OpenAIRE Datathon Report: SMS Team

Al Koudous Idrissou[1,2], Ali Khalili[1], and Peter van den Besselaar[2]

[1] Department of Computer Science, Vrije Universiteit Amsterdam, NL
[2] Department of Organization Sciences, Vrije Universiteit Amsterdam, NL
{o.a.k.idrissou,a.khalili, p.a.a.vanden.besselaar}@vu.nl

Abstract. This document briefly reports on the work done as part of the OpenAIRE datathon to show how OpenAIRE data could be of help for the pursuit of navigating across *Open Data Sources*. First, a quick *data quality* analysis directed us toward the need for *harmonizing* and *enriching* one important property used by OpenAIRE to describe its organisations: the property *country*. Doing so, it allowed us to move to the *data interlinking* step. For this purpose, we selected three external datasets namely GRID (Global Research Identifier Database), OrgRef (open data about academic & research organizations) and Cordis EU H2020 Organizations dataset. To stay on course with our goal, we performed a full-mesh data interlinking task across all datasets as only this allows navigating from one dataset to another. The interlinking was done using the Lenticular Lens approach, and then embedded in the SMS (Semantically Mapping Science) platform to enable end-users to browse and visualize the interlinked data to address research questions of their interest.

Keywords: OpenAIRE, Data Linking, Data Enrichment, Datathon

1 Data Enrichment

OpenAIRE contains 134528 organisations distributed across countries. Looking at the number of organisations described with a country information, it shows that only 63% Listing 1.1 of them use the country info. A detailed look at these organisations reveals that OpenAIRE uses both country code (e.g,. NG, US, UA, UNKNOWN) and DBpedia URIs (e.g., http://dbpedia.org/resource/Argentina) to describe the same geographical entity Listing 1.2.

```
#============================================================================#
# 1. OPENAIRE COUNTRY COVERAGE                                               #
# RESULT: 196, Distributed over 84833 organizations                          #
#============================================================================#
PREFIX voc: <http://lod.openaire.eu/vocab/>
SELECT (count(DISTINCT ?country) as ?total_Countries)
WHERE
{
   ?subject a voc:OrganizationEntity; voc:country ?country .
}
```

Listing 1.1. Statistics on OpenAire country coverage

```
#============================================================================#
### 2. OPENAIRE COUNTRY COVERAGE INCONCISTENCY                                #
# RESULT: 155, it affects ≃ 71% of the mentioned organizations (60055)       #
#  "NO"                                                                      #
#  http://dbpedia.org/resource/Argentina                                     #
#  http://dbpedia.org/resource/Luxembourg                                    #
#  http://dbpedia.org/resource/Cambodia                                      #
#  "CA"                                                                      #
#  http://dbpedia.org/resource/Turkey                                        #
#  "NG"                                                                      #
#  "US"                                                                      #
#  "UA"                                                                      #
#  "UNKNOWN"                                                                 #
#============================================================================#
PREFIX voc: <http://lod.openaire.eu/vocab/>
SELECT (count(DISTINCT ?country) as ?total_Countries)
WHERE
{
  # RESOURCE WITH DBPEDIA URI AS COUNTRY
  ?subject a voc:OrganizationEntity ; voc:country ?country .
  FILTER (isIRI(?country))
}
```

Listing 1.2. OpenAire country coverage inconsistencies statistics

For consistency, we added two new predicates to document *countryCode* and *countryName*. All OpenAIRE organizations with a correct country code are then enriched with the country code predicate (Listing 1.3). Using DBpedia, all organisations with a DBpedia URI are enriched with the property *countryName* and the actual country label extracted from DBpedia (Listing 1.4).

```
#============================================================================#
# 3. COUNTRY CODE PREDICATE                                                  #
# ENRICHING OPENAIRE BY ASSOCIATING THE EXISTING COUNTRY CODE                #
# VALUES TO A NEW AND MEANINGFUL PROPERTY: voc:countryCode                   #
# Insert into <https://www.openaire.eu>, 24778 (or less) triples -- done     #
#============================================================================#
prefix dbo: <http://dbpedia.org/ontology/>
PREFIX voc: <http://lod.openaire.eu/vocab/>
INSERT
{
   GRAPH ?g
   {
      ?subject voc:countryCode ?country.
   }
}
WHERE
{
  GRAPH ?g
  {
    ?subject a voc:OrganizationEntity; voc:country ?country .
    FILTER (isIRI(?country ) = "false"^^xsd:boolean)
  }
}
```

Listing 1.3. Country code enrichment

Finally, using RISIS's country harmonization dataset, we added respectively country name to organisations with country code and country code to organisations with country name (Listing 1.5, Listing 1.6). Doing so, out of the 63% of organisations documented with either country code or DBpedia URI, we now

have 54% of organisations documented with the correct country code and 62% of organisations documented with the appropriate country name label.

```
#============================================================================#
# 4. ENRICHING OPENAIRE WITH AN HARMONISED COUNTRY [COUNTRY NAME]            #
#============================================================================#
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX dbo: <http://dbpedia.org/ontology/>
PREFIX voc: <http://lod.openaire.eu/vocab/>
INSERT
{
  GRAPH <https://www.openaire.eu>
  {
     ?subject voc:countryName ?name.
  }
}
WHERE
{
  # RESOURCES WITH A DBPEDIA URI AS COUNTRY NAME
  GRAPH <https://www.openaire.eu>
  {
    ?subject a voc:OrganizationEntity; voc:country ?country .
    FILTER (isIRI(?country ))
  }

  # FIND THE DBPEDIA COUNTRY NAME LABEL
  {
    SELECT *
    {
       SERVICE <http://dbpedia.org/sparql>
       {
          ?country a dbo:Country ; rdfs:label  ?name .
          FILTER langMatches( lang(?name), "en" )
       }
    }#limit 10000
  }
} #limit 100
"Insert into <https://www.openaire.eu>, 73490 (or less) triples -- done" .

# TOTAL countryCode: 73613 DISTINCTS
# TOTAL countryName: 133351 AND 84516 DISTINCTS
PREFIX voc: <http://lod.openaire.eu/vocab/>
SELECT (COUNT(DISTINCT ?subject) AS ?TOTAL)
WHERE { GRAPH <https://www.openaire.eu> { ?subject voc:countryCode ?name. } }
```

Listing 1.4. Country name enrichment

```
#============================================================================#
# 5. ENRICHING OPENAIRE WITH [COUNTRY CODE] FROM [COUNTRY NAME]              #
#============================================================================#
PREFIX dbo: <http://dbpedia.org/ontology/>
PREFIX voc: <http://lod.openaire.eu/vocab/>
PREFIX cvoc: <http://rdf.risis.eu/datasets/countries/vocab/>
INSERT
{
  GRAPH <https://www.openaire.eu>
  {
    ?subject voc:countryCode ?code_2 .
  }
}
WHERE
{
  # RESOURCES WITH A PROPER NAME LABEL
  GRAPH <https://www.openaire.eu>
  {
    ?subject voc:countryName ?name .
  }
```

```
21
  # FIND THE COUNTRY CODE FOR THE SPECIFIED LABEL
23  {
      SELECT *
25    {
        SERVICE <http://stardog.risis.d2s.labs.vu.nl/annex/risis/sparql/query?>
27      {
          graph <http://risis.eu/dataset/countries>
29        {
            ?country cvoc:ISO3166_1_Alpha_2 ?code_2 .
31          ?country cvoc:official_name_en ?namex .
            BIND (STRLANG(?namex , "en") as ?name )
33        }
        }
35    }#limit 100
     }
37 } #limit 100
```

Listing 1.5. Country code enrichment

```
1 #================================================================================#
  # 6. ENRICHING OPENAIRE WITH AN HARMONISED COUNTRY                               #
3 # CODE PROPERTY AND CONCISTENT COUNTRY CODE VALUES                               #
  #================================================================================#
5
  prefix dbo: <http://dbpedia.org/ontology/>
7 PREFIX voc: <http://lod.openaire.eu/vocab/>
  PREFIX cvoc: <http://rdf.risis.eu/datasets/countries/vocab/>
9
  # ENRICHED OPENAIRE WITH countryName AND countryAltName
11 INSERT
  {
13  GRAPH <https://www.openaire.eu>
    {
15    ?subject voc:countryName ?name .
      ?subject voc:countryAltName ?altName .
17  }
  }
19
  WHERE
21 {
    # RESOURCES WITH A COUNTRY CODE PREDICATE
23  GRAPH <https://www.openaire.eu>
    {
25     ?subject voc:countryCode ?code_2 .
    }
27
    # ENRICHED INFORMATION FROM THE COUNTRIES DATABASE
29  {
      SELECT *
31    {
        SERVICE <http://stardog.risis.d2s.labs.vu.nl/annex/risis/sparql/query?>
33      {
          GRAPH <http://risis.eu/dataset/countries>
35        {
            ?country cvoc:ISO3166_1_Alpha_2 ?code_2 .
37          ?country cvoc:official_name_en ?name .
            ?country cvoc:name ?altName.
39        }
        }
41    }#limit 100
    }
43 } #limit 100
```

Listing 1.6. Country name enrichment

```
#=============================================================================#
# 7. OPENAIRE DEDUPLICATION LINKSET                                           #
# This affect some 11855 organisations (≃ 14%)                                #
#=============================================================================#
PREFIX voc: <http://lod.openaire.eu/vocab/>
PREFIX owl: <http://www.w3.org/2002/07/owl#>
INSERT
{
  GRAPH <http://openaire2018/linkset>
  {
    ?subject owl:sameAs ?object .
  }
}
WHERE
{
   ?subject owl#sameAs ?object .
}
```

Listing 1.7. Linkset creation

This *consistency*, *harmonisation* and *enrichment* procedures could be improved by using *inferencing* over *validated* discovered links or using the *sameAs* links included in the OpenAIRE database which account for 14% of overall organisations (Listing 1.7).

Overall, this task reveals that other measures need to be taken for making sure that the renaming 38% of OpenAIRE's organisations with neither CountryName nor CountryCode, are properly updated with the right and consistent values.

2 Entity Linking

In this step, we investigate how OpenAIRE connects to three other open datasets namely GRID, OrgRef and H2020 as depicted in Figure 1. To reach the entity linking goal, we used two strategies available in the Lenticular Lens tool [1]: linking based on cluster and linking based on refinement.

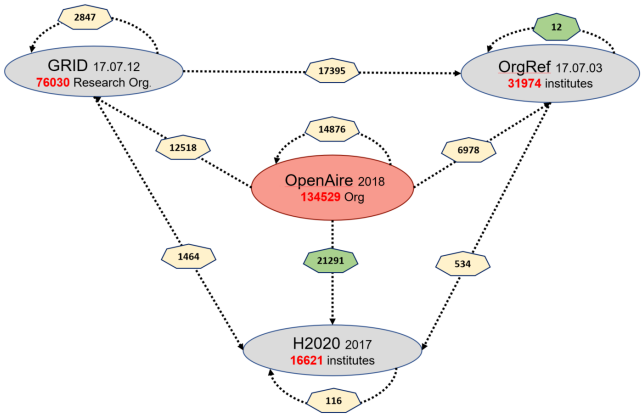

Fig. 1. Link Statistic Overview (Approximate Similarity)

2.1 Selected Datasets

The information provided here about the datasets utilised for the experiments conducted in this datathon were collected in January 2018. However, the datasets used are of earlier dates: grid-2017.07.12 Orgref-2017.07.03, OpenAire-2018.08.16, and H2020-2017.

Worldwide Organisations. The datasets (**Grid** [3] and **OrgRef** [4]) in this category provide some general information about particular organisations.

Grid, describes 80248 organisations across 221 countries using 12308 relationships. Only 17 countries (United States, United Kingdom, Japan, Germany, France, Canada, Czechia, China, India, Norway, Italy, Spain, Brazil, Russia, Switzerland, Sweden and Australia) within Grid host a thousand or more organisations. This account for 77% of its overall organisations. All organisations within Grid are assigned an address property value. While 96% of them have an organisation type (Company, Education, Healthcare, Nonprofit, Facility, Other, Government and archive), only 78% have geographic coordinates.

OrgRef collates existing data about the most important worldwide academic and research organisations (31000) from two main sources: Wikipedia and ISNI. These institutions are composed of universities, colleges, schools, hospitals, government agencies and companies involved in research.

European Organisations. Horizon 2020 is the largest EU Research and Innovation programme with the focuses on excellent sciences, industrial leadership and societal challenges. To this end, projects supported by the EU are documented in the **European Organisations' Projects H2020** database.

2.2 Lenticular Lenses Method

Lenticular Lenses (LL) is a context-sensitive alignment method which keeps track of provenance. It supports the use of several combinations of properties as identity criteria depending on the context and the use and combination of several alignment methods to generate candidate links. Furthermore, it allows for two types of analysis: (i) clustering alignments in a way that supports verifying the occurrence of an entity in several datasets, and (i) creating views (tables) on the datasets based on automated SPARQL queries that can be useful for users as a start point for more complex queries. The validation is currently supported in a fine-grained level, i.e. each pair of linked resources can be validated individually. A more advanced validation is under development, which will allow for manually validating a whole cluster at once or automatically validating several clusters based on metrics.

[3] `https://www.grid.ac/stats`
[4] `http://www.orgref.org/web/about.htm`

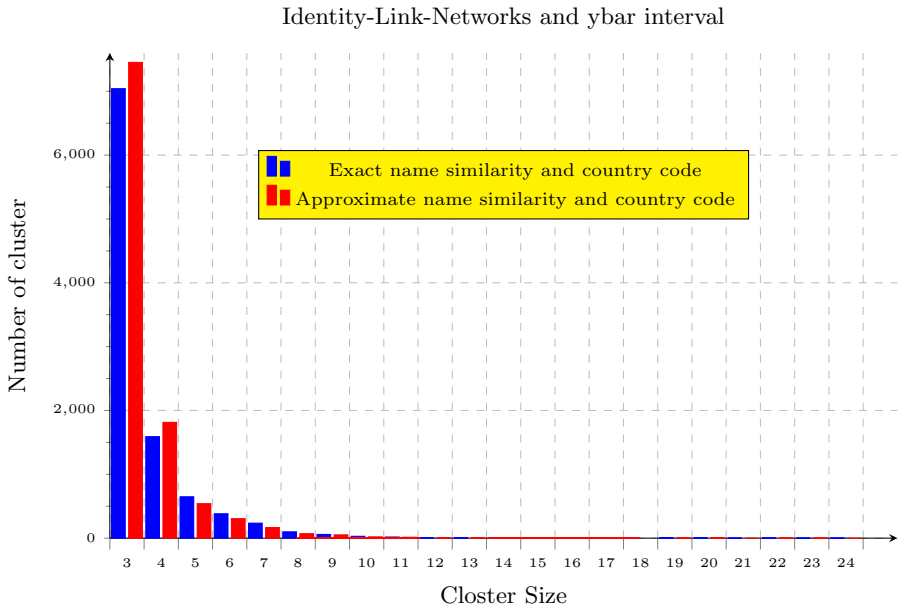

Fig. 2. Cluster size and cluster counts

2.3 Linking with exact string similarity and clustering

Here, the first step is to cluster together organisations that share the same country code or country name. Then, for each cluster, find organisations with the exact names.

Figure 2 shows for example that 7041 identity-link-clusters are found as three different representations of a unique entity. Figure 3 and Figure 4 are examples of identity-link-networks. In the first figure, the Mississippi State University is represented using six different URIs across three datasets (OpenAIRE, GRID and OrgRef). Out of these six mentioned of the same university, OpenAIRE refers to it in three different ways (duplicates).

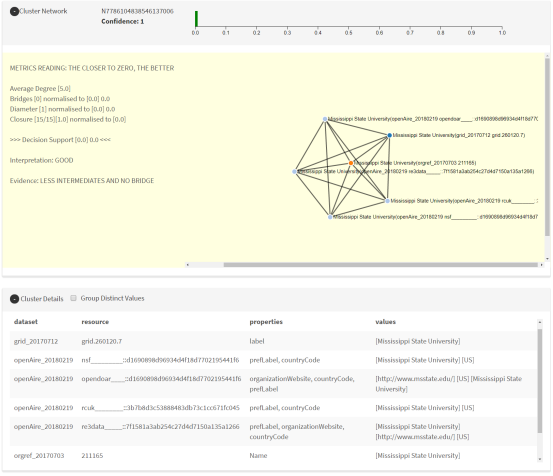

Fig. 3. Identity Link network: Mississippi State Uni.

Although the second figure is also an identity-link-network of six representations of a the unique real world Leipzig University, its network is created with 9 red dashed links. Links between entities are the materialisation of a relationship between the concerned entities. As these link are quantitative, the strength of a link is showed using coloured and dashed lines. A black link represents a 100% (exact match) match while the red dashed links represent a match less than 100%. The lower the strength gets, the more spaced the dash are. So, contrarily to Figure 3, even though Figure 4 is also a complete identity-link-network, there is a cloud of uncertainty on deciding whether Figure 4 is a true representation of the Leipzig University. This particular example is indeed a single representation of the Leipzig University. Anyhow, the point we try to make here is that, scenarios like this require an experts for the validation of the identity-link-network.

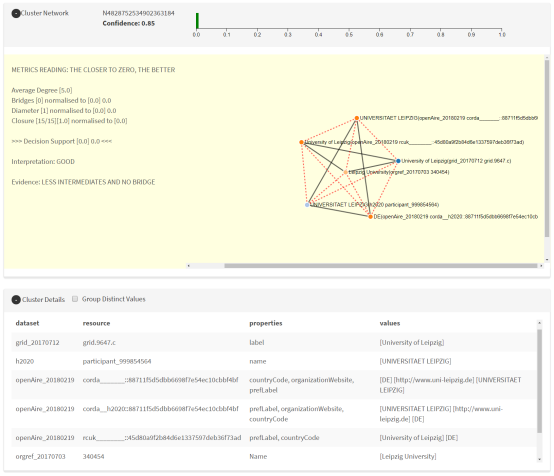

Fig. 4. Identity Link network: Leipzig University

In the next section, we describe how approximate links where computed during the experiments.

2.4 Linking with string approximation

In the second attempt to interlink our four selected open datasets using the Lenticular Lens tool, we first link organisations across theses datasets whenever their names are similar by 80% and then filter out links where the organisations are not within the same country. Table 1 shows in red the number of links found only using the approximate name similarity, and in black the number of links found after making sure that organisations on both sides of the links share the same geographical location. In Figure 1 we only illustrate the refined links. The figure also shows that links are not discovered only across datasets but also within the same dataset. The later accounts as de-duplication, and it could be seen that Orgref has the least duplicates (12 links found) while OpenAire has the most duplicates (14876 links found). Furthermore, it could be seen that in this linking scenario, more links are discovered between OpenAire and H2020. However, the number of discovered links between OpenAire and H2020 is more than the number of institute documented by H2020. This could be a negative impact of the duplicates within OpenAire. *Anyhow, it also looks promising as it shows a high potential in a better overlap, hence a better navigation between these two datasets.*

	OpenAire	Grid	OrgRef	H2020
OpenAire	34713 / 14876	30614 / 12518	16041 / 6978	26553 / 21291
Grid	30614 / 12518	16397 / 2847	25711 / 17395	2358 / 1464
OrgRef	16041 / 6978	25711 / 17395	7476 / 12	1130 / 534
H2020	26553 / 21291	2354 / 1464	1130 / 534	186 / 116

Table 1. Statistic of Links before (approximate similarity over organisation name) and after refinement (exact match over country name).

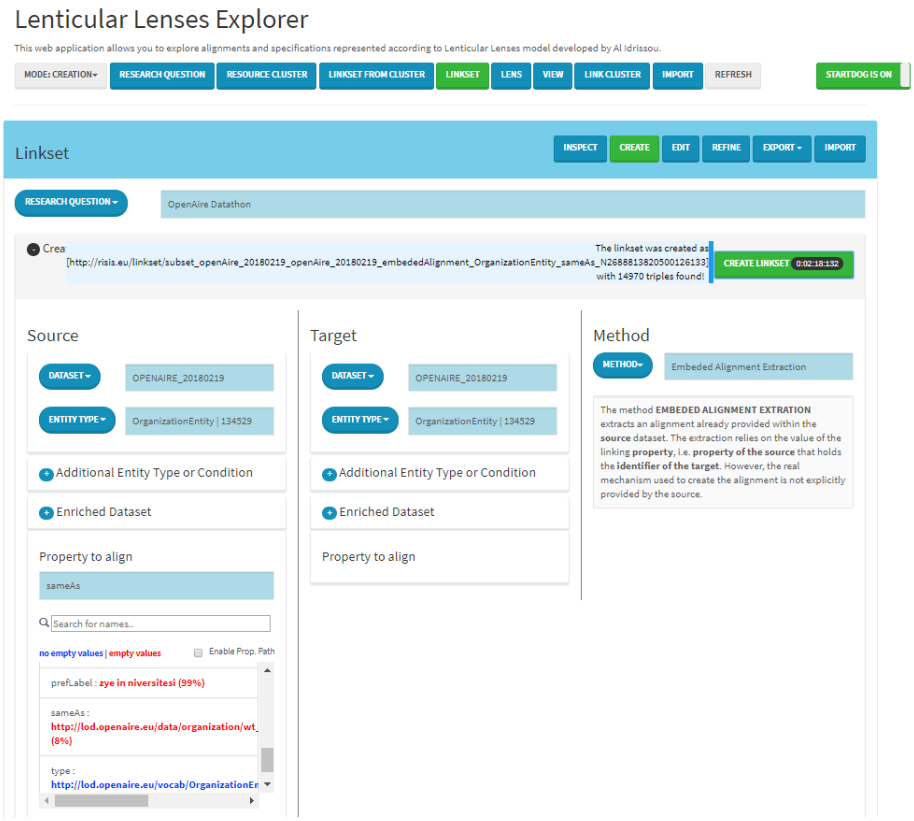

Fig. 5. Importing links using the Lenticular Lens

2.5 De-duplication

For the purpose of illustration, beside the de-duplication performed within each dataset, external duplicates could also be imported into the Lenticular Lens tool as displayed in Figure 5 where 14970 duplicates links where imported. Clustering

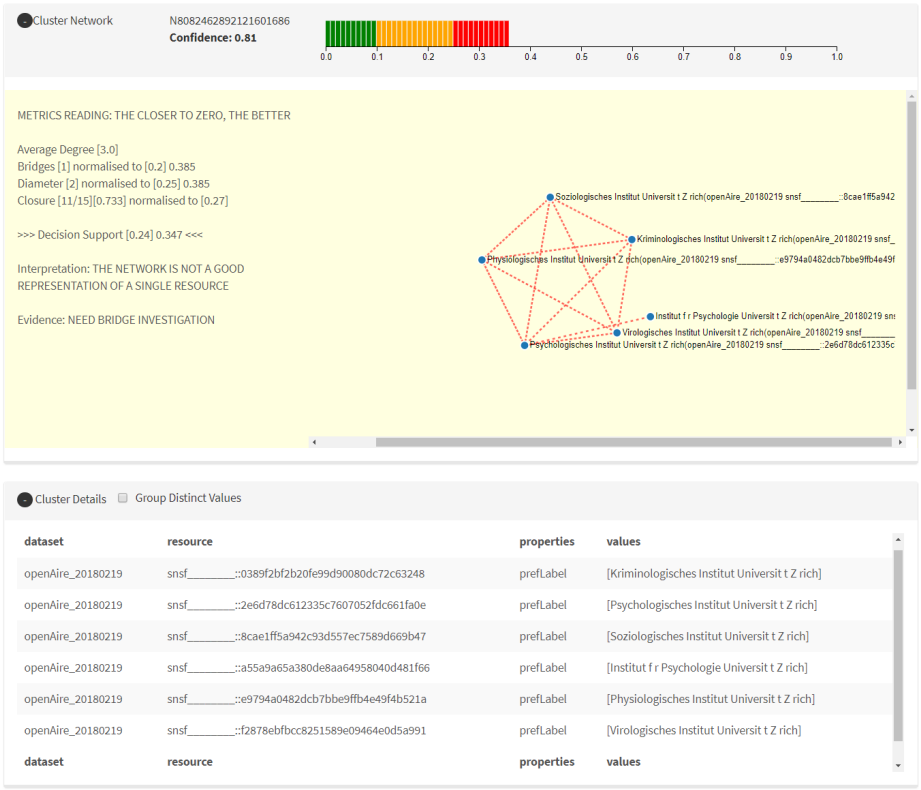

Fig. 6. Duplicate Detection

these links reveals a total of 14970 identity-link-networks. The detail as size and number of networks found is as follow: (3, 831) (4, 242) (9, 92) (6, 36) (7, 17) (8, 7) (9, 1) and (10, 8). Each and everyone of these networks form a full-mesh network, suggesting 14970 perfect identity-networks. However, a close look at the identity-link-networks of size six to ten revealed some errors in the de-duplicated set of links that come with the dataset. *This suggest another task of data-quality improvement that needs to be undertaken by the OpenAIRE data team*.

3 Browsing and Visualizing Data

After enriching and interlinking OpenAIRE, we embedded the enriched dataset in our SMS (Semantically Mapping Science) platform[2] available at `http://sms.risis.eu`. SMS allows end-users to browse data scattered over multiple datasets in an integrated way. The faceted browsing environment[3] provides a set of dynamic rendering components for users to mix and visualize the results while browsing data. User interaction are then translated to a set of SPARQL queries which are run in the background. For example, Figure 7) shows a screen-shot of the SMS browsing environment used to address the following research

question: *Find the list of educational institute in Germany that attended EU projects started in 2018.*

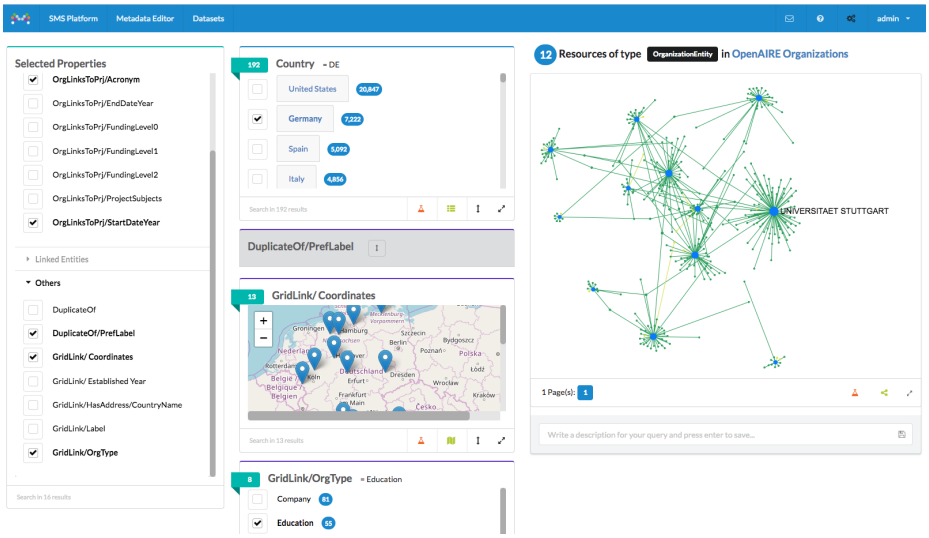

Fig. 7. Browsing the enriched OpenAIRE dataset on the SMS platform.

The WYSIWYQ (What You See Is What You Query) user interface in SMS enables users to save their queries in the system and turn them to interactive UIs once needed. You can reproduce the browsing environment used to answer the above query by visiting the `http://sms.risis.eu/browse/http%3A%2F%2Fsms.risis.eu%2Fd1502897809/http%3A%2F%2Fld-r.org%2Fconfigurations%2Fstate1519743335`. Listing 1.8 shows an example query generated by the SMS faceted browser.

Using user-friendly tools to search, browse and visualize OpenAIRE data enables more end-users to benefit from the potentials of OpenAIRE data for research and innovation studies. It also allows data curators to have a better understanding on the quality of data (for instance by looking at the geographical distribution of data and finding the lack of coverage in certain areas.).

```
PREFIX gridV: <http://www.grid.ac/ontology/>
PREFIX risisV: <http://risis.eu/alignment/predicate/>
PREFIX oaV: <http://lod.openaire.eu/vocab/>

SELECT DISTINCT ?s  ?title  ?ldr_ap1_Acronym ?ldr_ap2_PrefLabel WHERE {
  GRAPH <https://www.openaire.eu> {
    {
      SELECT DISTINCT ?s WHERE {
        GRAPH <https://www.openaire.eu> {
          ?s rdf:type oaV:OrganizationEntity .
          ?s oaVcountryCode> ?v1.
          ?s risisV:gridLink/<http://rdf.risis.eu/orgs/orgType> ?v2.
          ?s oaVorgLinksToPrj>/oaV:startDateYear ?v6.
          FILTER (str(?v1) IN ("""DE""") &&
          iri(?v2) IN gridV:Education) && str(?v6) IN ("""2018"""))
          ?s oaV:orgLinksToPrj/oaV:acronym ?ldr_ap1_Acronym .
          ?s risisV:duplicateOf/oaV:prefLabel ?ldr_ap2_PrefLabel .
        }
      }
      LIMIT 20 OFFSET 0
    }
    ?s oaV:orgLinksToPrj/oaV:acronym ?ldr_ap1_Acronym .
    ?s risisV:duplicateOf/oaV:prefLabel ?ldr_ap2_PrefLabel .
    OPTIONAL {
      ?s oaVprefLabel> ?title .
    }
  }
}
```

Listing 1.8. The SPARQL query generated by the SMS faceted browser

4 Conclusion

In our opinion, OpenAIRE quality still needs to be improved with regards to property homogeneity, de-duplication, a more descriptive URI schema. Using a human friendly interface to browse the OpenAIRE data can already facilitate the detection of quality issues. We successfully employed two tools namely Lenticular Lens and Linked Data Reactor to interlink, browse and define research questions on top of the OpenAIRE data.

This experiment tried to bring an idea of how OpenAIRE overlaps with the selected open data. We already interlinked OpenAIRE data with three relevant open dataset on organisations. Another major step has to be taken in the future is how to validate the automatically generated links. Once this step is done, this would be a great contribution to the link open data as one could freely and accurately navigate across datasets and answer more complex research questions.

References

1. A. K. Idrissou, R. Hoekstra, F. van Harmelen, A. Khalili, and P. van den Besselaar. Is my:sameAs the same as your:sameAs? In *Proceedings of the Knowledge Capture Conference on - K-CAP 2017*, pages 1–8, New York, New York, USA, 2017. ACM Press.
2. A. Khalili, P. V. den Besselaar, A. K. Idrissou, K. A. de Graaf, and F. van Harmelen. Semantically mapping science (SMS) platform. In *Proceedings of the First Workshop on Enabling Open Semantic Science co-located with 16th International Semantic*

Web Conference (ISWC 2017), Vienna, Austria, October 21st, 2017., pages 1–6, 2017.

3. A. Khalili, P. van Andel, P. van den Besselaar, and K. A. de Graaf. Fostering serendipitous knowledge discovery using an adaptive multigraph-based faceted browser. In *Proceedings of the Knowledge Capture Conference*, K-CAP 2017, pages 15:1–15:4, New York, NY, USA, 2017. ACM.