Assessing the Overlap of Science Knowledge Graphs: A Quantitative Analysis

Jenifer Tabita Ciuciu-Kiss and
Daniel Garijo

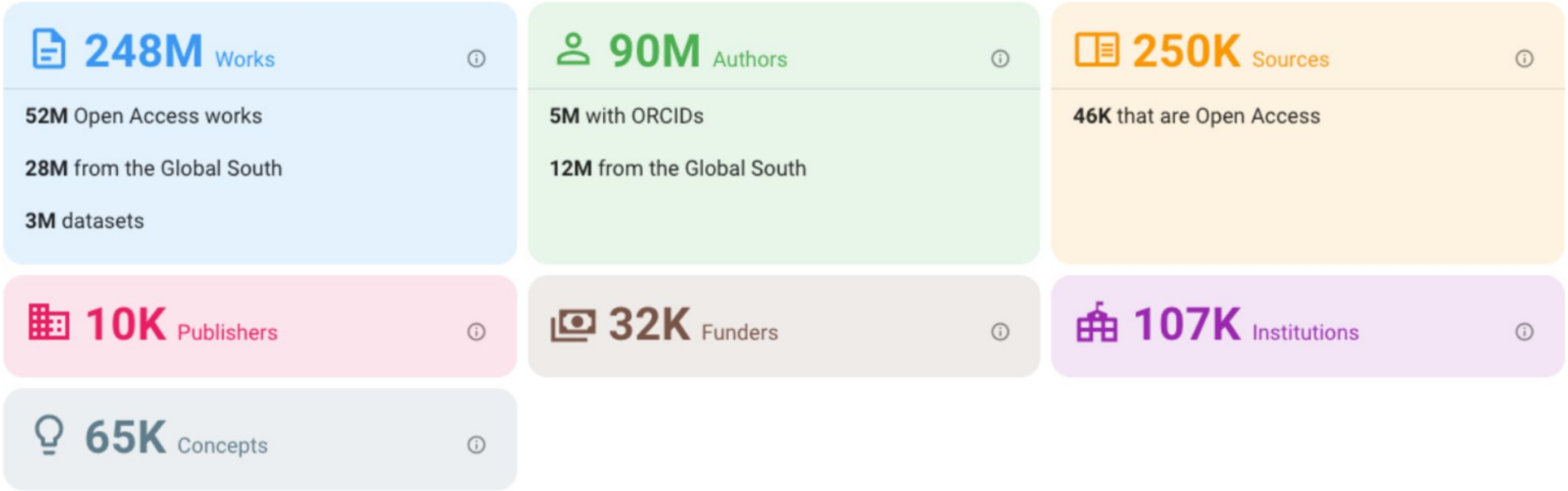
Ontology Engineering Group
Universidad Politécnica de Madrid, Spain

Science Knowledge Graphs (SKGs)

- Scientific knowledge base
- Vast amount of information
- E.g.: OpenAlex, OpenAIRE

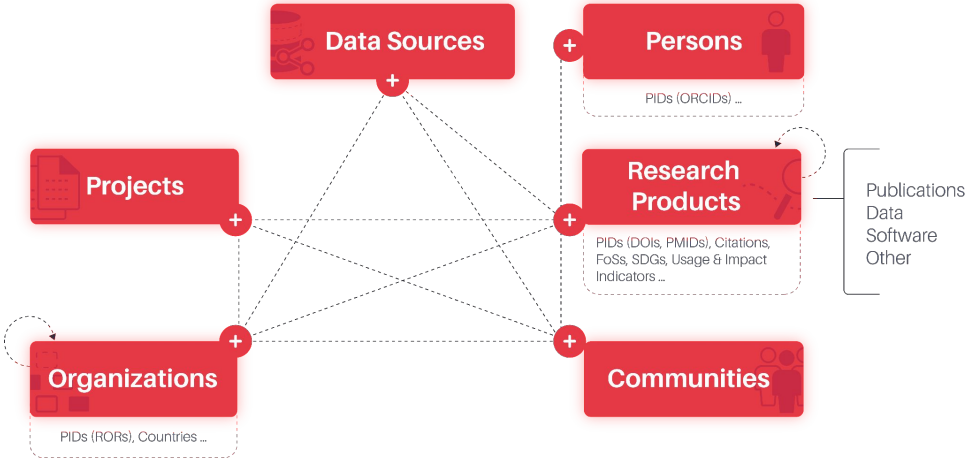
OpenAIRE

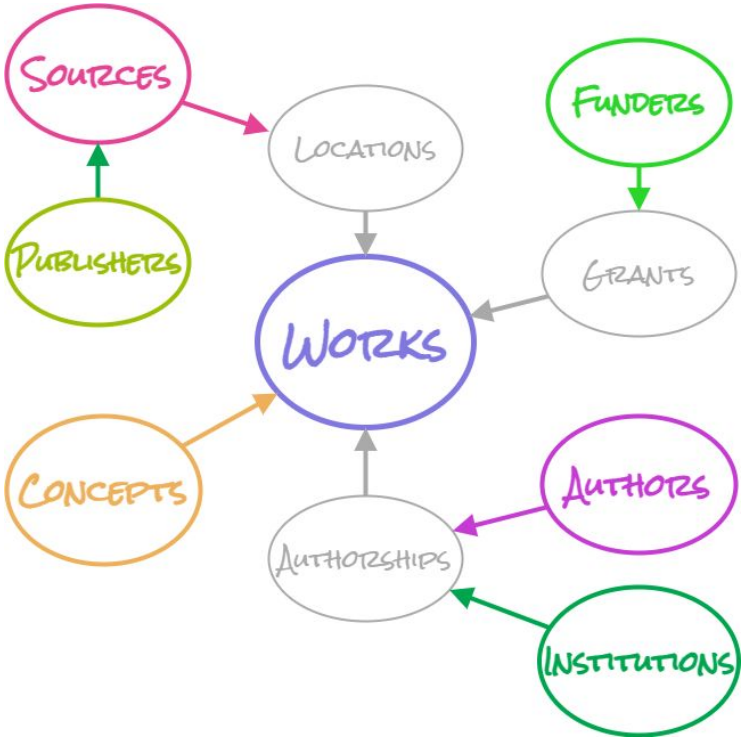
OpenAlex



Background

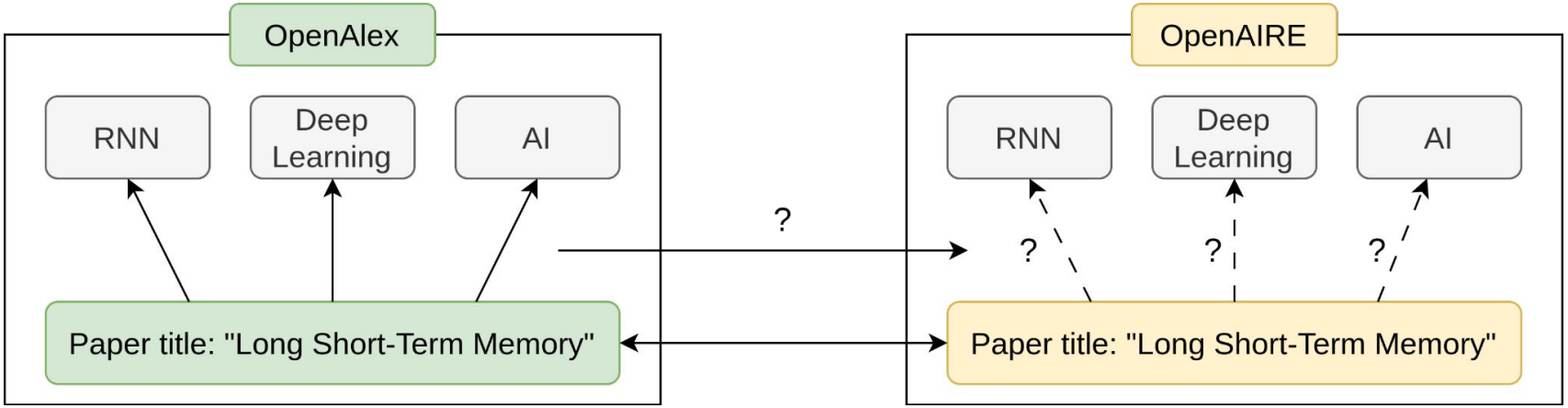
- Features of SKGs:
  - Millions of entities
  - Diverse structure and categorization
  - Integrate different sources





Problem

- Detect overlaps of papers among different SKGs
- Collect categorization of overlapping papers from the different SKGs
- Detect which categories may be related to each other



Proposed solution

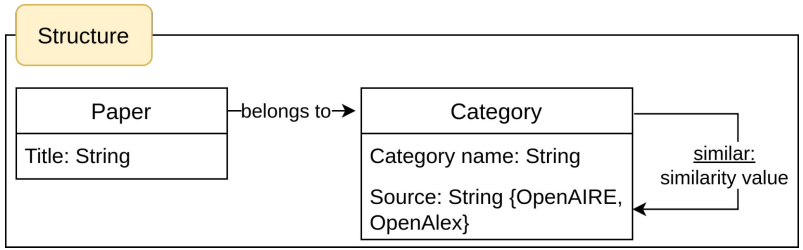
Quantitative bottom-up methodology to assess overlap among SKGs

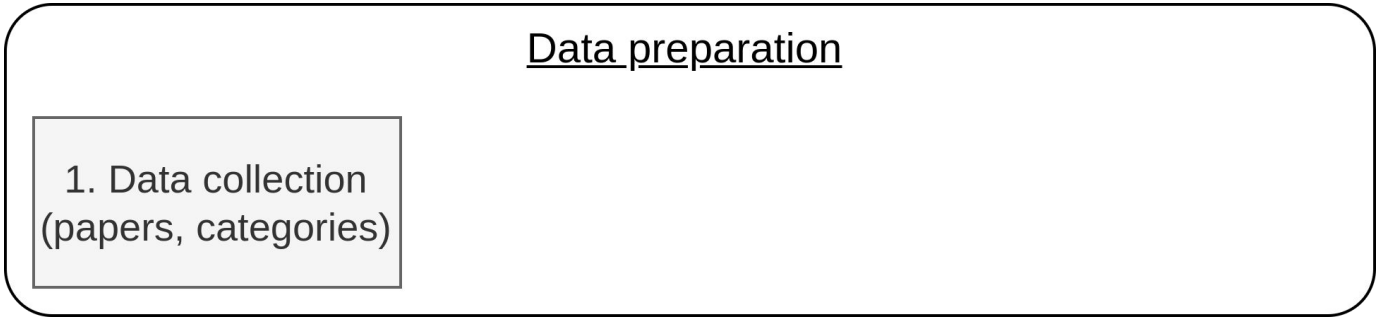
- Based on annotations collected from SKGs
- Counting with the number of papers belonging to similar categories
- PoC of ~100k papers in the AI domain

```
Paper title: "Long Short-Term Memory": {
Candidate categories: "openalex": [
            "recurrent neural network",
            "computer science",
            "backpropagation",
            "constant (computer programming)",
            "artificial neural network",
            "artificial intelligence",
            "algorithm",
            "term (time)",
            "deep learning",
            "physics",
            "quantum mechanics",
            "programming language"
        ],
Candidate categories: "openaire": [
            "long short term memory",
            "mean squared error",
            "statistics",
            "mathematics"
        ]
    }
```

Methodology

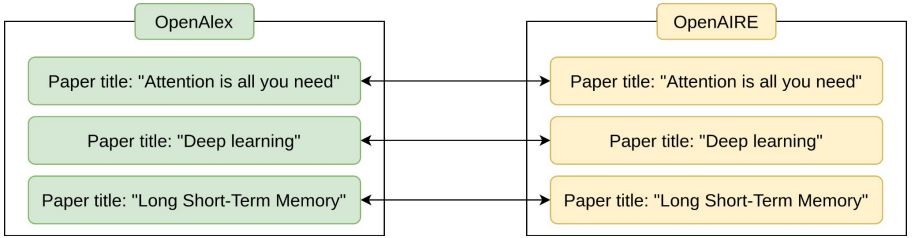
- Collect data from OpenAlex and OpenAIRE
  - Papers
  - Belonging categories
- Stored in the shown structure
- Using APIs of OpenAlex and OpenAIRE

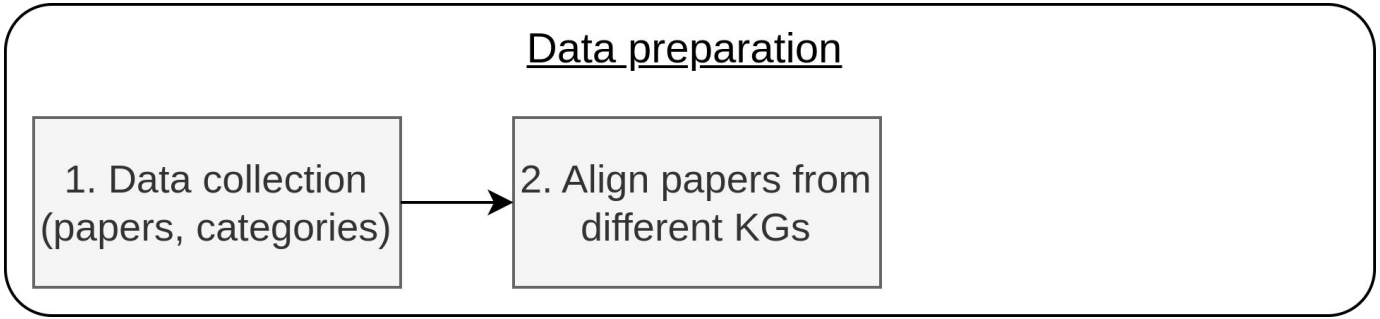




Methodology

- Align papers from different SKGs
  - Find papers that are present in all SKGs
  - Align based on an ID, e.g.: title or DOI

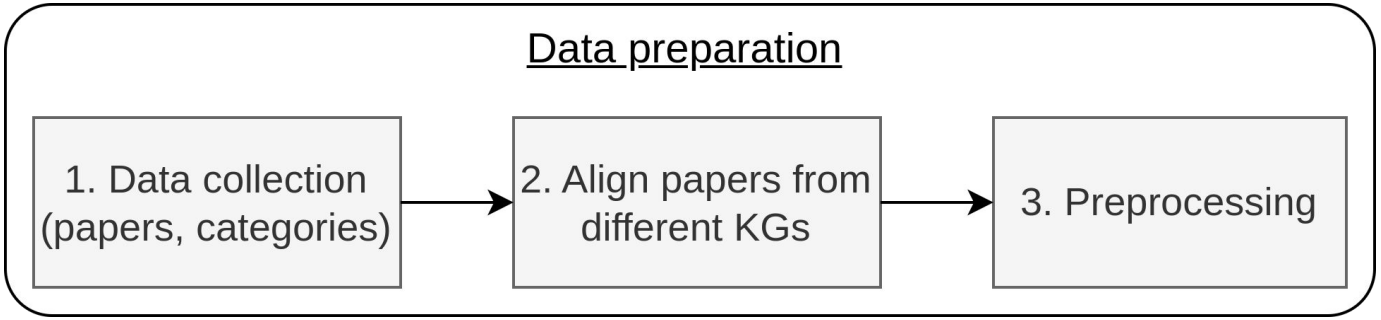




Methodology

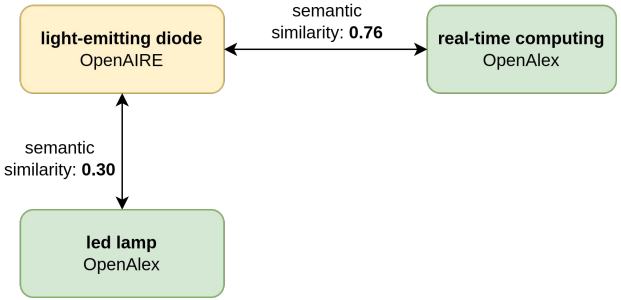
- Data preprocessing (paper titles, and category names)
  - Remove punctuations
  - Remove accents
  - Transform to lowercase
- The collected categories have to be **unified** for effective mapping

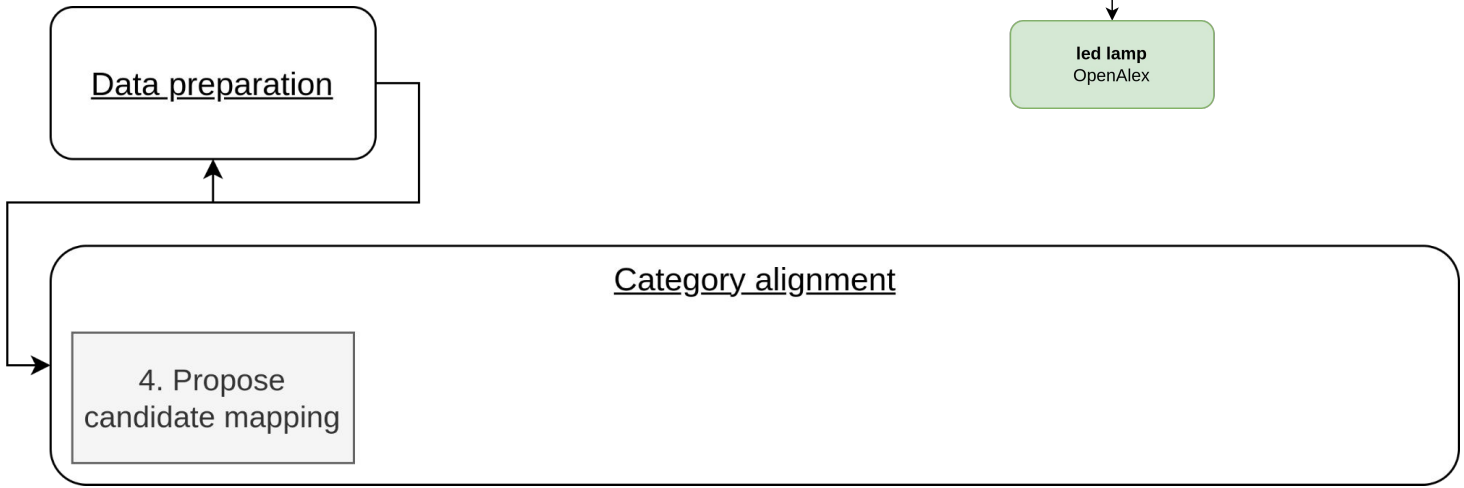
```
Example raw categories:
  - "computer.software_genre"
  - "0302 Clinical Medicine"
  - "business.industry"
  - "ddc:342"
```



Methodology

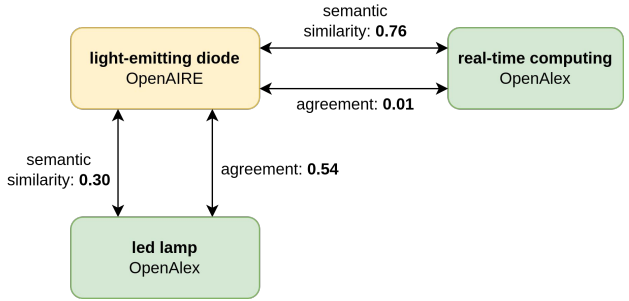
- Calculate semantic similarity based on GloVe embedding: [-1, 1]
- Propose candidate mapping
  - Exclude categories with <10 papers
  - Exclude complete overlaps
  - Exclude pairs with negative semantic similarity

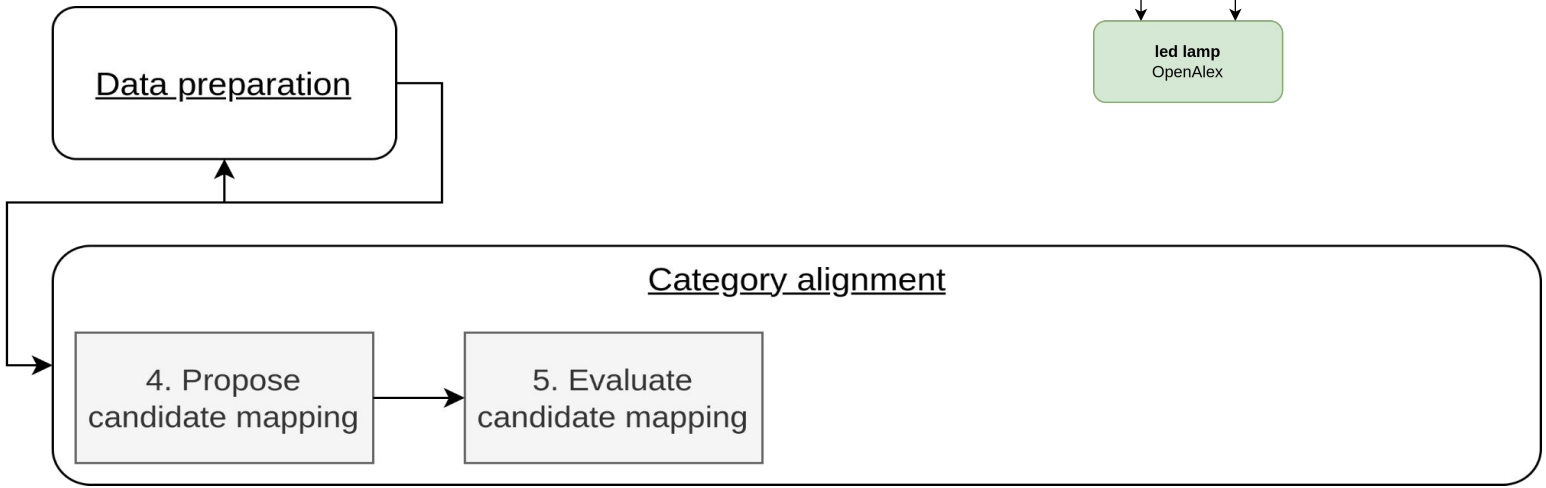




Methodology

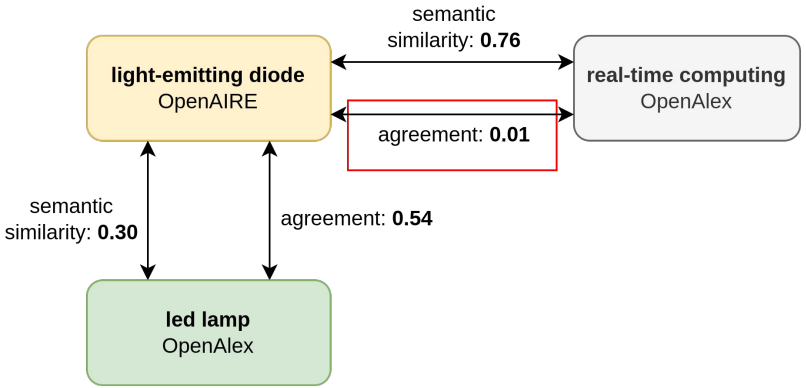
- Evaluate candidate mappings
- Based on two metrics
  - Semantic similarity of the mapped categories
  - Agreement (Intersection / Union of supporting papers)

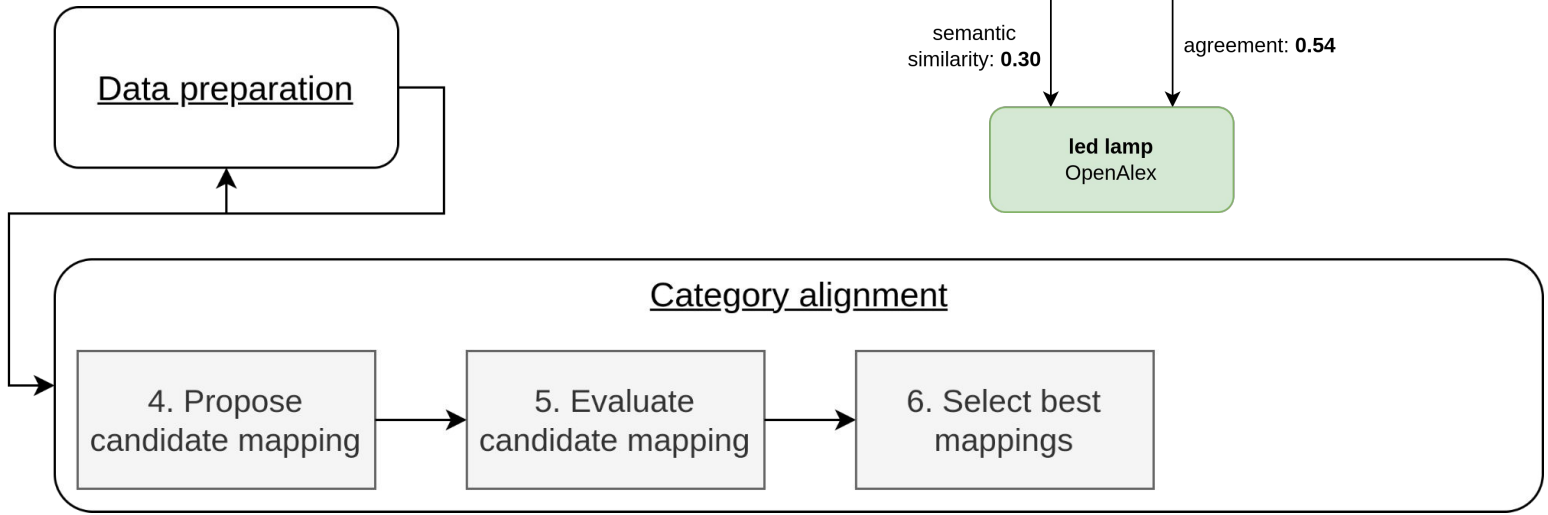




Methodology

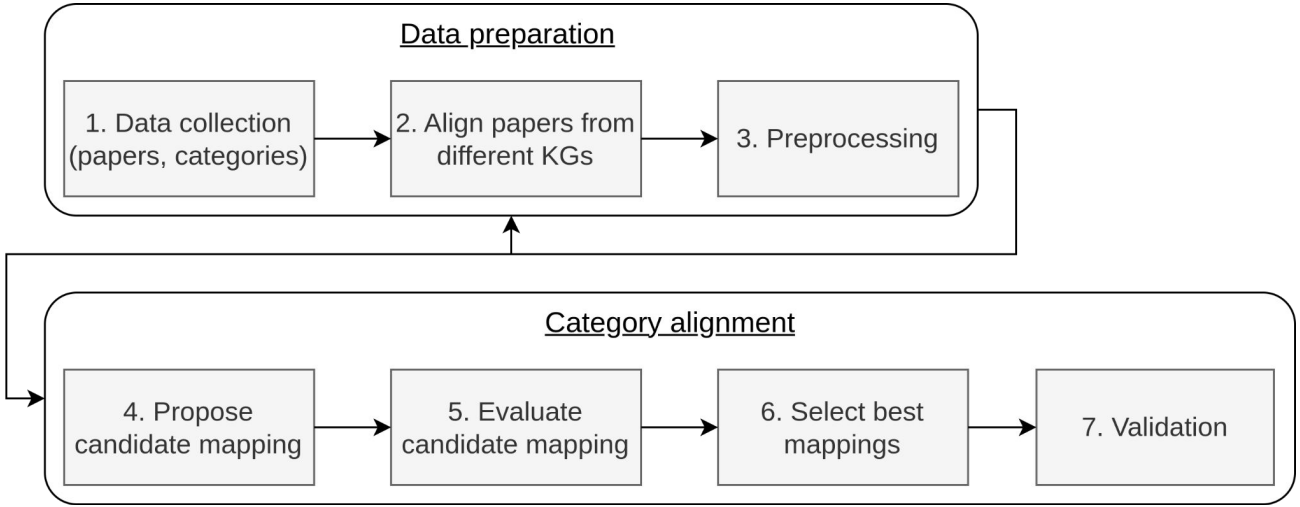
- Select the best mapping based on the 2 metrics:
  - Semantic similarity
  - Agreement
- Set a threshold of the agreement value: 0.5





Methodology

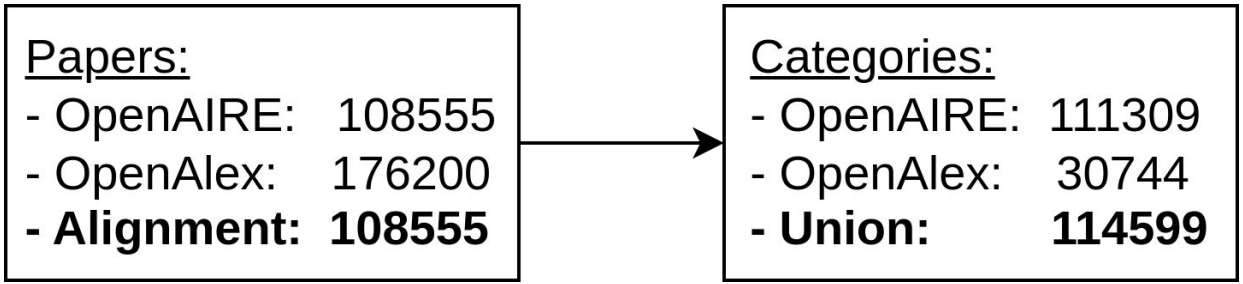
- Manual validation by experts
- 72 mappings proposed:
  - 14 misaligned labels:
    - e.g.: lasso (statistics) (OpenAIRE) – lasso (programming language) (OpenAlex)
  - 1 mismatch: peppered moth (OpenAIRE) – melanism (OpenAlex)



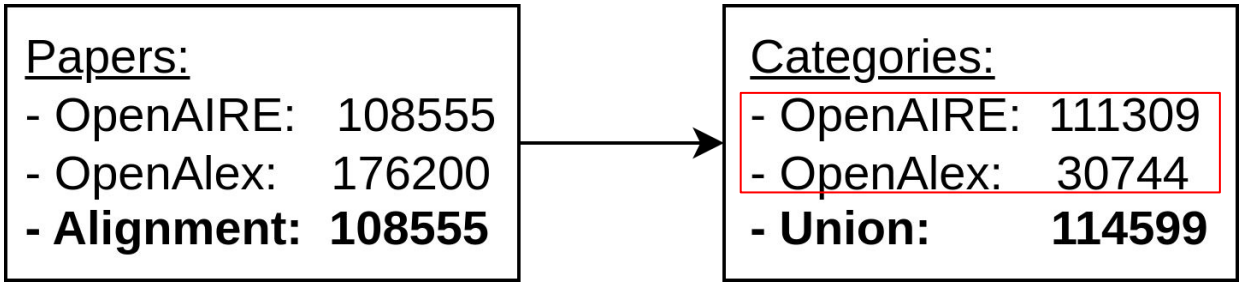
Data flow

Papers:
- OpenAIRE: 108555
- OpenAlex: 176200
- Alignment: 108555

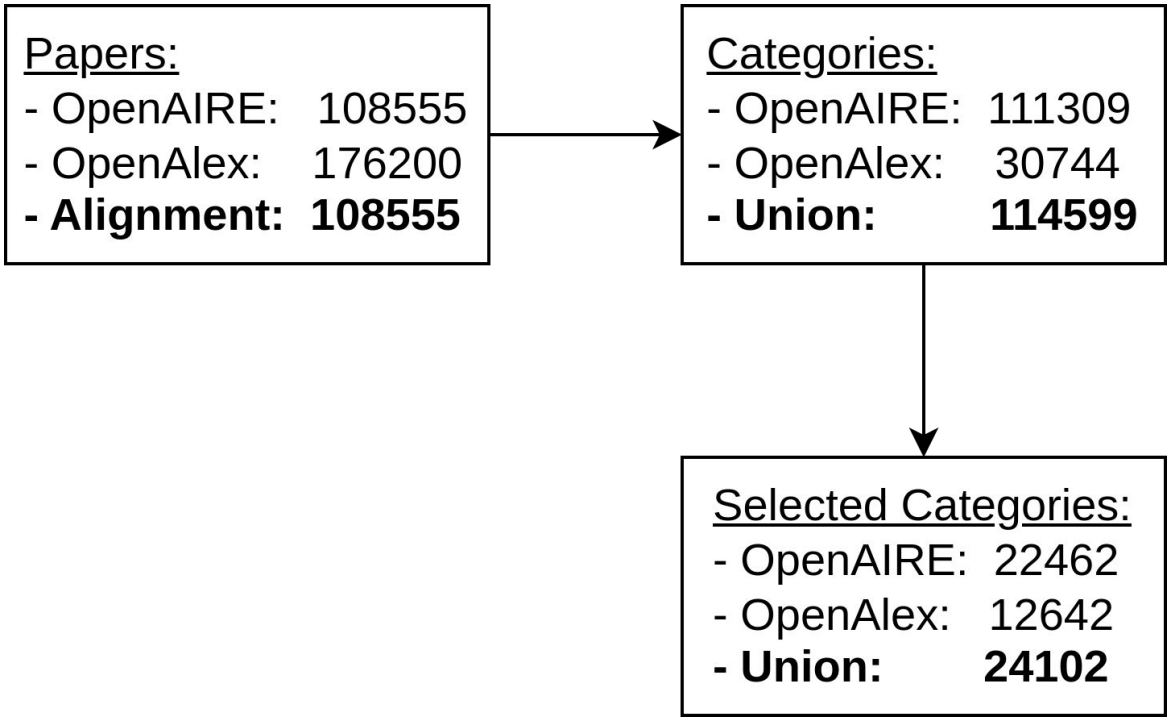
Data flow


Papers:
- OpenAIRE: 108555
- OpenAlex: 176200
- Alignment: 108555
Categories:
- OpenAIRE: 111309
- OpenAlex: 30744
- Union: 114599

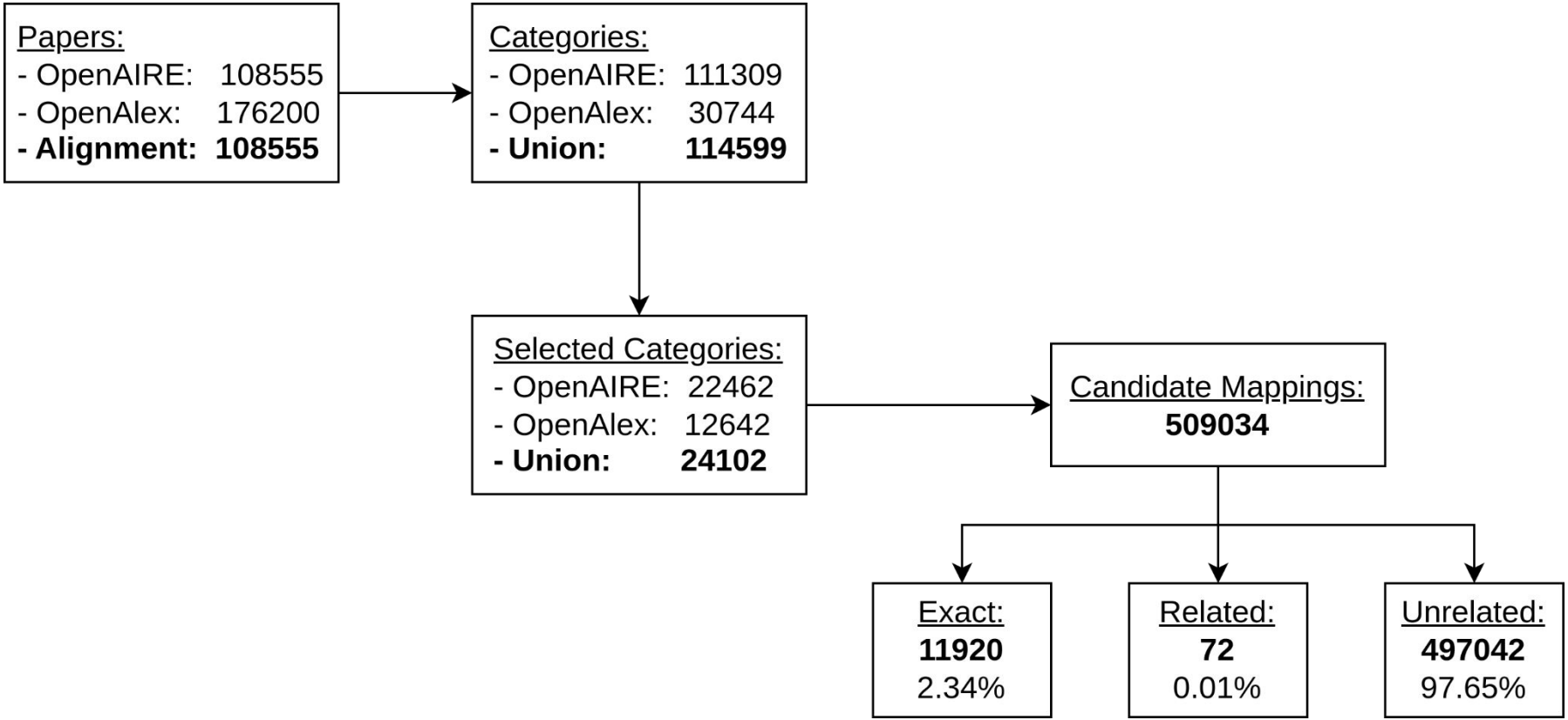
Data flow


Papers:
- OpenAIRE: 108555
- OpenAlex: 176200
- Alignment: 108555
Categories:
- OpenAIRE: 111309
- OpenAlex: 30744
- Union: 114599

Data flow

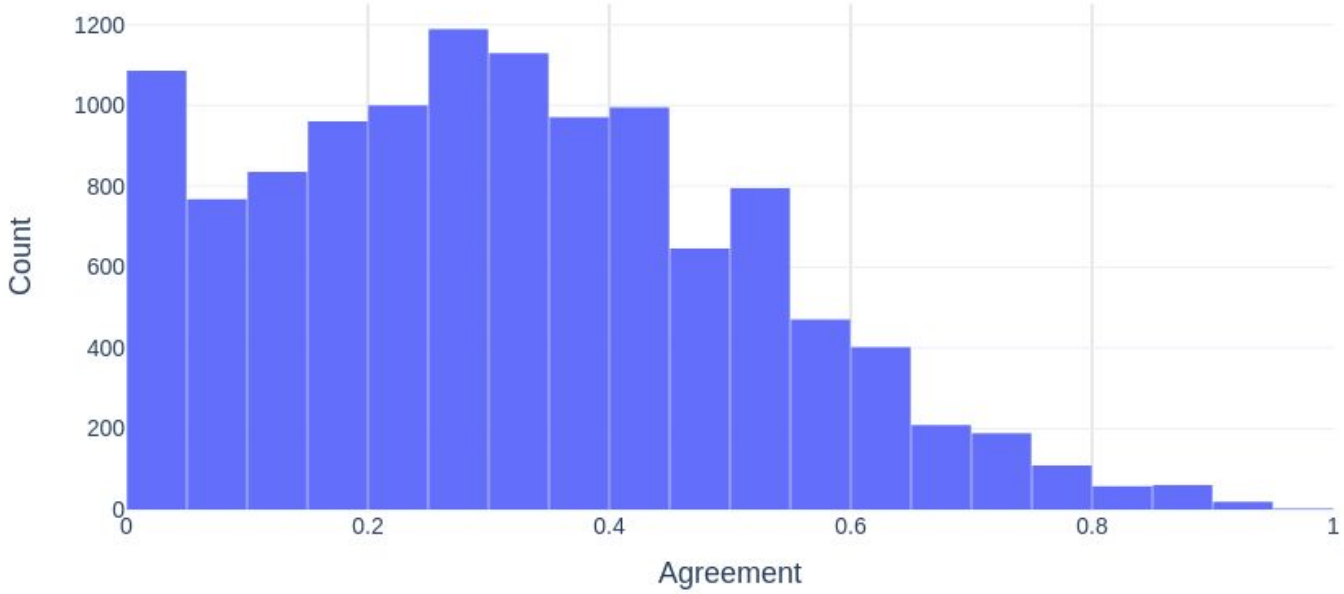

Papers:
- OpenAIRE: 108555
- OpenAlex: 176200
- Alignment: 108555
Categories:
- OpenAIRE: 111309
- OpenAlex: 30744
- Union: 114599
Selected Categories:
- OpenAIRE: 22462
- OpenAlex: 12642
- Union: 24102

Data flow


Papers:
- OpenAIRE: 108555
- OpenAlex: 176200
- Alignment: 108555
Categories:
- OpenAIRE: 111309
- OpenAlex: 30744
- Union: 114599
Selected Categories:
- OpenAIRE: 22462
- OpenAlex: 12642
- Union: 24102
Candidate Mappings:
509034
Exact:
11920
2.34%
Related:
72
0.01%
Unrelated:
497042
97.65%

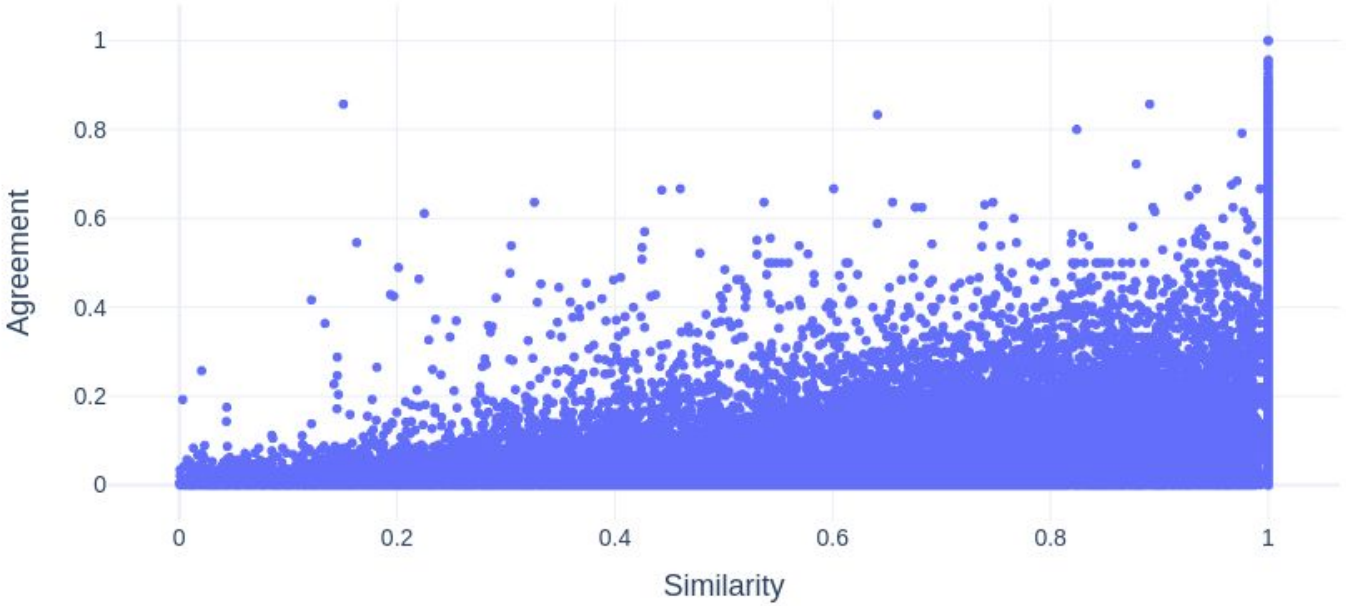
Results - Distribution of Agreement values

- Only overlapping categories are shown (**semantic similarity = 1.0**)
- **We can see a lack of agreement among the overlapping categories**



Results - Correlation of Agreement and Similarity

- **There is no correlation between the semantic similarity and agreement**



Conclusions and future work

- Developed a methodology to **quantitatively assess overlap among SKGs**
- PoC with ~100k papers
- 72 categories mapped
  - 14 misaligned labels:
    - e.g.: lasso (statistics) (OpenAIRE) – lasso (programming language) (OpenAlex)
  - 1 mismatch found: peppered moth (OpenAIRE) – melanism (OpenAlex)
- **Significant disagreement found among different SKGs**

- Extend the work to the **entire SKGs**
- Try different similarity embeddings (BERT, FastText, RoBERTA)

Assessing the Overlap of Science Knowledge Graphs: A Quantitative Analysis

Jenifer Tabita Ciuciu-Kiss and
Daniel Garijo

Ontology Engineering Group
Universidad Politécnica de Madrid, Spain

Number of mappings based on the agreement threshold