Assessing the Overlap of Science Knowledge Graphs: A Quantitative Analysis

Jenifer Tabita Ciuciu-Kiss[1][0000−0002−3170−6730] and
Daniel Garijo[1][0000−0003−0454−7145]

[1]Ontology Engineering Group, Universidad Politécnica de Madrid
jenifer.ciuciu-kiss@alumnos.upm.es, daniel.garijo@upm.es

Abstract. Science Knowledge Graphs (SKGs) have emerged as a means to represent and capture research outputs (papers, datasets, software, etc.) and their relationships in a machine-readable manner. However, different SKGs use different taxonomies, making it challenging to understand their overlaps, gaps and differences. In this paper, we propose a quantitative bottom-up analysis to assess the overlap between two SKGs, based on the type annotations of their instances. We implement our methodology by assessing the category overlap of 100,000 publications present both in OpenAlex and OpenAIRE. As a result, our approach produces an alignment of 71 categories and discusses the level of agreement between both KGs when annotating research artefacts.

Keywords: Scientific Knowledge Graph · Knowledge Graph · Taxonomy · Alignment

1 Introduction

As the volume of scientific literature increases, the need for scalable and efficient systems to navigate this extensive information becomes crucial. Science Knowledge Graphs (SKGs) [11] have emerged as a key tool for representing research entities (publications, people, organizations, datasets, software, etc.) their relationships and metadata in a machine-readable manner.

SKGs such as OpenAIRE [1] [23,24,19,18] and OpenAlex [2] [21] contain millions of entities describing publications and research outputs. One of the main challenges when using SKGs is identifying and resolving overlaps in categorization, which is critical for querying them consistently and reliably. This challenge is complex due to the diversity and volume of data within these KGs, requiring advanced methodologies for effective detection and resolution of overlaps. Understanding these overlaps and disagreements is essential for insights into the structure of scientific knowledge, highlighting patterns that are not immediately apparent due to data scale and diversity.

This paper proposes a quantitative bottom-up methodology to assess the overlap of SKGs categories, based on the annotations made on their instances.

[1] https://www.openaire.eu/
[2] https://openalex.org/

More specifically we aim to explore the overlap of the taxonomies used in scientific literature [22]. Our contributions include:

1. A novel methodology designed to explore the overlap between SKGs.
2. An implementation of the methodology, based on two SKGs to validate its effectiveness, resulting in 71 new aligned categories within these graphs.
3. An initial exploration study of the intersection of two SKGs, based on 100,000 papers that are jointly described in both of them.

As a proof of concept, we have applied our methodology to a subset of OpenAlex and OpenAIRE SKGs, in the AI domain. We chose OpenAlex for its extensive global database of academic research, and OpenAIRE for its European focus and its integration from heterogeneous data sources. This combination offers a comprehensive view of academic communication, providing a comprehensive dataset for our methodology.

The remainder of the paper is structured as follows. Section 2 describes our methodology, while Section 3 explains how we implemented our methodology by assessing OpenAlex and OpenAIRE. Section 4 discusses the results of our categorization analysis on both SKGs, Section 5 introduces relevant efforts to map taxonomies and ontologies, and Section 6 concludes the paper.

2 A Methodology for Assessing SKG Overlap

We propose a sequential process that evaluates the degree of overlap in SKGs and aims to develop a suite of potential mappings across various KGs, informed by the insights gained from the overlap assessment.

Our methodology is divided into two phases, detailed in Fig. 1. The initial phase (on top of Fig. 1) includes data collection, alignment of the different KG instances and preprocessing. This phase may be repeated and expanded as necessary to refine the dataset to an acceptable size and quality. After completing the dataset preparation, the category alignment phase starts(bottom of Fig. 1). This phase systematically proposes, evaluates, and selects the best mappings between categories based on existing paper annotations, producing a validated set of final mappings of overlapping categories.

2.1 Data preparation

To date, there is no available open dataset tailored for the quantitative analysis of overlaps within SKGs that considers associated papers and their categorizations. This gap requires the creation of a dataset for conducting a bottom-up quantitative analysis. The data preparation phase may be challenging due to 1) the size of SKGs and 2) the diverse structures and access methods of SKGs, which range from complete data dumps available on platforms like Zenodo[17] to those accessible only via REST APIs or SPARQL [5] queries. We detail the steps for data preparation below.

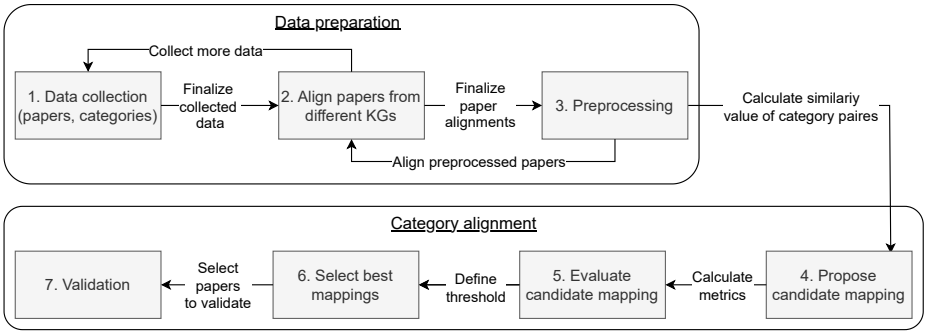

Fig. 1. Steps of the methodology for assessing the overlap of SKGs

Data collection encloses the aggregation of data from various KGs. Data must contain references to the research papers under analysis and their associated categories. Distinct unique identifiers (typically DOIs) may be used for publications, ensuring that these identifiers are consistent across all targeted KGs for data acquisition. Preliminary examination of the collection structures is compulsory to ensure the integrity and quality of the data obtained.

Align papers from different KGs using the gathered data and the chosen unique identifier (e.g. paper DOI, title). Complete alignment may not be feasible due to the heterogeneous nature of data across SKGs, yet a substantial portion of the data should be possible to align, given the overlap in the data sources.

Preprocessing of the gathered data entails multiple steps. First, we eliminate noise from category data, in order to avoid potential variations in character encodings and the presence of inconsistencies in category names and titles, such as inconsistent capitalization and the use of dashes. Cleaning the textual data enhances the alignment quality between the papers and their corresponding categories.

Following text cleansing, we remove underrepresented categories to streamline the later stages of the analysis by reducing its complexity. A category is considered underrepresented when the count of associated papers falls below a predetermined threshold. The value of this threshold is flexible and should consider the size of the dataset, the overall count of categories, and how papers are distributed among these categories.

2.2 Category alignment

This phase consists of three steps that generate an initial set of candidate mappings between categories:

Propose candidate mapping. To identify probable candidate mappings with significant relevance, a similarity model must be used to exclude mappings with semantic similarity below a designated threshold.

Text similarity may be computed using existing embedding techniques. For example, in our work we propose the `en_core_web_md` model in spaCy,[3] which employs GloVe word embeddings [12,20]. The similarity between category strings is determined by the cosine similarity of their vector representations:

$$\text{similarity}(\mathbf{A}, \mathbf{B}) = \frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\|\|\mathbf{B}\|}$$

where $\mathbf{A}$ and $\mathbf{B}$ are the vector representations of the category strings. The similarity ranges from -1 (opposite meaning) to 1 (identical meaning).

To maximize the number of candidate mappings, a similarity greater than 0 may be considered.

Evaluate candidate mappings using the following metrics:

- Number of papers belonging to the first category ($Support1$)
- Number of papers belonging to the second category ($Support2$)
- Number of papers belonging to both categories ($Intersection$)
- The ratio of $Intersection$ over $Union$ also called as IoU [7,15] ($Agreement$)

The $Agreement$ (i.e., $Intersection/Union$) is calculated using the $Intersection$ over $Union$ (where $Union = Support1 + Support2 - Intersection$) of papers belonging to a certain category in their respective SKG.

Select the best mapping based on the *Agreement*. Category mappings are classified into three types: exact, related, and unrelated. Exact matches, where the similarity score equals 1, represent identical categories across all KGs. Related matches exceed the established threshold of *Agreement*, suggesting a strong correspondence. Unrelated categories fall at or below the threshold, indicating a weaker or no relation. The threshold for *Agreement* is adjustable and upon various factors, notably the forthcoming manual validation. Although manual validation of all mappings would be ideal, resource limitations require setting a pragmatic threshold to minimize manual effort.

Validate candidate mappings by having domain experts manually review papers classified into the aligned categories.

3 Initial SKG Overlap Assessment: OpenAIRE and OpenAlex

This work uses two KGs as primary data sources: OpenAIRE (Open Access Infrastructure for Research in Europe [23,24,19,18]) and OpenAlex [21].

[3] `https://spacy.io/models/en#en_core_web_md`

OpenAIRE is a European Open Science infrastructure that aims to promote open scholarship and substantially improve the discoverability and reusability of research publications and data. The OpenAIRE KG integrates data from a wide range of research outputs, including publications, datasets, projects, and research organizations, facilitating a more interconnected and comprehensive understanding of European scientific research. The OpenAIRE API [4] allows access to a vast collection of scientific publications, datasets, projects, and funding information. In this work, we used the Search API [5] [6] to collect data on scientific publications and their categories, facilitating the quantitative analysis of categorization overlaps in KGs.

OpenAIRE is supported by the European Commission and various European entities. It aggregates data from a multitude of sources to build its comprehensive knowledge graph, including repositories, archives, and journals across Europe. As part of the European Open Science Cloud, OpenAIRE benefits from consistent updates and enhancements, ensuring its relevance and utility in the research community. SCINOBO[7] and other science taxonomies are used to classify the results.

OpenAlex is an open catalogue of the global research system, offering detailed information on academic papers, authors, institutions, etc. The platform indexes millions of research outputs, providing a rich dataset for analysis in various academic fields. The OpenAlex API [8] [9] provides access to their extensive dataset. This API enables querying and retrieving detailed information about academic works, supporting a wide range of scholarly analyses. In this work, the OpenAlex API was used to gather information on scientific papers and their categorization.

OpenAlex offers a dynamic dataset with weekly updates, incorporating the latest data from various public and proprietary sources, including academic publishers, preprint servers, institutional repositories, and databases. It aims to index the entirety of the scholarly record, offering an open, comprehensive view of global research output. By ingesting data from various sources, OpenAlex ensures a rich and varied dataset, which includes information on publications, authors, institutions, and citation metrics. Furthermore, OpenAlex aligns its dataset with Wikidata [25] categories. As a successor to the Microsoft Academic Graph [27], OpenAlex aims to provide a comprehensive, open resource for academia.

OpenAlex employs a taxonomy with 65,000 categories, as detailed in its README [10]. Further documentation elaborates on the classification model used

[4] https://graph.openaire.eu/docs/apis/home/
[5] https://graph.openaire.eu/docs/apis/search-api/
[6] https://api.openaire.eu/search/publications
[7] https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10192702/
[8] https://docs.openalex.org/
[9] https://api.openalex.org/works
[10] https://github.com/ourresearch/openalex-concept-tagging

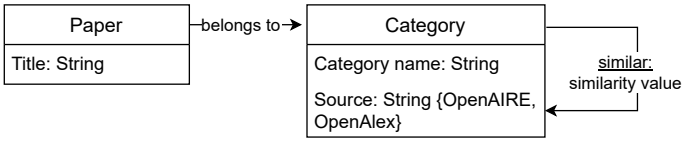

Fig. 2. Paper-category schema representation

by OpenAlex, providing comprehensive details. It is noted that the model attains a precision of 60%, an important consideration during our alignment efforts.

3.1 Paper-Category schema representation

Figure 2 represents the schema we used for storing the collected data, providing a structural basis for querying and retrieving information during the analysis. It outlines the associations between scientific papers and their respective categories within the database. Nodes marked as *Paper* are attributed with a *title*, while *Category* nodes encapsulate both the *category name* and the *source* attribute, which identifies whether the category is derived from the *OpenAIRE* or *OpenAlex* SKGs. The relational attribute *belongs to* connects papers to their relevant categories, and a second relational attribute *similar* binds category nodes together, equipped with a *similarity value* to express the level of similarity between category pairs. The Neo4j graph database[11] [26] was chosen store SKG data following our chosen representation.

3.2 SKG Overlap Analysis: Data preparation

The alignment of the collected papers from the KGs was conducted using their titles. Whenever multiple papers from the same source shared the same title, leading to potential matching conflicts, such occurrences were disregarded to maintain alignment precision.

A total of 108,555 papers were aligned, that were available in both KGs. On average, OpenAIRE assigns approximately 21 categories to each paper, whereas OpenAlex assigns around 18. Consequently, the dataset from OpenAIRE encompasses a larger number of categories. OpenAlex also assigns a confidence value to each of the assigned categories (not taken into account in this analysis).

The preprocessing primarily targeted the categories to ensure the text was clean for category alignment. The steps included removing Unicode characters, removing punctuation, and converting all text to lowercase. Through experimentation, it was determined that more extensive preprocessing did not significantly improve the results.

A threshold was defined for the minimal number of papers a category must be represented by to be included in the analysis. This threshold was set after evaluating the initial distribution of categories in OpenAIRE and OpenAlex and

[11] `https://neo4j.com/`

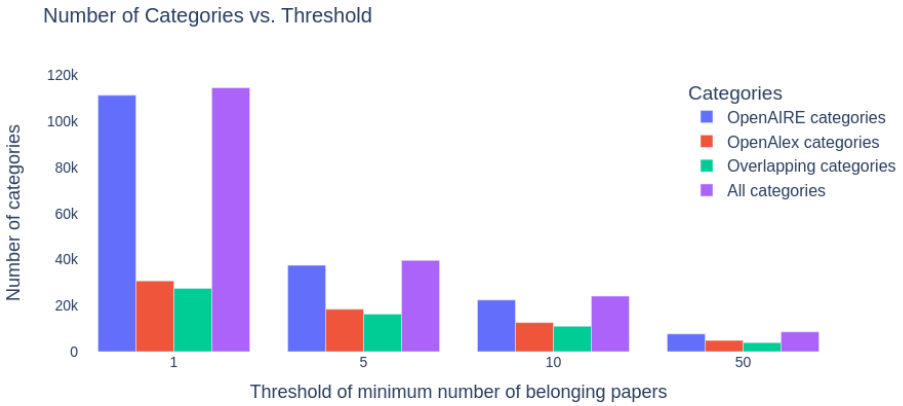

Fig. 3. The number of categories based on the threshold applied to the number of papers representing each category

testing various threshold levels. As depicted in Fig. 3, a threshold of 1 had no effect, while a threshold of 5 made a noticeable difference. A threshold of 10 was found to effectively refine the categories without excessively reducing their number. Consequently, we set the threshold at 10, thereby finalizing our dataset for further analysis stages.

3.3 SKG Overlap Analysis: Category Alignment

The category alignment phase involves proposing potential mappings between categories based on their similarity, evaluating these mappings against predefined metrics, and then refining the selection based on an agreement threshold to ensure only the most relevant mappings are considered.

While proposing the candidate mappings, a similarity threshold of 0.0 was selected, removing the mappings of opposite categories. This approach was chosen to ensure no potentially significant mappings were excluded at this early stage.

This approach resulted in 509,034 potential mappings. We then assessed these mappings using the metrics outlined in Section 2.2. To determine an appropriate threshold for the *Agreement* metric, we evaluated how the number of related matches varied with different threshold settings. As illustrated in Fig. 4, increasing the threshold reduces the count of mappings deemed related, with a notable decrease between 0.1 and 0.2. A plateau appears to occur between 0.4 and 0.5, beyond which the number of related matches dwindles to near zero, especially at a threshold of 0.9. Hence, we established an *Agreement* threshold of 0.5 (without imposing a limit on the 'Similarity' value), which identified 72 mappings as related.

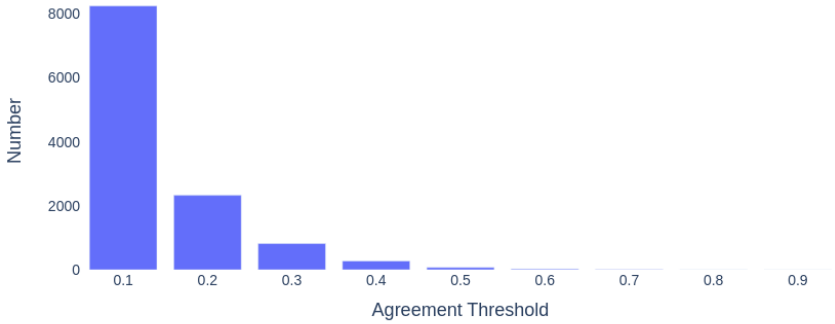

Fig. 4. Number of related matches based on the agreement threshold

We collected the papers that were associated with both of the matched categories for manual review. Following our validation process, the candidate mappings were inspected individually by two researchers, discussing the results until an agreement was reached.

4 Results

Fig. 5 illustrates the evolution of the data throughout our analysis, indicating how each step of the methodology impacts data quantity and analysis detail. A corpus of 176,200 papers from OpenAlex was collected, from which 108,555 were found to correspond with the OpenAIRE database entries. Following the paper alignment, an analysis of the categories was conducted. We defined a threshold, requiring a category to contain a minimum of 10 papers for consideration in the mapping process. This criterion resulted in a total of 12,642 categories from OpenAlex and 22,462 categories from OpenAIRE. There was considerable overlap among the categories, leading to the creation of 509,034 potential category mappings. Upon calculating the metrics described in Section 2.2 for each mapping, we categorized the mappings into three types: exact, related, and unrelated. Detailed in the bottom right of Fig. 5, under 0.1% of these mappings were considered related (counting 72). Meanwhile, 2.34% were identified as exact matches, signifying categories with a one-to-one correspondence across both KGs. The rest, 97.65% were classified as unrelated matches, which fall outside the relevant domain of this analysis.

In summary, there are 12,642 categories in OpenAlex, with 11,920 identified as exact matches, accounting for 94.23%. These categories also exist in OpenAIRE, directing our focus to matching the remaining 722 OpenAlex categories. We found 72 related categories, approximately 10% of the OpenAlex categories requiring matches.

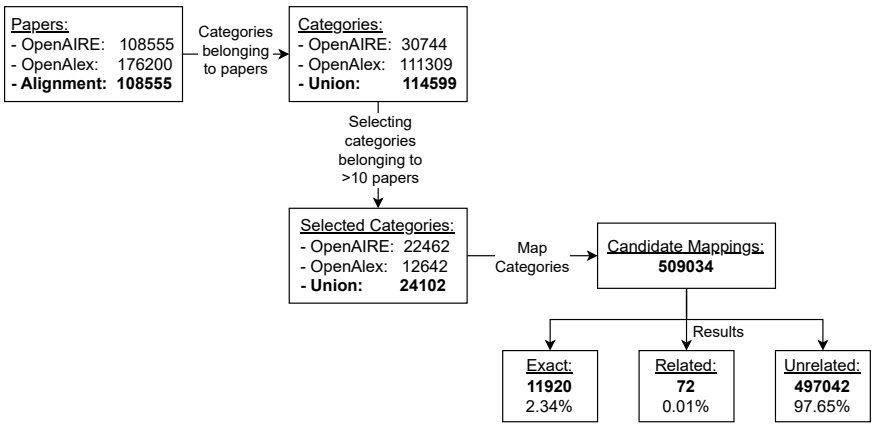

Fig. 5. Data flow from the initial collection to final category mapping analysis results

We manually examined these 72 mappings and observed that the labels do not always align (15 mappings). Upon further analysis of the overlapping papers for each SKG, we determined that the mappings remain plausible, although one of the label names may be incorrect. For instance, both 'lasso (statistics)' and 'lasso (programming language)' refer to papers related to lasso statistics. However, the label 'lasso (programming language)' in OpenAlex is used incorrectly for the reviewed papers, which all correspond to lasso statistics. We also identified 1 example of correlation, but not causation between categories: 'melanism' from OpenAlex and 'peppered moth' from OpenAIRE both refer to a collection of papers studying the melanism in a concrete species of moth. Therefore, of the initial 72 mappings we proposed, 14 were identified with misaligned labels referring to the same papers and 1 exhibited correlation without causation. This underscores the importance of our methodology in identifying candidate mappings that require expert validation.

Further, we investigated the relationship between the *Similarity* and the *Agreement* metrics, with findings illustrated in Fig. 6. Interestingly, there appears to be no significant correlation between these two metrics, indicating a high level of disagreement when annotating research publications with concepts.

Another interesting takeaway is the distribution of the *Agreement* values of the exact matches, shown in Fig. 7. Despite the presence of identical categories across both KGs, there is a strong disagreement among the papers that belong to these categories (i.e., the same papers have different category annotations). Additional work is needed to assess if the confidence values assigned in OpenAlex categories affect these findings (e.g. removing low-confidence categories).

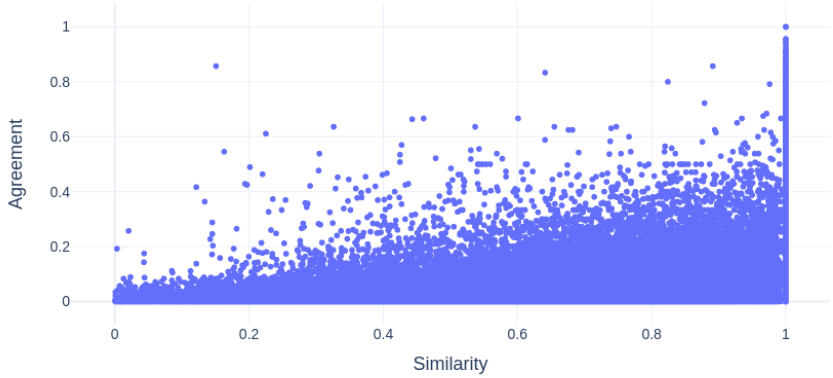

Fig. 6. Correlation of Similarity and Agreement

The scripts [2] [12] used to carry out our methodology and analysis are available online under the MIT license. The results of the matches can be found in Zenodo [3]. [13]

In summary, our analysis yielded 3 main findings. First, we identified 71 (72 mappings proposed - 1 mismatch: 'melanism' from OpenAlex and 'peppered moth' from OpenAIRE) newly aligned categories across two SKGs. Secondly, we observed a notable lack of correlation between the 'Similarity' and 'Agreement' metrics. Finally, our research revealed that the presence of identical categories in both SKGs does not guarantee agreement on category assignments.

5 Related work

Ontology alignment [6] is a subset of KG alignment and involves matching concepts, relationships, and instances across different ontologies to enable knowledge integration, facilitating a unified view of knowledge across various domains. This section explores significant contributions to the field of KG and ontology alignment.

5.1 KG alignment based on embeddings

Several methods leverage embedding techniques to enhance the interoperability and integration of heterogeneous knowledge bases.

ITransE [29] is an approach for embedding knowledge from various KGs, applicable to cross-lingual KG alignment. The method builds on TransE [1], learning embeddings for entities and relations, and then mapping these embeddings

[12] `https://github.com/kuefmz/define_taxonomy`
[13] `https://zenodo.org/records/10974512`

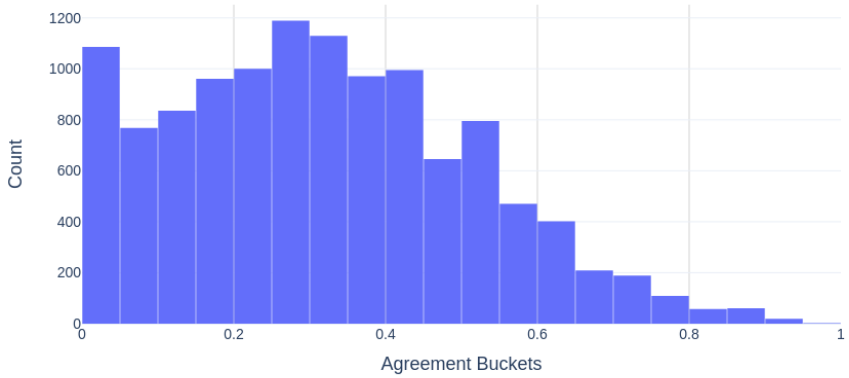

Fig. 7. Distribution of Agreement values of the exact mappings (Similarity = 1)

to a shared space using predefined entity alignments. ITransE updates these embeddings through an iterative process as it discovers new entity alignments, requiring uniform relations across all involved KGs for alignment execution.

JE [9] learns embeddings for multiple KGs in a single vector space to align entities. The method employs initial entity alignments to associate two KGs and modifies the TransE model to include an entity alignment loss in its loss function, allowing the alignment process.

In [8] the authors present a KG embedding method for entity alignment, a crucial task for integrating knowledge from various KGs. Their work provides a comprehensive meta-level analysis of popular embedding methods, identifying statistically significant correlations between different embedding methods and meta-features extracted from KGs. This rigorous analysis offers a unique perspective on the effectiveness and efficiency of various embedding methods in real-world KG settings, addressing critical questions about the assumptions and sensitivities of these methods to different KG characteristics.

The publications presented above focus on using embedding methods for aligning KGs, utilizing models to project entities and relations into a unified vector space. These methods often depend on pre-aligned data. In contrast, we introduce a quantitative approach that focuses on a bottom-up analysis to assess overlaps directly based on the categorization of scientific literature within the KGs. However, we base our work on these techniques to calculate similarity between categories.

5.2 KG alignment based on machine learning

SelfLinKG [16] introduces an approach for enhancing the connectivity and utility of scientific KGs. This work leverages self-supervised learning techniques to

identify and establish links between disparate KGs, facilitating a more integrated and comprehensive representation of scientific knowledge. By employing self-supervision, the authors demonstrate significant improvements in the accuracy and efficiency of KG linking, offering valuable insights into the potential of machine learning in KG integration.

Cross-lingual KG alignment [28] presents method for aligning KGs across different languages using graph convolutional networks (GCNs). This work [28] focuses on the challenge of matching entities in multilingual KGs, an essential task for enhancing cross-lingual interoperability and integration of information. Their approach involves training GCNs to embed entities from different languages into a unified vector space, where alignment is determined based on the proximity of entity embeddings. This method leverages both structural and attribute information of entities, aiming to improve the accuracy and efficiency of cross-lingual KG alignment.

In [13], the authors explore the application of KGs and attention mechanisms in bag-level relation extraction, providing a quantitative analysis of their impact. This study contributes a new dataset and proposes a framework to evaluate how KGs and attention mechanisms affect the extraction process, offering insights that could inform the development of more effective relation extraction methods.

All these methods leverage the power of machine learning to identify patterns and establish connections within and across KGs, contributing to a richer, more interconnected web of knowledge. However, they often require substantial training data and can sometimes obscure the interpretability of the alignment process. These methods adapt and evolve through learning patterns in the data, which, while effective, can introduce complexities in understanding why specific alignments are suggested. Our quantitative approach sidesteps these challenges by employing a straightforward, bottom-up analysis that directly assesses the categorizations of papers in KGs. Our method offers a clear, logical pathway to understanding alignments, grounded in the inherent structure and content of the KGs themselves, rather than inferred patterns from machine learning models.

6 Conclusions and Future Work

This work proposed a quantitative bottom-up analysis to assess the overlaps among different KGs, using OpenAIRE and OpenAlex as primary data sources. The findings underscore a notable divergence in the categorization and alignment of KGs despite their reliance on similar resources and methodologies. Surprisingly, even when these KGs draw upon comparable datasets and aim to represent similar domains, the divergence in their categorization frameworks is substantial.

This study successfully proposed a set of mappings that are likely to be related, offering a new perspective on the interconnectedness of these KGs. However, it is imperative to note that the proposed mappings are preliminary and require further validation by domain experts to ensure their accuracy and relevance. This validation is crucial for ensuring the mappings' utility in enhancing the interoperability and integration of KGs in the realm of scientific research.

The future direction of this research involves expanding the scope to complete the analysis of the entire OpenAlex and OpenAIRE KGs and expand to other SKGs (e.g. ORKG [14], AI-KG [4], Crossref [10]), thereby enriching the dataset and enhancing the robustness of the findings. Furthermore, we plan to enhance our experiments by employing various embeddings to eliminate biases and ascertain the similarity between terms. An important aspect for the future is to consider additional data available in the KGs. OpenAlex provides the confidence values for categories, which we did not incorporate in our work. By integrating more KGs and more data from the KGs, a more comprehensive understanding of the overlaps and divergences across different knowledge domains may be achieved. We intend to broaden our analysis by incorporating an inter-annotator agreement metric, which will serve not only as an additional measure but also as a tool for validation. Furthermore, we plan to enhance our experiments by employing various embeddings to ascertain the similarity between terms.

Moreover, our goal is to delve deeper into AI-related papers, extracting and analyzing their categorizations to propose a refined set of mappings. This effort will involve a systematic collection and analysis of AI research outputs across various KGs, followed by the application of advanced alignment and mapping techniques. The ultimate goal is to construct a more interconnected and semantically rich network of KGs, facilitating a more integrated and accessible repository of scientific knowledge.

Acknowledgements

This work is supported by the Madrid Government (Comunidad de Madrid - Spain) under the Multiannual Agreement with Universidad Politécnica de Madrid in the line Support for R&D projects for Beatriz Galindo researchers, in the context of the VPRICIT, and through the call Research Grants for Young Investigators from Universidad Politécnica de Madrid.

References

1. Bordes, A., Usunier, N., Garcia-Duran, A., Weston, J., Yakhnenko, O.: Translating embeddings for modeling multi-relational data. In: Burges, C., Bottou, L., Welling, M., Ghahramani, Z., Weinberger, K. (eds.) Advances in Neural Information Processing Systems. vol. 26. Curran Associates, Inc. (2013), https://proceedings.neurips.cc/paper_files/paper/2013/file/1cecc7a77928ca8133fa24680a88d2f9-Paper.pdf
2. Ciuciu-Kiss, J.T.: Scripts used to define taxonomy in the al/ml domain. kuefmz/define_taxonomy: v0.2 (Apr 2024). https://doi.org/10.5281/zenodo.10987999
3. Ciuciu-Kiss, J.T., Garijo, D.: Assessing the Overlap of Science Knowledge Graphs: A Quantitative Analysis — exact and related matches (Apr 2024). https://doi.org/10.5281/zenodo.10974512
4. Dessì, D., Osborne, F., Reforgiato Recupero, D., Buscaldi, D., Motta, E., Sack, H.: 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 19. pp. 127–143. Springer (2020)
5. DuCharme, B.: Learning SPARQL: querying and updating with SPARQL 1.1. " O'Reilly Media, Inc." (2013)
6. Euzenat, J., Shvaiko, P., et al.: Ontology matching, vol. 18. Springer (2007)
7. Everingham, M.: The pascal visual object classes challenge 2007. In: http://www. pascal-network. org/challenges/VOC/voc2007/workshop/index. html (2009)
8. Fanourakis, N., Efthymiou, V., Kotzinos, D., Christophides, V.: Knowledge graph embedding methods for entity alignment: experimental review. Data Mining and Knowledge Discovery **37**(5), 2070–2137 (2023)
9. Hao, Y., Zhang, Y., He, S., Liu, K., Zhao, J.: A joint embedding method for entity alignment of knowledge bases. In: Knowledge Graph and Semantic Computing: Semantic, Knowledge, and Linked Big Data: First China Conference, CCKS 2016, Beijing, China, September 19-22, 2016, Revised Selected Papers 1. pp. 3–14. Springer (2016)
10. Hendricks, G., Tkaczyk, D., Lin, J., Feeney, P.: Crossref: The sustainable source of community-owned scholarly metadata. Quantitative Science Studies **1**(1), 414–427 (2020)
11. Hogan, A., Blomqvist, E., Cochez, M., d'Amato, C., Melo, G.D., Gutierrez, C., Kirrane, S., Gayo, J.E.L., Navigli, R., Neumaier, S., et al.: Knowledge graphs. ACM Computing Surveys (Csur) **54**(4), 1–37 (2021)
12. Honnibal, M., Montani, I.: spacy 2: Natural language understanding with bloom embeddings, convolutional neural networks and incremental parsing. `https://spacy.io/` (2017)
13. Hu, Z., Cao, Y., Huang, L., Chua, T.S.: How knowledge graph and attention help? a quantitative analysis into bag-level relation extraction. arXiv preprint arXiv:2107.12064 (2021)
14. Jaradeh, M.Y., Oelen, A., Prinz, M., Stocker, M., Auer, S.: Open research knowledge graph: a system walkthrough. In: Digital Libraries for Open Knowledge: 23rd International Conference on Theory and Practice of Digital Libraries, TPDL 2019, Oslo, Norway, September 9-12, 2019, Proceedings 23. pp. 348–351. Springer (2019)
15. Li, X., Wang, W., Hu, X., Li, J., Tang, J., Yang, J.: Generalized focal loss v2: Learning reliable localization quality estimation for dense object detection. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 11632–11641 (2021)
16. Liu, X., Mian, L., Dong, Y., Zhang, F., Zhang, J., Tang, J., Zhang, P., Gong, J., Wang, K.: Oag know: Self-supervised learning for linking knowledge graphs. IEEE Transactions on Knowledge and Data Engineering **35**(2), 1895–1908 (2021)
17. Manghi, P., Atzori, C., Bardi, A., Baglioni, M., Schirrwagen, J., Dimitropoulos, H., La Bruzzo, S., Foufoulas, I., Mannocci, A., Horst, M., Czerniak, A., Kiatropoulou, K., Kokogiannaki, A., De Bonis, M., Artini, M., Ottonello, E., Lempesis, A., Ioannidis, A., Summan, F.: OpenAIRE Research Graph: Dumps for research communities and initiatives (Jun 2022). https://doi.org/10.5281/zenodo.6638478, `https://doi.org/10.5281/zenodo.6638478`
18. Manghi, P., Atzori, C., Bardi, A., Baglioni, M., Schirrwagen, J., Dimitropoulos, H., La Bruzzo, S., Foufoulas, I., Mannocci, A., Horst, M., et al.: Openaire research graph dump (2022)
19. Manghi, P., Bardi, A., Atzori, C., Baglioni, M., Manola, N., Schirrwagen, J., Principe, P., Artini, M., Becker, A., De Bonis, M., et al.: The openaire research graph data model. Zenodo (2019)

20. Pennington, J., Socher, R., Manning, C.D.: Glove: Global vectors for word representation. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). pp. 1532–1543 (2014)
21. Priem, J., Piwowar, H., Orr, R.: Openalex: A fully-open index of scholarly works, authors, venues, institutions, and concepts. arXiv preprint arXiv:2205.01833 (2022)
22. Rayleigh, J.W.S.B.: Scientific papers, vol. 1. University Press (1899)
23. Rettberg, N., Schmidt, B.: Openaire-building a collaborative open access infrastructure for european researchers. LIBER Quarterly: The Journal of the Association of European research libraries **22**(3), 160–175 (2012)
24. Rettberg, N., Schmidt, B.: Openaire: Supporting a european open access mandate. College & Research Libraries News **76**(6), 306–310 (2015)
25. Vrandečić, D., Krötzsch, M.: Wikidata: a free collaborative knowledgebase. Communications of the ACM **57**(10), 78–85 (2014)
26. Vukotic, A., Watt, N., Abedrabbo, T., Fox, D., Partner, J.: Neo4j in action, vol. 22. Manning Shelter Island (2015)
27. Wang, K., Shen, Z., Huang, C., Wu, C.H., Dong, Y., Kanakia, A.: Microsoft academic graph: When experts are not enough. Quantitative Science Studies **1**(1), 396–413 (2020)
28. Wang, Z., Lv, Q., Lan, X., Zhang, Y.: Cross-lingual knowledge graph alignment via graph convolutional networks. In: Proceedings of the 2018 conference on empirical methods in natural language processing. pp. 349–357 (2018)
29. Zhu, H., Xie, R., Liu, Z., Sun, M.: Iterative entity alignment via joint knowledge embeddings. In: IJCAI. vol. 17, pp. 4258–4264 (2017)