

# Developing Knowledge Graph Based On OpenAire Database

---

# Table of contents

1. Goal
2. What is the OpenAIRE Graph?
3. Dataset overview
4. Data structure
5. Implementation steps
6. Visualization of Knowledge Graph
7. Usage examples
8. Sources

# Goal

To develop a knowledge graph based on the OpenAIRE dataset to extract information about the most important research papers.

# What is the OpenAIRE Graph?



The OpenAIRE Graph is a free and open resource that brings together and interlinks hundreds of millions of metadata records from over 100k data sources trusted by researchers.

The dataset selected contains metadata records of the OpenAIRE Graph for the research community “Technological University Network (TU-NET).”

TU-NET is a network for the Irish Technological Universities to share expertise, information and resources

# Data structure

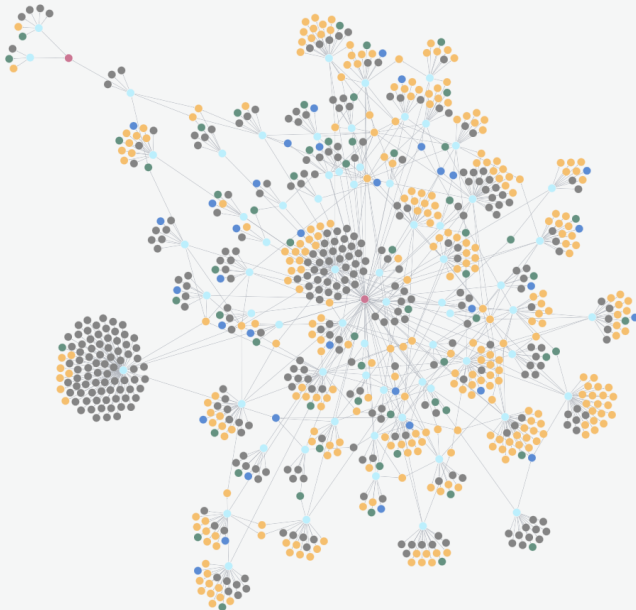
The dataset is stored in multiple archive files, where each line represents a metadata record. Below is an example of a single line from the dataset:

```
{
  "author": [
    {
      "fullName": "Hegarty, Nora",
      "name": "Nora",
      "surname": "Hegarty",
      "rank": 1,
      "pid": null
    },
    {
      "fullName": "Carbery, Alan",
      "name": "Alan",
      "surname": "Carbery",
      "rank": 2,
      "pid": null
    },
    {
      "fullName": "Hurley, Tina",
      "name": "Tina",
      "surname": "Hurley",
      "rank": 3,
      "pid": null
    }
  ],
  "openAccessColor": null,
  "publiclyFunded": true,
  "type": "publication",
  "language": {
    "code": "eng",
    "label": "English",
    "country": "IE",
    "label": "Ireland"
  },
  "provenance": {
    "provenance": "Inferred by OpenAIRE",
    "trust": "0.85"
  },
  "subjects": [
    {
      "subject": {
        "scheme": "keyword",
        "value": "NONE OF THESE"
      }
    }
  ],
  "provenance": {
    "provenance": "Harvested",
    "trust": "0.99"
  },
  "mainTitle": "Learning by Doing: Re-designing the First Year Information Literacy Programme at WIT Libraries",
  "description": "The purpose of this paper is to describe the process involved in re-designing Waterford Institute of Technology (WIT) Libraries' information literacy programme for first year students. It is written by some of the members of the library learning support team, who deliver the programme. It describes the steps involved in the programme's development and design, discusses the pedagogical principles that influenced the initiative, and summarises the evaluations we have undertaken to date. These evaluations have yielded positive informal and formal feedback from the students and lecturers who participated in the programme. The value of a pedagogically sound, active learning approach to information literacy training is highlighted throughout the results. By providing constructive solutions for incorporating active learning into library user education programmes, this paper is expected to be a useful source of practical information for libraries in similar positions, of similar scale, faced with similar challenges. It is likely to be of particular interest to librarians involved in information literacy education.",
  "publicationDate": "2009-12-01",
  "source": {},
  "format": {
    "application/pdf"
  },
  "contributor": {},
  "coverage": [
    {
      "bestAccessRight": {
        "code": "c_abf2",
        "label": "OPEN",
        "scheme": "http://vocabularies.coar-repositories.org/documentation/access_rights/"
      },
      "id": "dedup_wf_002:1216e03a73cdcc5572982bba297eda5",
      "originalId": "oai:repository.wit.ie:1431",
      "oai:generic.eprints.org:1431",
      "pid": "dateOfCollection": "2024-07-05T19:14:43.524",
      "lastUpdateTimestamp": "1728173418144",
      "indicators": {
        "citationImpact": {
          "citationCount": 0.0,
          "influence": 2.9837197E-9,
          "popularity": 5.9487684E-10,
          "impulse": 0.0,
          "citationClass": "C5",
          "influenceClass": "C5",
          "popularityClass": "C5"
        },
        "context": [
          {
            "code": "eu-conexus",
            "label": "European University for Smart Urban Coastal Sustainability",
            "provenance": {
              "provenance": "Inferred by OpenAIRE",
              "trust": "0.8"
            },
            "code": "tunet",
            "label": "ITU-NET",
            "provenance": {
              "provenance": "Inferred by OpenAIRE",
              "trust": "0.8"
            }
          }
        ],
        "collectedFrom": {
          "key": "opendoar____:4daa3db355ef2b0e64b472968cb70f8d",
          "value": "SETU Open Access Repository"
        },
        "instance": {
          "pid": [
            {
              "accessRight": {
                "code": "c_abf2",
                "label": "OPEN",
                "scheme": "http://vocabularies.coar-repositories.org/documentation/access_rights/"
              },
              "openAccessRoute": null,
              "type": "Article",
              "url": "http://repository.wit.ie/1431/",
              "publicationDate": "2009-12-01",
              "referenced": "peerReviewed",
              "hostedBy": {
                "key": "opendoar____:4daa3db355ef2b0e64b472968cb70f8d",
                "value": "SETU Open Access Repository"
              },
              "collectedFrom": {
                "key": "opendoar____:4daa3db355ef2b0e64b472968cb70f8d",
                "value": "SETU Open Access Repository"
              }
            },
            {
              "isGreen": true,
              "isInDiamondJournal": false
            }
          ]
        }
      }
    }
  ]
}
```

## Implementation steps

1. Read each line from all archive files and store it in a variable.
2. Convert the data into CSV format by selecting relevant fields. Each CSV file represents a node in the knowledge graph.
3. Preprocess and clean the data to ensure consistency and accuracy.
4. Connect to Neo4j Aura (console) using the Neo4j library.
5. Use Cypher queries to establish relationships between the nodes.
6. Upload the processed data into the Neo4j database.

# Visualization of Knowledge Graph





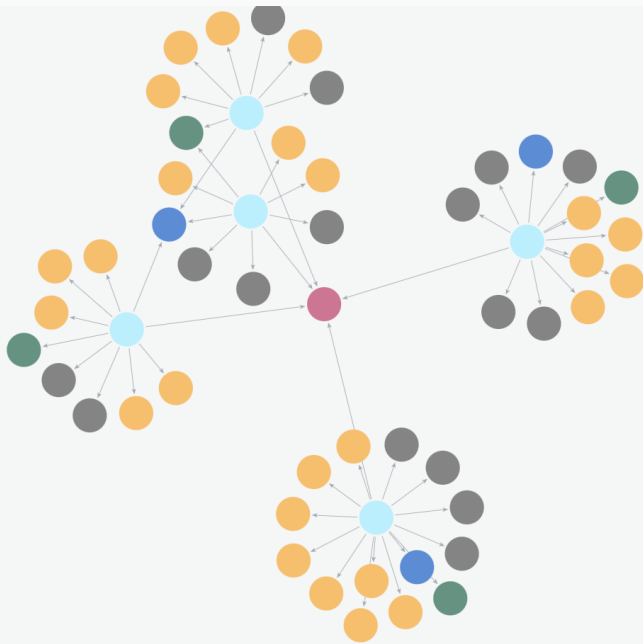
## Usage examples

The following Cypher query retrieves the most cited records in English:

```
MATCH (r:Record)-[:HAS_LANGUAGE]->(l:Language)
WHERE r.citationCount IS NOT NULL AND l.language = 'English'
WITH r
ORDER BY r.citationCount DESC
LIMIT 5
```

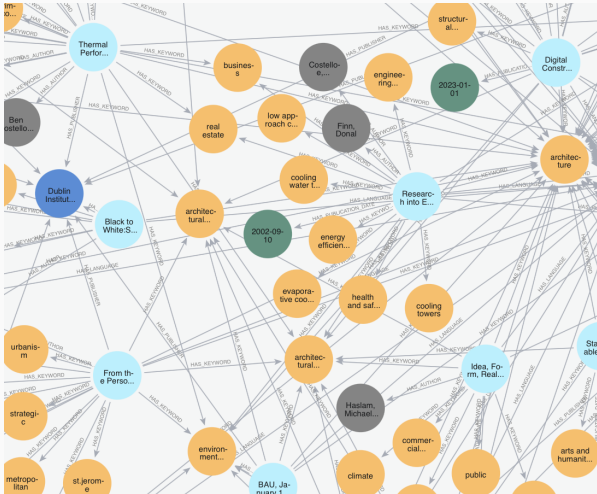
```
MATCH p=(r)-[rel]-(m)
RETURN p
```

## Usage examples



## Usage examples

All records associated with the keyword architecture



1. OpenAire Graph - <https://graph.openaire.eu>
2. Neo4j documentation - <https://neo4j.com/docs/getting-started/languages-guides/neo4j-python/>
3. Dataset source - <https://zenodo.org/records/13135167>