individual research contributions using RDF and structured templates. Unlike fully automated SKGs, ORKG relies on manual, community-driven annotations where users describe publications through semantically rich triples that capture the problem, method, and result of a study. Annotations are made using predefined templates that align with scholarly discourse elements, enabling fine-grained semantic modeling of contributions. The system is designed to increase interpretability and transparency of research metadata, supporting both manual entry and semi-automated extraction tools. While its manual approach limits coverage compared to large-scale automated graphs, the semantic depth and precision of ORKG annotations make it especially valuable for comparative analyses.

1.1.4. PwC - Papers with Code

PwC[8] is a domain-specific SKG focused on the AI/ML research landscape. It integrates scientific publications, benchmark datasets, evaluation results, and source code into a coherent, task-driven knowledge graph. Each paper is linked to tasks and methods, with annotations derived via a hybrid pipeline combining automated extraction and human curation. PwC sources its papers primarily from arXiv and Crossref, and then connects them to relevant benchmarks and method families, drawing from curated taxonomies that reflect the evolving state of the field. The PwC dataset is regenerated daily, ensuring that new papers and updated annotations are continuously incorporated. Labels are reviewed and maintained by moderators and contributors from the research community, which supports high-quality and fine-grained annotations useful for reproducibility studies and trend analysis. Metadata and category information are available through the PwC platform and associated GitHub repositories.[9] Previous work has examined the consistency and accuracy of existing method quality in PwC, highlighting both the strengths and the limitations in coverage and granularity [31].

[3] https://graph.openaire.eu/docs/apis/home/
[4] https://graph.openaire.eu/docs/apis/search-api/
[5] https://api.openaire.eu/search/publications
[6] https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10192702/
[7] https://www.orkg.org/
[8] https://paperswithcode.com/
[9] https://github.com/paperswithcode

2. Related Work

This section describes the diverse resource categorization techniques employed by different SKGs, highlighting their methodological differences and implications for consistency in annotations. SKGs employ varied annotation strategies to assign semantic categories to scholarly works. These strategies differ in automation level, interpretability, and domain specificity. Below, we categorize these approaches into four main types: rule-based/metadata, NER/linking, topic modeling/classification, and hybrid/manual curation.

2.1. Metadata- and taxonomy-based classification

Platforms like OpenAIRE [16, 17] and Crossref [20] rely on repository metadata and established taxonomies (e.g., OECD Fields of Science [32], SCINOBO [33], ACM CSS [34, 35]) to assign broad subject categories to publications [14, 15]. This strategy enables scalable and harmonized annotation of research outputs using well-established classification schemes, contributing to metadata interoperability and integration across repositories. However, it typically produces coarse-grained labels that may lack domain specificity. Notably, many SKGs using metadata-based strategies provide limited documentation about how repositories map local tags to global taxonomies, which introduces opacity into the categorization pipeline.

2.2. Named-entity recognition (NER) and entity linking

Some SKGs extract entities directly from unstructured text using NER and linking to external knowledge bases (e.g., Wikidata [36], MeSH [37][10]) [38]. This approach enables direct annotation of domain-specific entities (e.g., genes, diseases, methods), which is particularly useful in specialized fields like biomedicine, for instance, biomedical SKGs often detect gene, disease, or method mentions via concept recognition tools, followed by normalization to identifiers. These annotations can enhance semantic granularity and support knowledge integration. However, such pipelines are unevenly documented, with some SKGs omitting details about their training data and linking heuristics [19, 39]. This opacity complicates reproducibility and comparison across systems.

2.3. Topic modeling and supervised classification

Large-scale SKGs such as OpenAlex apply ML to cluster publications and assign topics based on features extracted from titles, abstracts, venue names, and citation networks. These annotations can enhance semantic granularity and support knowledge integration, for example, OpenAlex's topic pipeline uses network clustering to define topic communities, labels them via LLMs, and employs a deep learning classifier to annotate works[11]. Such approaches can surface emerging topics and latent structure in scientific literature. However, they may result in inconsistent granularity, with some topics being overly broad and others overly specific.

2.4. Hybrid human-in-the-loop annotation

Domain-specific SKGs, such as PwC, combine automated matching of papers to predefined task/method taxonomies with manual curation by community moderators1. This hybrid approach leverages the scalability of automation while incorporating expert validation to improve annotation accuracy. Such strategies strike a balance between breadth and quality: automated systems offer scalability, while manual review ensures semantic precision. Similarly, ORKG uses user-defined templates (e.g., method, result, etc.), filled manually, to produce highly structured and semantically rich metadata[12]. These annotations provide detailed and interpretable representations of research contributions. This yields precise

[10] https://www.nlm.nih.gov/mesh/meshhome.html
[11] https://docs.openalex.org/api-entities/topics
[12] https://orkg.org/stats

results, but coverage remains limited, and the user-dependent process is not uniformly documented across contributions. Moreover, manual curation may introduce subjective bias, as moderators may apply labels based on individual interpretations, experience, or familiarity with specific research areas.

Implications of strategy diversity on annotation consistency The diversity of annotation strategies across SKGs introduces both strengths and challenges for metadata consistency. Metadata-based systems produce generalized labels with limited thematic depth; topic modeling yields probabilistic but uneven concept assignments, and NER/linking systems vary based on entity recognition quality and KB integration. Hybrid manual approaches deliver semantically rich labels but lack scalability and full traceability, particularly when documentation of contributor workflows is absent. Understanding the trade-offs of each strategy is essential for improving interoperability and annotation quality. Our work contributes to this area by performing a fine-grained, paper-level comparison across SKGs, focusing on semantic overlap, divergence, and contextual usage of categories.

3. Methodology

To assess the consistency and semantic validity of annotations across SKGs, we designed a two-stage evaluation pipeline: (1) data collection and preparation, and (2) comparative analysis based on manual validation of annotation correctness. All annotations were verified by a domain expert against the content of each publication by manually inspecting titles and abstracts. To mitigate potential bias in manual validation, we followed predefined rules (see Section 3.1), avoided adding new annotations, and applied a conservative exclusion strategy. Future work will incorporate multiple annotators and inter-annotator agreement to strengthen the reproducibility of our results.

3.1. Data Collection

In the first stage of our methodology, we assembled a dataset of AI-related research papers that are jointly annotated across four prominent Scientific Knowledge Graphs (SKGs): ORKG, OpenAlex, OpenAIRE, PwC. These SKGs were selected to represent a spectrum of annotation strategies: from fully manual (ORKG), to hybrid human-in-the-loop (PwC), to fully automated systems (OpenAlex and OpenAIRE). This diversity allowed us to conduct a balanced comparison of annotation behavior across different design paradigms.

Each SKG defines its own scope and indexing strategy, focusing on different disciplines, sources, or publication types, which naturally leads to variations in which papers are included and how they are annotated. As a result, it is common for a paper to appear in one SKG but not another, or to be indexed without any associated annotations. This diversity in coverage is expected and reflects the design priorities of each graph rather than inconsistencies.

To enable a controlled comparative analysis, we selected only papers that were indexed and annotated by all four SKGs. Although the SKGs use distinct terminology for annotations, such as “tasks” and “methods” (PwC), “research problems” (ORKG), or broader “subjects” (OpenAIRE, OpenAlex)—we refer to all such labels uniformly as annotations throughout this paper..

Constructing a dataset with full parallel annotations across multiple SKGs required scanning a large candidate pool of papers and applying an iterative filtering process. Only those papers for which each SKG provided at least one annotation were retained for the final analysis. The resulting dataset and its properties are described in detail in Section 4.

Paper Selection To build the dataset, we began by compiling a broad pool of AI-related research papers published between 2023 and 2025. Candidate papers were selected based on the authors’ domain expertise and covered a wide range of topics within Artificial Intelligence. For each paper, we attempted to match entries across the four selected SKGs using persistent identifiers (primarily DOIs) and title matching. Because each SKG has a different scope and indexing strategy, many papers were not fully

covered in all four. Some were absent from one or more SKGs, while others were indexed but lacked relevant annotations. To ensure a fair and controlled comparison, we retained only papers that (1) were indexed in all four SKGs and (2) had at least one annotation from each. This filtering process was applied iteratively to an initial, approximately 200 papers, resulting in a final set of 70 papers that satisfied the completeness criteria for comparative analysis.

Categorization Retrieval Once the final set of papers was selected, we retrieved their corresponding annotations from each of the four SKGs. The retrieval process was adapted to the access mechanisms and data availability of each source. Specifically:

- For OpenAlex, we used the official API[13] to retrieve topic and concept annotations associated with each paper's DOI.
- For OpenAIRE, we queried the Search API[14] to obtain subject classifications based on the OECD Fields of Science taxonomy.
- For ORKG, we used the public SPARQL endpoint[15] to extract structured annotations based on predefined semantic templates. In particular, we retrieved the `hasResearchProblem` and `hasMethod` fields from each research contribution.
- For PwC, we used a local data dump[16] accessed on July 1, 2025, to extract annotations for tasks and methods associated with each paper.

Annotations were collected and stored separately for each SKG, preserving their original format, structure, and terminology. No filtering or transformation was applied during this stage, to ensure that the data remained faithful to its source. This raw annotation set served as the input for the normalization and comparative analysis steps described in the following sections.

Initial Dataset: Normalization The initial dataset was constructed directly from the raw annotations retrieved from each SKG. To ensure basic consistency across sources, all annotation labels were transformed to lowercase. In addition, whenever annotations were expressed using OECD Fields of Science (FoS) codes or non-standard descriptors, these were replaced with their corresponding standard FoS labels. No further transformations, filtering, or reformatting were applied at this stage. This version of the dataset preserves the original annotation behavior of each SKG and is referred to as the *initial dataset* in the rest of the paper.

Gold-Standard Dataset: Manual Validation To assess annotation correctness, we manually curated a gold-standard by reviewing each paper's title and abstract. Validation was performed by the first author (a PhD researcher specializing in AI and SKGs with over 5 years of experience). Each annotation was evaluated against three criteria: (i) semantic relevance to the paper's research problem or method, (ii) domain specificity (avoiding overly generic categories such as 'science'), and (iii) contextual accuracy. Borderline cases were conservatively excluded. This rule-based approach was adopted to minimize subjective bias. Annotations that were relevant were marked as correct and therefore kept in the final gold-standard dataset, while those that were off-topic, overly generic, overly specific, or misleading were marked as incorrect. In borderline cases, we adopted a conservative approach and excluded such annotations from the gold-standard.

During this process, we also identified cases where the abstracts retrieved from certain SKGs were incomplete, incorrect, or contained metadata artifacts. These cases were corrected manually using the official abstracts from publisher websites or arXiv [40] to ensure that our validation was based on accurate representations of the paper content.

[13] https://api.openalex.org/works
[14] https://api.openaire.eu/search/publications
[15] https://orkg.org/sparql
[16] https://paperswithcode.com/about

No new annotations were introduced during this step; the gold-standard only reflects corrections to existing labels and underlying metadata. This version of the dataset [41] is used in our evaluation of annotation accuracy in Section 5.

3.2. Comparative Analysis

To evaluate the consistency and semantic appropriateness of category annotations across SKGs, we conducted a comparative analysis that combines quantitative metrics with qualitative interpretation. This twofold approach allowed us to assess both the overall annotation performance of each SKG and the nature of discrepancies that emerge when multiple SKGs describe the same publication.

Using the gold-standard dataset, we first evaluated annotation correctness in terms of precision, recall, and F1-score for each SKG. These metrics quantify the alignment between the annotations and expert-validated labels, forming the basis of the results presented in Section 5.

To complement the evaluation, we then analyzed the types of inconsistencies that commonly arise across SKGs. For this purpose, we read and interpreted the title and abstract of each paper to identify recurring annotation issues and anomalies. In particular, we distinguish four main types of inconsistencies:

- **Coverage inconsistency**: Cases where a paper was present in all four SKGs but one or more SKGs provided no annotation. Importantly, no new categories were introduced during our process; coverage was judged solely based on whether each SKG offered at least one valid label.
- **Label mismatch**: Use of different terms to describe the same concept (e.g., "NER" vs. "Named Entity Recognition"), reflecting differences in vocabulary and annotation conventions.
- **Granularity difference**: One SKG uses broad categories (e.g., "Computer Vision") while another applies fine-grained concepts (e.g., "Panoptic Segmentation"), complicating direct comparison.
- **Incorrect category assignment**: A category is clearly misaligned with the content of the paper—such as labeling an NLP paper as "Computer Vision"—often due to automatic inference errors or misinterpreted metadata.

Each inconsistency was documented at the paper level, and summary statistics were compiled to capture their distribution across the dataset. Representative examples and edge cases are discussed in Section 6 to illustrate common pitfalls, semantic drift, and limitations in current SKG annotation practices. All code used for dataset construction and analysis is available on GitHub [42] and also published as a snapshot on Zenodo [41].

4. Dataset

The two datasets used in our analysis were constructed following the methodology described in Section 3. Both the initial dataset and the manually curated gold-standard dataset are publicly available on Zenodo [41]. Both datasets include the exact same set of 70 AI-related research papers from 2023–2025, each annotated by all four SKGs. Table 1 presents key statistics for the initial dataset and the manually curated gold-standard dataset.

Table 1
Summary statistics for the initial and gold-standard datasets.

Metric	Initial Dataset	Gold-Standard Dataset
Total annotations (across SKGs)	2756	1046
Average annotations per paper	39.37	14.94
Average annotations per paper per SKG	9.84	3.78
Unique category labels	728	300

Initial dataset reflects the raw annotations retrieved from each SKG, with only minimal normalization applied, such as converting to lowercase and replacing classification codes when necessary. This version of the dataset contains a total of 2,756 annotations, which corresponds to an average of 39.37 annotations per paper and 9.84 annotations per paper per SKG. The dataset includes 728 unique category labels, illustrating the broad topical coverage and terminological diversity across SKGs. However, this volume also introduced substantial noise, redundancy, and inconsistency, particularly in cases where overly generic or highly specific terms inflated the annotation count.

The **Gold-standard dataset** builds on the initial version by incorporating manual validation of each annotation. Using the title and abstract of each paper, we assessed whether the assigned categories accurately reflected the main research topic or contribution. Annotations deemed off-topic, overly broad, overly specific, or ambiguous were removed. In a few cases, missing/incorrect abstracts were also corrected manually. This refinement reduced the dataset to 1,046 total annotations—an average of 14.94 annotations per paper and 3.78 per paper per SKG. The number of unique category labels decreased to 300, resulting in a cleaner and more semantically coherent label set suitable for evaluation purposes.

The contrast between the two datasets highlights the tendency of automated and hybrid SKG pipelines to overgenerate annotations. While the initial dataset captures the full breadth of current SKG outputs, the gold-standard version provides a human-validated benchmark that filters out noise and prioritizes interpretability.

5. Results

This section presents the results of our analysis, structured around the research questions (RQs) introduced in Section 1. Each subsection restates the corresponding RQ and provides a detailed response based on comparative findings on the proposed gold-standard dataset consisting of 70 AI-related research papers, annotated across the four SKGs.

5.1. RQ1: How do annotation strategies differ across SKGs?

To examine how annotation practices vary across SKGs, we analyzed the average number of annotations per paper and the number of unique category labels for each graph, both before and after gold-standard curation. Table 2 summarizes these results. The initial dataset reveals significant differences in annotation strategies across SKGs. PwC assigns the highest number of categories per paper (16.73 on average), followed by OpenAlex (12.39), OpenAIRE (7.43), and ORKG (2.83). While OpenAlex exhibits the broadest vocabulary with 277 unique categories, ORKG, despite its lower per-paper average—maintains 133 distinct labels, pointing to a focused yet diverse annotation strategy. After manual curation, annotation counts dropped significantly across all SKGs, ranging from 4.66 (PwC) to 1.93 (ORKG), reflecting reductions of over 50% in all cases. Overall, the results confirm that SKGs differ widely in both the volume and nature of their annotations. Automated systems like OpenAlex and OpenAIRE aim for broad coverage, but differ in granularity and topical precision. OpenAlex produces more annotations and a broader vocabulary, while OpenAIRE remains more conservative. PwC, despite being hybrid, assigns the highest number of categories per paper, suggesting a bias toward exhaustive labeling that introduces redundancy. In contrast, ORKG assigns far fewer annotations on average, but maintains a high number of distinct category labels, indicating a more targeted and semantically diverse annotation strategy.

To further explore how SKGs align in their annotation choices, we computed the number of overlapping categories assigned per paper for each pair and triplet of SKGs. Table 3 summarizes the total number of shared annotations observed across 70 papers. Overlaps were relatively sparse, underscoring the inconsistency in how different SKGs annotate the same paper. The highest pairwise agreement occurred between OpenAlex and OpenAIRE (71 overlapping categories), likely due to their shared reliance on automated subject classification at a broad scope. In contrast, overlap between ORKG and OpenAIRE was negligible (1 overlap), reflecting their distinct coverage and semantic focus. Notably, there were zero papers where all four SKGs assigned at least one identical category, and only two triple

Table 2
Annotation statistics per SKG in the initial and gold-standard datasets.

Metric	PwC	OpenAlex	OpenAIRE	ORKG
Initial Dataset				
Avg. categories per paper	16.73	12.39	7.43	2.83
# Unique categories	198	277	157	133
Gold-Standard Dataset				
Avg. categories per paper	4.66	4.84	3.67	1.93
# Unique categories	119	96	38	75

combinations (PwC–OpenAlex–ORKG and OpenAlex–OpenAIRE–ORKG) resulted in even a single shared category across 70 papers. These results reinforce the conclusion that while SKGs may annotate the same papers, they do so using divergent taxonomies and strategies, limiting interoperability and semantic alignment.

Table 3
Category overlaps across SKGs in the **gold-standard dataset**. Values represent total number of overlapping categories across 70 papers.

SKG Combination	Pairwise Overlap
PwC & OpenAlex	18
PwC & OpenAIRE	4
PwC & ORKG	7
OpenAlex & OpenAIRE	71
OpenAlex & ORKG	3
OpenAIRE & ORKG	1

5.2. RQ2: How accurate are the annotations compared to a manually curated gold-standard?

To assess annotation correctness, we evaluated how well the labels assigned by each SKG aligned with the manually curated gold-standard. For each of the 70 AI-related papers, the gold-standard contains only those annotations that were present in the original SKG outputs and judged to be semantically correct based on the paper's title and abstract. No new categories were added—evaluation was performed purely within the set of originally retrieved labels.

Table 4
Annotation precision, recall and F1 score per SKG, including total and average annotations per paper in both datasets.

Metric	PwC	OpenAlex	OpenAIRE	ORKG
Initial Dataset				
Total Annotations	1171	867	520	198
Avg. Annotations per Paper	16.73	12.39	7.43	2.83
Gold-Standard Dataset				
Total Annotations	317	339	257	133
Avg. Annotations per Paper	4.66	4.84	3.67	1.93
Precision	0.27	0.39	0.35	0.66
Recall	0.99	1.00	0.72	0.98
F1-score	0.42	0.56	0.47	0.79

As shown in Table 4, SKGs vary significantly in annotation quality. PwC and OpenAlex achieve nearly

perfect recall (0.99 and 1.00, respectively), meaning most relevant labels are included in their original outputs. However, both suffer from low precision, 0.27 for PwC and 0.39 for OpenAlex, indicating a high proportion of irrelevant, redundant, or overly specific annotations. These results reflect their emphasis on breadth and automated category expansion.

ORKG, in contrast, produces far fewer annotations but with significantly higher semantic alignment, yielding the highest precision (0.66) and F1-score (0.79). OpenAIRE falls between the two extremes, offering a trade-off between coverage and correctness with a precision of 0.35 and recall of 0.72.

Labeling methodology. An annotation was marked as *correct* if it matched a label in the gold-standard exactly (case-insensitive). All comparisons were made using string matching; minor spelling differences (e.g., *modeling* vs. *modelling*) were considered incorrect. Duplicates were collapsed and not penalized. Importantly, the evaluation focuses solely on correctness relative to the gold-standard, it does not assess whether the remaining categories are optimal or comprehensive. For instance, PwC might include appropriate categories that are semantically correct yet filtered out due to being overly specific or redundant.

Metric computation. Metrics were computed using the `scikit-learn` functions `precision_score`, `recall_score`, and `f1_score`, with `zero_division=0`. For each SKG and paper, we created a binary vector over all labels (union of predicted and gold) to indicate presence/absence. These vectors were concatenated across the 70 papers and aggregated into global metrics, offering a strict but comparable evaluation of annotation quality across SKGs.

5.3. RQ3: What types of annotation inconsistencies occur most frequently?

Since the gold-standard dataset was created by manually filtering the initial annotations, the most frequent inconsistency was overannotation—labels that did not align with the paper's content. No new categories were added; instead, irrelevant, redundant, overly generic, or excessively specific annotations were removed. A small number of coarse-grained categories were refined into more precise terms, and typographical variants (e.g., British vs. American spelling) were harmonized for consistency.

Table 5 summarizes the extent of label filtering per SKG. For each graph, we report the number of initial annotations, retained gold-standard annotations, incorrect labels removed, and the average number of incorrect annotations per paper. Except for two isolated cases, all papers contained at least one correct annotation per SKG, validating the application of the evaluation metrics.

Table 5
Annotation overassignment per SKG: number of annotations before and after manual filtering, total removed, and average number of incorrect annotations per paper.

Metric	PwC	OpenAlex	OpenAIRE	ORKG
Initial Annotations	1018	801	479	184
Gold-standard Annotations	243	311	228	123
Overassigned Annotations Removed	779	490	251	61
Avg. Incorrect Annotations per Paper	12.56	7.90	5.11	0.87

The results show that PwC and OpenAlex contributed the most noise, with 779 and 490 incorrect annotations respectively—averaging over 12 and nearly 8 errors per paper. OpenAIRE followed with moderate overannotation, while ORKG was the most conservative, averaging fewer than one incorrect label per paper. Overall, 1,581 out of 2,482 initial annotations (64%) were removed, highlighting the need for improved quality control, context-sensitive categorization, and curated vocabularies to enhance SKG reliability.

6. Discussion

This section reflects on the findings in light of the three RQs, emphasizing key patterns and providing representative examples to illustrate coverage, accuracy, and consistency differences across SKGs.

6.1. RQ1 – Category annotations across SKGs

The SKGs in our study differed significantly in annotation strategies and category granularity. PwC and OpenAlex assigned more categories per paper (17 and 12 on average, respectively), while OpenAIRE provided broader disciplinary coverage, and ORKG applied minimal but more conservative annotations.

A central finding was the trade-off between quantity and specificity. For instance, the paper *"Gemini 1.5"* (DOI: 10.48550/arxiv.2403.05530) received rich model-specific labels from PwC, while ORKG applied only "finding pre-trained large language model" and OpenAIRE relied on broad terms such as "computer" and "information sciences". Similarly, the paper *"MiniCPM"* (DOI: 10.48550/arxiv.2404.06395) was annotated with task-specific terms like "domain adaptation" in PwC, but only "generic" in ORKG.

In hybrid SKGs like PwC, taxonomy cleanliness emerged as a concern. For example, *"Generating Benchmarks for Factuality Evaluation"* (DOI: 10.48550/arxiv.2307.06908) was annotated with both "language modeling" and "language modelling", highlighting the absence of normalization. Duplicate entries like these complicate downstream semantic analysis.

Surface-level term matching occasionally led to significant misclassifications. The paper *"Pride and Prejudice: LLM Amplifies Self-Bias"* (DOI: 10.18653/v1/2024.acl-long.826) triggered legal and political science categories in OpenAlex due to title keywords like "prejudice", despite being an ML study.

Annotation coverage, specificity, and taxonomic coherence differ substantially across SKGs. While automated and hybrid systems offer breadth, they introduce term redundancy and thematic drift. Manual systems like ORKG provide focused but sparse coverage. Context-aware disambiguation and taxonomy standardization are needed to improve interoperability. Notably, annotation types differ substantially across SKGs. ORKG emphasizes problem/method/result triples, OpenAIRE and OpenAlex assign broad subject categories, while PwC offers task- and method-level keywords. Although ORKG also provides research field classifications, we excluded them as they overlap strongly with OpenAIRE and OpenAlex and would bias the comparison.

6.2. RQ2 – Accuracy compared to the gold-standard

To assess annotation relevance, we compared each SKG to a manually curated gold-standard. PwC achieved the highest recall but suffered from low precision (27%). ORKG, though sparse, demonstrated the highest precision (66%), followed by OpenAIRE and OpenAlex.

Many misclassifications stemmed from overgeneralization or superficial keyword matching. For example, OpenAlex labeled the paper *"Enhancing Text-Based Knowledge Graph Completion"* (DOI: 10.1016/j.knosys.2024.112155) with "paleontology" and "mechanical engineering"—labels likely derived from unrelated co-occurring terms. In contrast, PwC provided precise annotations like "graph embedding" and "contrastive learning".

Another case is *"OLMo"* (DOI: 10.18653/v1/2024.acl-long.841), where OpenAlex provided both accurate and noisy categories such as "topic modeling" and "meteorology". However, the presence of confidence values in OpenAlex allowed for assessing reliability—an advantage over other SKGs.

High annotation density does not guarantee high relevance. Systems like PwC maximize coverage but include considerable noise, while ORKG captures essential concepts at the expense of completeness. Confidence scoring and validation pipelines are key to improving accuracy.

6.3. RQ3 – Types of Annotation Inconsistencies

Our comparative analysis identified four primary types of annotation inconsistencies:

- **Coverage inconsistency:** Some SKGs failed to annotate otherwise well-covered papers. For instance, ORKG labeled the paper *"MiniCPM"* only as "generic", missing the technical depth captured by PwC and OpenAlex.
- **Incorrect assignment:** Irrelevant labels were particularly common in automated systems. OpenAlex annotated *"MiniCPM"* with "meteorology" and "geography", while the actual topic concerns model scaling strategies.

- **Granularity mismatch:** Labels ranged from very general (e.g., "computer science") in OpenAIRE to extremely specific (e.g., "contrastive learning") in PwC, complicating comparisons and integration.
- **Label noise and duplication:** PwC's community-driven labels sometimes included errors. The paper *"Phi-3 Technical Report"* (DOI: 10.48550/arxiv.2404.14219) was accurately annotated with terms like "attention mechanisms" but also included the nonsensical category "15 ways to contact how can I speak to someone at delta airlines", likely a mislabeling or spam entry.

These inconsistencies were tracked per paper and summarized across the dataset. Over 60% of all raw annotations were filtered out during gold-standard construction, underscoring the need for quality assurance. They mainly stem from taxonomy misalignment, inconsistent curation, and limited validation. A key ongoing research challenge is therefore how to ensure robust quality control, category normalization, and richer annotation metadata to improve trust in these systems.

7. Conclusion and Future Work

This study provides an in-depth comparison of categories (i.e., task and method annotations) across four prominent Scientific Knowledge Graphs (SKGs) (ORKG, OpenAlex, OpenAIRE, and PwC) on a shared set of 70 AI-related publications. By analyzing a manually curated dataset with parallel annotations from each SKG, we reveal substantial variation in category annotation coverage, granularity, and semantic alignment. Our dataset is limited in size, but to our knowledge, it is the first cross-SKG category annotation comparison study for SKG comparison.

Our findings suggest that PwC offers the most comprehensive and fine-grained annotations, likely due to its hybrid strategy combining automated extraction with manual validation. OpenAIRE, in contrast, employs a coarse-grained taxonomy that emphasizes domain-level categorization. ORKG exhibits high semantic precision where annotations are present, but its reliance on manual input results in limited coverage. OpenAlex provides broad topic coverage through automated classification, but occasionally assigns contextually inappropriate labels, likely due to literal keyword matching and limited disambiguation.

Although more work is needed to confirm our findings outside the pool of articles selected in our dataset, the discrepancies found highlight the challenges of aligning annotations across SKGs and point to broader research challenges in annotation design, vocabulary standardization, and semantic interoperability for consistent categorization and improved cross-graph metadata integration.

Future work will expand our dataset beyond 70 AI-related publications to include additional domains, enabling broader generalization. We plan to compute inter-SKG agreement metrics and apply clustering to uncover structural inconsistencies. We also aim to assess inter-annotator agreement for the gold labels to validate their reproducibility. Finally, we plan to explore schema mapping and ontology alignment strategies to reconcile differences in labeling schemes and abstraction levels.

Declaration on Generative AI

All research content and ideas are original by the authors. We acknowledge the use of ChatGPT for supervised grammar checks and minor paragraph rewording.

Acknowledgements

The authors would like to thank the EVERSE project (GA 101129744) under the European Union's Horizon Europe Programme (HORIZON-INFRA-2023-EOSC-01-02).

References

[1] A. A. Salatino, A. Mannocci, F. Osborne, Detection, analysis, and prediction of research topics with scientific knowledge graphs, in: Predicting the dynamics of research impact, Springer, 2021, pp. 225–252. URL: https://doi.org/10.1007/978-3-030-86668-6_11. doi:10.1007/978-3-030-86668-6_11.

[2] P. Manghi, A. Mannocci, F. Osborne, D. Sacharidis, A. Salatino, T. Vergoulis, New trends in scientific knowledge graphs and research impact assessment, Quantitative Science Studies 2 (2021) 1296–1300. URL: https://doi.org/10.1162/qss_e_00160. doi:10.1162/qss_e_00160. arXiv:https://direct.mit.edu/qss/article-pdf/2/4/1296/2007915/qss$_{e0}$0160.*pdf*.

[3] H. Shema, J. Bar-Ilan, M. Thelwall, Research blogs and the discussion of scholarly information, PLoS ONE 7 (2012) e35869. URL: https://doi.org/10.1371/journal.pone.0035869. doi:10.1371/journal.pone.0035869, epub 2012 May 11.

[4] M. Dodge, R. Kitchin, Codes of life: Identification codes and the machine-readable world, Environment and Planning D: Society and Space 23 (2005) 851–881. URL: https://doi.org/10.1068/d378t. doi:10.1068/d378t. arXiv:https://doi.org/10.1068/d378t.

[5] D. Dessí, F. Osborne, D. Reforgiato Recupero, D. Buscaldi, E. Motta, Cs-kg: A large-scale knowledge graph ofnbsp;research entities andnbsp;claims innbsp;computer science, in: The Semantic Web – ISWC 2022: 21st International Semantic Web Conference, Virtual Event, October 23–27, 2022, Proceedings, Springer-Verlag, Berlin, Heidelberg, 2022, p. 678–696. URL: https://doi.org/10.1007/978-3-031-19433-7_39. doi:10.1007/978-3-031-19433-7_39.

[6] J. Schickore, Scientific discovery, in: E. N. Zalta (Ed.), The Stanford Encyclopedia of Philosophy, winter 2022 ed., Metaphysics Research Lab, Stanford University, 2022. URL: https://plato.stanford.edu/entries/scientific-discovery/, first published March 6, 2014; substantive revision October 31, 2022.

[7] P. Langley, H. A. Simon, G. L. Bradshaw, J. M. Zytkow, Scientific discovery: computational explorations of the creative process, MIT Press, Cambridge, MA, USA, 1987.

[8] A. Jacobsen, R. de Miranda Azevedo, N. Juty, D. Batista, S. Coles, R. Cornet, M. Courtot, M. Crosas, M. Dumontier, C. T. Evelo, C. Goble, G. Guizzardi, K. K. Hansen, A. Hasnain, K. Hettne, J. Heringa, R. W. Hooft, M. Imming, R. Jeffery, Keith G. an Kaliyaperumal, M. G. Kersloot, C. R. Kirkpatrick, T. Kuhn, I. Labastida, B. Magagna, P. McQuilton, N. Meyers, A. Montesanti, M. van Reisen, P. Rocca-Serra, R. Pergl, S.-A. Sansone, L. O. B. da Silva Santos, J. Schneider, G. Strawn, M. Thompson, A. Waagmeester, T. Weigel, M. D. Wilkinson, E. L. Willighagen, P. Wittenburg, M. Roos, B. Mons, E. Schultes, Fair principles: Interpretations and implementation considerations, Data Intelligence 2 (2020) 10–29. URL: https://doi.org/10.1162/dint_r_00024. doi:10.1162/dint_r_00024. arXiv:https://direct.mit.edu/dint/article-pdf/2/1-2/10/1893430/dint$_{r0}$0024.*pdf*.

[9] A.-L. Lamprecht, L. Garcia, M. Kuzak, C. Martinez, R. Arcila, E. M. D. Pico, V. D. D. Angel, S. van de Sandt, J. Ison, P. A. Martinez, P. McQuilton, A. Valencia, J. Harrow, F. Psomopoulos, J. L. Gelpi, N. C. Hong, C. Goble, S. Capella-Gutierrez, Towards fair principles for research software, Data Science 3 (2020) 37–59. URL: https://doi.org/10.3233/DS-190026. doi:10.3233/DS-190026. arXiv:https://doi.org/10.3233/DS-190026.

[10] M. Barker, N. P. C. Hong, D. S. Katz, A.-L. Lamprecht, C. Martinez-Ortiz, F. Psomopoulos, J. Harrow, L. J. Castro, M. Gruenpeter, P. A. Martinez, T. Honeyman, Introducing the fair principles for research software, Scientific Data 9 (2022) 622. URL: https://doi.org/10.1038/s41597-022-01710-x. doi:10.1038/s41597-022-01710-x.

[11] D. Newman, K. Hagedorn, C. Chemudugunta, P. Smyth, Subject metadata enrichment using statistical topic models, in: Proceedings of the 7th ACM/IEEE-CS Joint Conference on Digital Libraries, JCDL '07, Association for Computing Machinery, New York, NY, USA, 2007, p. 366–375. URL: https://doi.org/10.1145/1255175.1255248. doi:10.1145/1255175.1255248.

[12] L. Tartar, The General Theory of Homogenization: A Personalized Introduction, volume 7 of *Lecture Notes of the Unione Matematica Italiana*, Springer Berlin Heidelberg, Berlin, Heidelberg, 2010. doi:10.1007/978-3-642-05195-1.

[13] J. Priem, H. Piwowar, R. Orr, OpenAlex: A fully-open index of scholarly works, authors, venues, institutions, and concepts, 2022. URL: http://arxiv.org/abs/2205.01833. doi:10.48550/arXiv.2205.01833, arXiv:2205.01833 [cs].

[14] N. Rettberg, B. Schmidt, Openaire - building a collaborative open access infrastructure for european researchers, LIBER Quarterly: The Journal of the Association of European Research Libraries 22 (2012) 160–175. URL: https://liberquarterly.eu/article/view/10641. doi:10.18352/lq.8110.

[15] N. Rettberg, B. Schmidt, Openaire: Supporting a european open access mandate, College Research Libraries News 76 (2015) 306–310. URL: https://crln.acrl.org/index.php/crlnews/article/view/9326. doi:10.5860/crln.76.6.9326.

[16] P. Manghi, A. Bardi, C. Atzori, M. Baglioni, N. Manola, J. Schirrwagen, P. Principe, The openaire research graph data model, 2019. URL: https://doi.org/10.5281/zenodo.2643199. doi:10.5281/zenodo.2643199.

[17] P. Manghi, C. Atzori, A. Bardi, M. Baglioni, J. Schirrwagen, H. Dimitropoulos, S. La Bruzzo, I. Foufoulas, A. Mannocci, M. Horst, A. Czerniak, K. Iatropoulou, A. Kokogiannaki, M. De Bonis, M. Artini, A. Lempesis, A. Ioannidis, N. Manola, P. Principe, T. Vergoulis, S. Chatzopoulos, D. Pierrakos, Openaire graph dump, 2022. URL: https://doi.org/10.5281/zenodo.7488618. doi:10.5281/zenodo.7488618.

[18] M. Y. Jaradeh, A. Oelen, K. E. Farfar, M. Prinz, J. D'Souza, G. Kismihók, M. Stocker, S. Auer, Open research knowledge graph: Next generation infrastructure for semantic scholarly knowledge, in: Proceedings of the 10th International Conference on Knowledge Capture, K-CAP '19, Association for Computing Machinery, New York, NY, USA, 2019, p. 243–246. URL: https://doi.org/10.1145/3360901.3364435. doi:10.1145/3360901.3364435.

[19] D. Dessì, F. Osborne, D. Reforgiato Recupero, D. Buscaldi, E. Motta, H. Sack, Ai-kg: An automatically generated knowledge graph of artificial intelligence, in: The Semantic Web – ISWC 2020: 19th International Semantic Web Conference, Athens, Greece, November 2–6, 2020, Proceedings, Part II, Springer-Verlag, Berlin, Heidelberg, 2020, p. 127–143. URL: https://doi.org/10.1007/978-3-030-62466-8_9. doi:10.1007/978-3-030-62466-8_9.

[20] G. Hendricks, D. Tkaczyk, J. Lin, P. Feeney, Crossref: The sustainable source of community-owned scholarly metadata, Quantitative Science Studies 1 (2020) 414–427. URL: https://doi.org/10.1162/qss_a_00022. doi:10.1162/qss_a_00022. arXiv:https://direct.mit.edu/qss/article-pdf/1/1/414/1760913/qss$_{a0}$0022.pdf.

[21] J.-P. Vergne, T. Wry, Categorizing categorization research: Review, integration, and future directions, Journal of Management Studies 51 (2014) 56–94. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/joms.12044. doi:https://doi.org/10.1111/joms.12044. arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1111/joms.12044.

[22] E. Rosch, Principles of categorization, in: E. Rosch, B. B. Lloyd (Eds.), Cognition and Categorization, Lawrence Elbaum Associates, 1978, pp. 27–48.

[23] R. Guha, R. McCool, E. Miller, Semantic search, in: Proceedings of the 12th International Conference on World Wide Web, WWW '03, Association for Computing Machinery, New York, NY, USA, 2003, p. 700–709. URL: https://doi.org/10.1145/775152.775250. doi:10.1145/775152.775250.

[24] G. Shani, A. Gunawardana, Evaluating recommendation systems, in: Recommender Systems Handbook, 2011. URL: https://api.semanticscholar.org/CorpusID:435521.

[25] R. Dattakumar, R. Jagadeesh, A review of literature on benchmarking, Benchmarking: An International Journal 10 (2003) 176–209. URL: https://doi.org/10.1108/14635770310477744. doi:10.1108/14635770310477744. arXiv:https://www.emerald.com/bij/article-pdf/10/3/176/140450/14635770310477744.pdf.

[26] T. A. Trikalinos, G. Salanti, E. Zintzaras, J. P. A. Ioannidis, Meta-analysis methods, Advances in Genetics 60 (2008) 311–334. doi:10.1016/S0065-2660(07)00413-0.

[27] J. T. Ciuciu-Kiss, D. Garijo, Assessing the overlap of science knowledge graphs: A quantitative analysis, in: Natural Scientific Language Processing and Research Knowledge Graphs: First International Workshop, NSLP 2024, Hersonissos, Crete, Greece, May 27, 2024, Proceedings, Springer-Verlag, Berlin, Heidelberg, 2024, p. 171–185. URL: https://doi.org/10.1007/978-3-031-65794-8_11.

doi:10.1007/978-3-031-65794-8_11.

[28] C. Goutte, E. Gaussier, A probabilistic interpretation of precision, recall and f-score, with implication for evaluation, in: Proceedings of the 27th European Conference on Advances in Information Retrieval Research, ECIR'05, Springer-Verlag, Berlin, Heidelberg, 2005, p. 345–359. URL: https://doi.org/10.1007/978-3-540-31865-1_25. doi:10.1007/978-3-540-31865-1_25.

[29] M. Altman, P. N. Cohen, The scholarly knowledge ecosystem: Challenges and opportunities for the field of information, Frontiers in Research Metrics and Analytics 6 (2021) 751553. URL: https://doi.org/10.3389/frma.2021.751553. doi:10.3389/frma.2021.751553.

[30] K. Wang, Z. Shen, C. Huang, C.-H. Wu, Y. Dong, A. Kanakia, Microsoft academic graph: When experts are not enough, Quantitative Science Studies 1 (2020) 396–413. URL: https://doi.org/10.1162/qss_a_00021. doi:10.1162/qss_a_00021. arXiv:https://direct.mit.edu/qss/article-pdf/1/1/396/1760880/qss$_{a0}$0021.*pdf*.

[31] J. T. Ciuciu-Kiss, D. Garijo, A study of the categories used in 'papers with code', in: ESWC 2025 Workshops and Tutorials Joint Proceedings, volume 3977 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2025. URL: https://ceur-ws.org/Vol-3977/NSLP-03.pdf.

[32] OECD, Frascati Manual 2015: Guidelines for Collecting and Reporting Data on Research and Experimental Development, OECD Publishing, Paris, 2015. URL: https://doi.org/10.1787/9789264239012-en. doi:10.1787/9789264239012-en.

[33] N. Gialitsis, S. Kotitsas, H. Papageorgiou, Scinobo: A hierarchical multi-label classifier of scientific publications, in: Companion Proceedings of the Web Conference 2022, WWW '22, Association for Computing Machinery, New York, NY, USA, 2022, p. 800–809. URL: https://doi.org/10.1145/3487553.3524677. doi:10.1145/3487553.3524677.

[34] N. Coulter, Acm's computing classification system reflects changing times, Commun. ACM 40 (1997) 111–112. URL: https://doi.org/10.1145/265563.265579. doi:10.1145/265563.265579.

[35] B. Rous, Major update to acm's computing classification system, Commun. ACM 55 (2012) 12. URL: https://doi.org/10.1145/2366316.2366320. doi:10.1145/2366316.2366320.

[36] D. Vrandečić, M. Krötzsch, Wikidata: a free collaborative knowledgebase, Commun. ACM 57 (2014) 78–85. URL: https://doi.org/10.1145/2629489. doi:10.1145/2629489.

[37] C. E. Lipscomb, Medical subject headings (mesh), Bull Med Libr Assoc. (2000). URL: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=35238, 88(3): 265–266.

[38] A. Gopan, Z. Kobti, Become: A framework for node classification in social graphs, Procedia Comput. Sci. 251 (2024) 208–215. URL: https://doi.org/10.1016/j.procs.2024.11.191. doi:10.1016/j.procs.2024.11.191.

[39] H. Santos, P. Pinheiro, J. P. McCusker, S. M. Rashid, D. L. McGuinness, Scikg: Tutorial on building scientific knowledge graphs from data, data dictionaries, and codebooks, in: ESWC 2023 Workshops and Tutorials Joint Proceedings, Hersonissos, Greece, 2023. URL: https://tetherless-world.github.io/scikg-eswc-2023/, tutorial co-located with the 20th Extended Semantic Web Conference.

[40] P. Ginsparg, Arxiv at 20, Nature 476 (2011) 145–147. URL: https://doi.org/10.1038/476145a. doi:10.1038/476145a.

[41] J. T. Ciuciu-Kiss, Parallel skg annotations and gold-standard for 70 ai papers, 2025. URL: https://doi.org/10.5281/zenodo.16422144. doi:10.5281/zenodo.16422144.

[42] J. T. Ciuciu-Kiss, kuefmz/skg_metadata_comparative_analysis: v1.0, 2025. URL: https://doi.org/10.5281/zenodo.16422339. doi:10.5281/zenodo.16422339.