Linking Scholarly Datasets — the EOSC Perspective*

Marcin Wolski[1][0000−0001−5550−7739], Antoni Klorek[1], Cezary Mazurek[1][0000−0002−8715−9326], and Anna Kobusińska[2][0000−0002−3501−2840]

[1] Poznan Supercomputing and Networking Center, Poland
[2] Poznań University of Technology, Poland

Abstract. A plethora of publicly available, open scholarly data has paved the way for many applications and advanced analytics on science. However, a single dataset often contains incomplete or inconsistent records, significantly hindering its use in real-world scenarios. To address this problem, we propose a framework that allows linking scientific datasets.The resulting connections can increase the credibility of information about a given entity and serve as a link between different scholarly graphs. The outcome of this work will be used in the European Open Science Cloud (EOSC) as a base for introducing new recommendation features.

Keywords: Big scholarly datasets · Entity linking · EOSC · Microsoft Academic Graph · OpenAIRE Graph.

1 Introduction

In recent years, with the development of many novel research fields, and the rapid growth of digital publishing, the term Big Scholarly Data (BSD) has been coined and become increasingly popular [23]. Big Scholarly Data was introduced to reflect the size, diversity and complexity of the vast amounts of data associated with scholarly undertakings, such as journal papers, conference proceedings, degree theses, books, patents, presentation slides, and experimental data from research projects [8, 23]. These data collections have millions of co-authors, papers, citations, figures and tables, and massive scale-related data produced by scholarly networks and digital libraries. The use of the BSD has gained immense importance lately, particularly with the advent of multi-disciplinary research projects, which use BSD to discover research collaboration, expert finder and recommender systems [9].

One of the ongoing efforts to deliver a virtual, distributed research data repository and related services is the EOSC - European Open Science Cloud. The EOSC resources comprise outcomes of research efforts, such as published papers. As a primary dataset of scientific resources for the EOSC [1], the OpenAIRE

* Supported by the *EOSC Future* project, co-funded by the EU Horizon 2020 Programme INFRAEOSC-03-2020/101017536

ICCS Camera Ready Version 2023
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-35995-8_43

Research Graph (OARG) dataset was employed. However, they also contain software and e-infrastructure services, such as computational power, storage, and network to support scientific experiments [7], which sets EOSC apart from other environments.

In this paper, we propose the OARGLink framework that enhances the OARG dataset with the content provided by open scholarly datasets to improve the accessibility and composability of the EOSC resources. AMiner, Microsoft Academic Graph (MAG) [6], and DBLP are examples of the datasets which are widely used for various research purposes [23]. Unfortunately, linking the OARG and open records is a demanding task. Despite these datasets offering information about the corresponding scientific resources, such information usually varies from one dataset to another. For instance, one dataset can provide only the basic information about scholarly resources, such as title, authors, published year etc. At the same time, the other dataset can store such information as the number of citations or the content of the abstract. As a result, the enrichment of the EOSC resources will further enhance its features, such as the intelligent discovery of the EOSC resources and smart recommendations [20] and, in the future, introduce new capabilities such as suggesting research collaborations. By combining resources of EOSC and open datasets, we will make the corresponding resources more credible and enable the usage of scientific networks in EOSC services.

The structure of this paper is as follows. In Section 2, the EOSC and the chosen open scholarly datasets are introduced. Section 3 presents the related work. Next, in Section 4, we outline a problem of entity linking, and discuss the general idea of the proposed OARGLink framework for connecting the OARG and MAG datasets. Furthermore, Section 5 presents the data flow in the proposed solution, and Section 6 describes how the data processing algorithms were customized based on a few trials with sample databases. Section 7, presents results that describe the efficiency of the implemented solution, known limitations and the method of results verification. Next, Section 8 provides the discussion on the proposed solution. Finally, Section 9 presents conclusions and future work.

2 The EOSC and open scholarly datasets

2.1 The EOSC and the EOSC Future project

The EOSC is an ongoing effort to connect the existing European e-infrastructures, integrate cloud solutions and provide a coherent point of access to various public and commercial services in the field of academic research [4]. The EOSC also is a key acronym for various European R&D projects related to Open Science on the national, regional and European levels. These projects aim at engaging researchers to utilize a web of scientific resources that are open and Findable, Accessible, Interoperable, and Reusable (FAIR). The regional EOSC initiatives and aggregators contribute to the data collection on the users of available resources (e.g. the EOSC-Nordic). Along with the variety of stakeholders of the EOSC ecosystem, several different roles address the needs of these stakeholders, such as research infrastructures, technology providers, service providers, data managers, researchers, policymakers (including funders), and everyday users [3].

The vision of the EOSC Future project is to deliver an operational EOSC Platform with an integrated environment consisting of data, professionally provided services, available research products and infrastructure that will be accessed and used by the European research community. At the heart of the EOSC Platform are the users who provide and exploit EOSC resources. Users include researchers, resource providers, research and technology enablers, trainers and policymakers. The EOSC Portal is considered a universal open hub for EOSC users. The EOSC portal offers public and commercial e-infrastructure services, including distributed and cloud computing resources and the EOSC Research Products. The number of currently registered services in the EOSC Portal is almost 400[3]. The OARG is now the provider of Research Products for the EOSC Platform, delivering around 150 million publications, datasets, research-supporting software, configurations and other products[4]. The estimated size of the target population of the EOSC is roughly 2 million, including 1.7m researchers covering all major fields of science and levels of seniority [2].

2.2 Scholarly datasets

In general, scholarly datasets are published by organisations which own and develop the platforms for scholarly data management. Such platforms are usually exposed to the scientific community as academic search engines or digital libraries [23]. Their primary function is to crawl documents from the Web, extract useful information from them, and then store and index them in a coherent repository [12, 22]. Typically, there are various scholarly applications implemented on top of these repositories.

Table 1. Basic features of a few of the selected open scholarly datasets

Dataset	**Entities**
OpenAIRE Research Graph (OARG)	140m publications, 16m research data, 286k research software items, 175k organizations
Open Academic Graph (OAG)	MAG: 240m publications, 243m authors & AMiner: 184m publications, 113m authors
AminerNetwork	2m publications (8m citations), 1.7m authors, 4.3m collaboration relationships
DBLP	6.22m publications
OpenAlex	249m publications, 103m authors, 226k venues, 108k institutions

Table 1 presents the basic statistics of the selected datasets. In addition to search engines and digital libraries, many datasets with scientific and social networks have been published so far. Scientific and academic social networks, such as Mendeley, Academia, LinkedIn or ResearchGate, are being utilized by

[3] https://marketplace.eosc-portal.eu/services

[4] https://graph.openaire.eu/

ICCS Camera Ready Version 2023
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-35995-8_43

the users to enhance their knowledge about other users in the networks and find new collaborators for their current or future projects [11].

As of the end of 2021, Microsoft stopped developing MAG. The end of this vast and prospering data source raised concerns, so the nonprofit organisation OurResearch developed the OpenAlex (OA) [14], a fully open catalogue of the global research system. The primary OpenAlex source of data is MAG, but the developers make use of other sources such as Crossref or Pubmed. The OA data is constantly updated using available repositories, databases, and internet scrapers. Despite storing information, the authors created a website and API that researchers may easily use to obtain desired records. To the best of our knowledge, the OA is the most extensive open–source platform currently available.

Every dataset has a different graph structure of its entities. OpenAIRE shares entities about publications, datasets, software and other research products. Information about authors is kept inside these tables. OAG, made from Aminer and MAG, contains, despite a publication entity with essential authors' data, a separate table with authors' details, a table with affiliations data and a table with venues. MAG also has different tables with publications and authors. Moreover, this dataset contains many other entities such as venues, journals, affiliations, conferences, etc. AminerNetwork shares publication data, and there is additional information about authors in a separate entity. What is more, there is an entity that contains connections between coauthors. DBLP contains entities about publications, authors, journals and conferences. In turn, MAG provides publications, authors, venues, institutions and concepts. These entities have connections between them all.

3 Related Work

In the existing solutions, a single dataset is often linked with another one, creating a new dataset that combines entities from these sources. To link two large scholarly datasets, MAG and AMiner, the scientists developed a framework to create the Open Academic Graph (OAG) [24]. The resulting dataset contains data from sources and links between publications, authors and venues. This work aimed to build a large, open-knowledge, linked entity graph. Another example is a dataset containing publications from various scientific disciplines built on the base of the arXiv.org resources [16]. The primary purpose of this effort was to link the publications to the MAG to enrich the metadata information. As a result, a freely available dataset with annotated and extracted citations was proposed to be used by researchers and practitioners. Furthermore, a dedicated database model, based on the ResearchGate (RG) data source, was prepared for implementing collaborators finding system [15]. The model comprises two parts: one for designing a consistent, collaborator-finding system and the other that contains different relations between the pair of users. RG dataset has been collected from Jan. 2019 to April 2019 and includes raw data of 3980 RG users.

In all mentioned systems and approaches to linking scholarly data, the difficulty lies in dealing with entity matching, which aims to identify records that

DOI: 10.1007/978-3-031-35995-8_43

belong to the same entity. Entity matching (record linkage) is very important for data integration and cleaning, e.g. removing duplicates. The difficulty arises mainly from heterogeneous and poor-quality data. The existing algorithms for linking entities can be divided into two categories: classification-based and rule-based [10]. The first one tries to determine if two records are the same by classifying them into the same entity or assigning them the same label. Here, Machine Learning and Deep Learning methods are often used. The models learn patterns from training data and then apply them to solve the given problem of entity matching. The challenge in this method is the preparation of high-quality training examples. The ruled-based category is the deterministic approach to link entities. By setting the number of rules, the records are compared against them and if the rules are fulfilled, the records are classified as the same entity. The difficulties are in setting rules not too strict and not too loose to link, preferably all the same records and not omitting the correct ones. Also, many rule-based approaches cannot cope with missing values and require numerous preprocessing steps.

4 The OARGLink framework — problem statement and general idea

In this paper, *the OARGLink framework* dedicated to linking OARG records (publications) with the records from open scholarly datasets was proposed. The framework aims to create a coherent database of the pair (identical) publications[5].

There are some common problems that usually have to be faced during the integration of heterogeneous databases [19]. Among them, based on analyzing the scholarly datasets' structure and content, we identified problems that can also impact the process of connecting the OARG and open scholarly data. Firstly, there is high dimensional scholarly data that, among other issues, impose computational challenges. Moreover, some of the scholarly data is incomplete, inaccurate or unreasonable. Finally, a problem with data integrity occurs. Unique identifiers between records of two scholarly datasets often do not exist. Moreover, different scholarly datasets have other data structures and fields, meaning they are encoded differently. Moreover, the problem of linking scholarly data usually deals with differences in spelling, formatting and proper fulfilment of the metadata of a given record. The existing datasets often contain incomplete or noisy records (resulting from, e.g. encoding issues in the source PDFs). There does not also exist standard schema used for storing key attributes among the datasets (such as the author's name can be saved as a full name or as abbreviations). Furthermore, connecting scholarly datasets consisting of millions of records always poses computational challenges. As a result, the process of linking scholarly datasets is a non-trivial and challenging task [13].

Taking the above arguments into account, we focused our efforts on the identification of the same publications among different datasets and thereby

[5] https://gitlab.pcss.pl/eosc-extra/scholarlydata

DOI: 10.1007/978-3-031-35995-8_43

omitting other existing entities such as venues, journals, etc. A journal article can be identified by the journal name, volume, issue number, and starting page number. However, for a large fraction of open-access scholarly papers crawled from the Web, such information is usually not available. Empirically, a paper entity can be uniquely identified by four header fields: title, authors, year, and venue, in which the venue is a conference or a journal name [17].

In the paper, the Record Linkage Method for combining data was proposed to find the connections between OARG publications with the records from the MAG scholarly dataset. The proposed solution can also be considered a relaxed Deterministic (Exact) Matching Method [18]. The publication's year and the number of authors must be exact; one title has to be a part of the second title, and at least one author has to be matched precisely.

For the entity linking among OARG and MAG, we took advantage of a dedicated dump of OARG with a pre–selected set of resources for EOSC (OARG–EOSC). This dump contains 1.7m papers, 2.63m datasets, 223k software and 19.6k other research products[6]. The DOI in the OARG-EOSC dataset is present in 1558130 publications, which is 91.037% of records. As a result, taking this field as one to connect entities from different datasets would result in omitting many records already at the start. However, other fields of interest, i.e. publication year, title and authors, are present in every record so that they can be used to identify the same entity among datasets.

5 Data processing flow in the OARGLink

The proposed algorithm processes the scholarly data in three phases depicted in Figure 1. In the algorithm the following general approach for connecting entities in heterogeneous datasets is applied:

1. The fields that might be useful during connection are identified. By leaving only essential fields and eliminating the not needed ones, it is ensured that the next steps would not suffer from high-dimensional data.
2. Data cleaning process is applied — the records with essential fields missing are deleted; any data inconsistencies are identified.
3. Data are modified and unified, e.g. it is checked whether in both datasets a date is stored in the same type and order and changes are applied if needed.

Currently, the framework is implemented and evaluated for connecting the records of OARG with AminerNetwork and OARG with MAG.

PHASE 0: Data extraction and cleaning

The structure of the publications' metadata differs depending on the dataset, and the number of fields varies from publication to publication. However, some fields are usually shared among them and almost always present. The aim of Phase 0 is to leave only essential fields. These are the paper id, title, authors and publication date. If needed, adequate processing is performed to convert the above fields to be in the proper format:

[6] https://sandbox.zenodo.org/record/1094615#.Y_8k6tLMKRQ

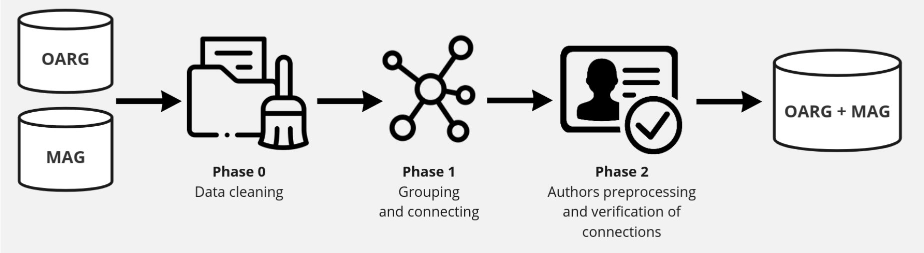

Fig. 1. Data processing in the OARGLink framework

- if the date is written fully, only the year is left,
- if the name and surname of an author are separated, e.g. by a semicolon, they are split into separate words,
- publications with no authors or year are erased

This cleaned data is saved in JSON format.

PHASE 1: Linking datasets by publications' title

In the first phase of the linking procedure, the following steps are performed:

1. The cleaned data is loaded from both datasets — OARG and the one to connect with it.
2. Dictionaries are created for each dataset — the publications in the dictionaries are grouped by the number of authors and the publication year. This grouping minimizes the number of needed comparisons between publications' titles. When the year or the number of authors of publications do not match, there is no point in comparing the titles of these records. Simultaneously, during the same iteration, titles are changed to lowercase.
3. Publications are linked by checking if the title from the external dataset is contained in the OARG publication's title (checking performed only between the same number of authors and year groups, as described in the second point).
4. Ids of connected pairs are stored and saved as JSON files.

PHASE 2: Authors' preprocessing and verification of connected pairs by authors' comparison

The pair of records obtained in Phase 1 links the publications with the same titles, number of authors, and year of publications. These pairs might not always be exact, as authors may differ. Consequently, the second phase verifies and performs additional processing of the authors by performing the following steps:

1. The linked pairs of publications are loaded.
2. Authors from these publications are extracted.

3. Authors are preprocessed in the following way: first, the authors are concatenated into a single string, and the punctuation characters, double spaces (if present), and space from the end of a string (if present) are deleted. Next, all characters are lowercase, and a string is split into separate words. Finally, single–letter words are deleted, and words are sorted alphabetically.
4. Pairs of related publications are selected in which at least one first or last name matches. The proposed solution does not verify whether all authors match because, even with complex preprocessing, many typos or special foreign characters influence the obtained results. After the analysis, this method allows for the maximisation of the number of connected pairs and, at the same time, minimizes wrongly connected ones.
5. Obtained pairs of ids are saved in the JSON file.

The Python programming language was the primary technology to implement *the OARGLink*. Built-in libraries made it possible to implement compelling data reading, cleaning and manipulation methods. We used dictionaries as the primary data structures to store publications' information.

6 The OARGLink customization

Before connecting the complete datasets of the OARG and the MAG, the data processing algorithms of *the OARGLink* were customized by making a few trials with data samples.

Firstly, an attempt to connect the OpenAIRE with the AminerNetwork was made. Next, the same approach was used to connect the OpenAIRE with the MAG dataset taken as a part of the OAG. After linking each pair of dataset samples, the results were evaluated against the accuracy (i.e. checked if the connected results contain identical publications) and the processing time. Then, the basic algorithm was modified, and the experiments were re-run to check the results.

6.1 Linking samples of the OARG and AminerNetwork

The existing OpenAIRE dump file containing 189000 publications was used during the linking process. All publications were taken as a sample to conduct the trial run. The full Aminer dump contains over 2 million publications — a sample of half a million was used to perform the linking. AminerNetwork publication data was downloaded from the Aminer website[7].

After the completion of two phases, the proposed algorithm found 60 connections. Every connected record was checked manually, and a single mistake was identified. An incorrectly connected pair is the following:

– Publication from OARG:
Title: *"Introduction"*, Authors: *Salim Yusuf, Michael Gent, Genell Knatterud, Michael Terrin*, Year: *2002*.

[7] https://www.aminer.org/aminernetwork

DOI: 10.1007/978-3-031-35995-8_43

- Publication from AminerNetwork:
  Title: *"Introduction"*,Authors: *Michael Bauer, Gene Hoffnagle, Howard Johnson, Gabriel Silberman*,Year: *2002*.

The mistake originated from the fact that despite having the same titles, number of authors and published year, there are authors with the same names in both entities. Although the found problem rarely occurs, the proposed algorithm classified these publications as the same. Summing up this trial, the algorithm achieved 98.3% precision. However, some publications might have been omitted, e.g. titles had typos, the same publications were stored under slightly different titles, etc.

The linking procedure ran with an Intel Core i5 CPU with 8GB RAM and lasted less than 25 seconds.

6.2 Linking samples of the OpenAIRE and MAG

The EOSC OpenAIRE dump contains 1.7m publications. The whole dataset was included in the experiment as it will give us correct judgment for future work.

The MAG was chosen because it contains far more records than the Aminer part of the OAG, so looking from a higher perspective, it should result in more linkings and better results and may be used as a separate data source for the EOSC portal in the future. Ten batches of 100k publications were sampled from the MAG (about 240m publications in total) for experimental purposes and connected one by one with the OARG.

Five hundred and seven connections were found, and only one error occurred, so the algorithm achieved 99.8% precision. The error is as follows:

- Publication from OARG:
  Title: *"Approximate Solution Of Some Mixed Boundary Value Problems Of The Generalized Theory Of Couple-Stress Thermo-Elasticity"*, Authors: *Chumburidze, Manana, Lekveishvili, David*, Year: *2014*.
- Publication from MAG:
  Title: *"Mixed Boundary Value Problems"*, Authors: *Fioralba Cakoni, David Colton*, Year: *2014*.

The error occurred because the second title is included in the first one, the number of authors is the same, the published year is identical, and there are authors with the same name in both examples.

The OARG connection time, with a single 100k batch and the Intel Core i5 CPU configuration with 8GB RAM was less than 90 seconds.

7 The overall results and their verification

After setting up the optimal configuration for the proposed *the OARGLink* framework, we ran it on sample data of OARG-EOSC and MAG. This intermediate step determined the estimated time needed to connect the complete

ICCS Camera Ready Version 2023
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-35995-8_43

datasets. Also, the additional experiments did ensure that the algorithm could cope with the more significant amount of data and possible differences in the metadata. The results from the experiment and estimations are presented in Table 2.

Table 2 presents the estimated number of connections and the time needed to accomplish the whole processing. These estimations were calculated based on experimental results and with the assumption of linear growth of these two factors. However, the sample of data taken to the experiment may slightly differ from the whole dataset. As a result, the final number of obtained connections, as well as the time needed, may vary. It's worth mentioning that times were measured on a device with Intel Core i5 CPU and 8GB RAM and with fixed sizes of batches with publications. So, by presenting the estimations, we also assumed that the final operations would be performed on the same or similar machine and using batches of the same sizes. Eventually, we proceeded to the last step to make connections between EOSC OARG (1.7m publications) and full MAG (240m publications). As we calculated from the estimations, it should take about 60 hours and around 121k connected pairs should be obtained.

Table 2. Experiments results and estimations

	Results		**Estimations**
	OARG (1.7m) with MAG (100k)	OARG (1.7m) with 10x MAG (100k)	OARG (1.7m) with whole MAG (240m)
# of final connections	50	507	121k
time assuming batches 1.7m and 100k	90s	15min	60h

The linking process started with downloading the whole MAG dataset to the local computer. Compressed files had almost 150GB; this number increased over three times after uncompressing. Next, all data had to be split into smaller files. We divided each file, which contained over 5m publications, into ones that had 500k of them. Then, the cleaning was done, and the connecting procedure started. The results are presented in Table 3.

Table 3. Full connection results

	OARG	**MAG**	**OARG+MAG**
# of records	1.7m	240m	496168

As a result, over 29% of OARG's publications were connected with their MAG's equivalents (around four times more than we had estimated). The procedure took 4 days and ran on two machines with 4-core processors and 8GB RAM. The processing time was 50% longer than we had estimated. In some batches, there were far more publications to compare. It was caused by the vast number of the same publications' year or the same number of authors among the documents.

Table 4. The presence of DOI in the connected records

# of connected publications (OARG+MAG)	496168
# of OARG's publications with DOI	458438
# of MAG's publications with DOI	433208

We took advantage of DOI (Digital Object Identifier) to verify if the obtained results were correct and whether the connected pairs of the papers from both datasets referred to the same article. The basic statistics about related publications and the presence of DOI are presented in Table 4.

The proposed verification procedure comprised the following steps:

1. The pairs of connected records in which both publications possess DOIs were found
2. DOIs in these pairs were compared and it was calculated how many of them are exact and how many differ.
3. A subset of connected pairs where DOI is different was taken and the related publications were manually analyzed

The results show that 92.4% of connected publications on the OARG's side have DOI. Moreover, in 87.3% of connected publications on the MAG's side, DOI is present.

Table 5 presents the found records (OARG and MAG), which possess DOI in both (corresponding) publications. The connected records were thoroughly analyzed, where DOI was different, although the proposed framework had merged the corresponding publications together. We found out, among others, that both publications can be the same in the title, authors, published date and content, but they were published at different conferences or in other journals. As a result, the document's formatting or structure may differ, or some additional footnotes may be present. So, the articles are the same but published under different DOIs. Nevertheless, the obtained results reveal that 82.5% of connected pairs simultaneously have DOI in both publications. Also, in 93% of these pairs, DOIs are exact, and in 7% of these pairs, DOIs differ.

Table 5. The presence of DOI in BOTH of the connected publications

# of pairs with DOI in both publications	409575
# of pairs where DOIs are exact	380933
# of pairs with different DOIs	28642

8 Discussion

The quality of the entity matching task is a trade-off between precision and coverage. In the case of EOSC, the target was to achieve high precision (we accept the risk of omitting some theoretically existing connections between records). This is because when the target datasets are significant, a small fraction of false positives in the sample data may eventually lead to many false matching. In turn, the wrongly connected records would lead to degrading user experience (measured in EOSC as, e.g. a mean consumer feedback satisfaction score for the presented publication records).

The most apparent method to connect entities with publications in different datasets is to use a universal identifier. Currently, the DOI is recognized as Persistent Identifier (PID) for publications [5]. DOI is a unique and never-changing string assigned to online (journal) articles, books and other works, and it is usually present in the datasets. But, since DOI is not always present in every record of analysed datasets, we argue that such an identifier cannot be used unambiguously for entity linking in scholarly systems.

Our chosen method for connecting common records from different scholarly datasets follows rule-based patterns for integrating heterogeneous databases. Among the considered approaches for implementing the entity linking system, a solution proposed by [24] was the most promising to be re-used. The authors implemented an algorithm based on locality-sensitive hashing (LSH) and Convolution Neural Networks (CNN) to connect publications. However, considering the structure and the format of the publications' metadata in OARG (i.e. lack of some fields in OARG, e.g. venue in the publication's metadata), the considered solution would not work out of the box. The algorithm would not be precise, and the results wouldn't be satisfied without the significant changes in the existing source codes and additional efforts to construct the artificial training set. Another problem indicated while analyzing the considered approach was related to the time consumption and the extensive need for computational resources of this solution. A robust GPU-based hardware infrastructure would be needed to run the proposed algorithm efficiently.

Furthermore, it is still a challenge to evaluate the entity linking algorithms formally. In the paper [24], the authors manually label venue training data (around 1000 records) and construct artificial complex training data for papers and venues. A similar approach was used by [21], who matched CiteSeerX against DBLP and obtained 236 matching pairs. They also noticed the need to apply data cleaning tasks to filter out highly unhealthy data on top of a supervised approach.

Taking the above into account, we argue that the most effective method for linking Big Scholarly Datasets for EOSC is to combine two different approaches: (1) a rule-based approach based on the precise data matching (which includes preprocessing phase with data cleaning) and (2) the DOI pairing to identify the same publications. As a result, a more significant number of connections are meant to be identified. What is essential intersecting these two approaches should minimize potential errors. The algorithm used in the proposed framework found more connections than the method of connecting publications by only existing DOIs. Incorrect DOIs or their absence doesn't determine the accuracy of our solution.

The results of our experimental work with item-based recommendations based on connected records (MAG-OARG and OpenAlex-OARG) clearly show its usefulness in supporting EOSC-related scenarios. In particular, the OpenAlex dataset has multiple fields present in the metadata of a publication, so adding them to OARG records and using them in the EOSC portal brings additional benefits. Data about citations, references and related works are especially interesting here. This information could enhance the network of interconnections between publications and positively influence the quality of recommendations. Other data, e.g. venues, publishers, and different identifiers, could enrich the presented information at the portal and allow for a deeper understanding of a given record. As a result, EOSC users will see more information about the record and be able to use the portal as the primary source of their research.

The previously mentioned poor quality of data in source datasets may raise concerns if missing fields, such as some authors or wrongly stated year of publication, will result in omitting some matching which should be matched as equal entities. The methods for dealing with some data inconsistencies were stated in the framework's description, but unfortunately, it is impossible to consider all possible errors in input data. The OARGLink mainly focuses on correctness, not returning false positive examples. As a result, precision should be high. Conversely, omitted examples may be classified as false negatives so that the recall measure may be slightly lower.

9 Summary and future work

This paper studied the vital problem of linking large-scale scholarly datasets. Connecting heterogeneous scholarly datasets is not a trivial task. Firstly, they consist of various entities in academic graphs, such as author, paper, or organisation entities. Moreover, the structure of familiar entities such as papers or authors differs. Secondly, observing ambiguous values of the same attributes is widespread, such as authors' names or publications' titles. Finally, the scale of datasets is large, usually with millions of entities.

As a part of our work, we analyzed a few scholarly datasets to obtain more information about their content and structure. On top of the OARG, which is the primary data source for EOSC, we selected a few others: the OAG, which is one of the most significant sources of publications and other scholarly information currently available on the Internet; AminerNetwork as its structure is well defined and potentially ideal for the experimental purposes, and eventually DBLP, which constitutes a popular among scientists open scholarly dataset. We narrowed down the work to papers and authors. Firstly, we evaluated the implemented solution with samples of data. Experimental results show that our solution OARGLink achieved at least 98.3% precision for connecting OARG with Aminer and 99.8%

ICCS Camera Ready Version 2023
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-35995-8_43

accuracy for connecting OARG with MAG. The coverage hasn't been evaluated as a part of this work.

In the final work of connecting OARG for EOSC (1.7 million publications) with MAG (240 million publications), we achieved almost half a million connected pairs, over 29% of the used OARG dataset. The results indicate that the database with linked records can be used as a valuable source of information for further usage in EOSC.

The method proposed in this paper can be reused for linking other scholarly repositories. For example, social network databases can be linked and explored to resolve more accurate bindings between users by taking advantage of specific academic relationships. Furthermore, we have also recognized the need for applying a similar approach in digital libraries. Digital libraries often combine various sources with digital books from different providers into coherent records. In such a case, multiple editions of the same book, sometimes having other titles, must be merged into one entry and presented to the reader.

The primary goal for future work is to use the connected pair of records as a base for implementing new recommendation scenarios for EOSC. On top of that, there are many other potential directions for further development on *the OARGLink framework*. For example, the framework can be adapted to a large-scale computational environment (i.e. adapted to using scalable technologies such as Apache Spark and Hadoop) and run to connect the full OARG (140 million) and MAG datasets. In the future, additional experiments will be conducted to perform an extended evaluation of the implemented method, such as experiments with noisy data to determine the actual coverage.

References

1. Almeida, A.V.d., Borges, M.M., Roque, L.: The european open science cloud: A new challenge for europe. In: Proceedings of the 5th International Conference on Technological Ecosystems for Enhancing Multiculturality. TEEM 2017, CM (2017)
2. Anca Hienola (ICOS), John Shepherdson (CESSDA ERIC), B.W.C.: D5.2a eosc front-office requirements analysis. Tech. rep. (2022)
3. Barker, M., Manola, N., Gaillard, V., Kuchma, I., Lazzeri, E., Stoy, L., Piera, J.: Digital skills for fair and open science: Report from the eosc executive board skills and training working group (2021)
4. Budroni, P., Claude-Burgelman, J., Schouppe, M.: Architectures of knowledge: The european open science cloud. ABI-Technik **39**(2), 130–141 (2019)
5. Cousijn, H., Braukmann, R., Fenner, M., Ferguson, C., van Horik, R., Lammey, R., Meadows, A., Lambert, S.: Connected research: The potential of the pid graph. Patterns **2**(1), 100180 (2021)
6. Färber, M., Ao, L.: The microsoft academic knowledge graph enhanced: Author name disambiguation, publication classification, and embeddings. Quant. Sci. Stud. **3**(1), 51–98 (2022)
7. Ferrari, T., Scardaci, D., Andreozzi, S.: The open science commons for the european research area. Earth Observation Open Science and Innovation. ISSI Scientific Report Series **15**, 43–68 (2018)

8. Giles, C.L.: Scholarly big data: information extraction and data mining. In: Proceedings of the 22nd ACM international conference on Information & Knowledge Management. pp. 1–2 (2013)
9. Khan, S., Liu, X., Shakil, K., Alam, M.: A survey on scholarly data: From big data perspective. Information Processing & Management **53**, 923–944 (07 2017)
10. Kong, C., Gao, M., Xu, C., Qian, W., Zhou, A.: Entity matching across multiple heterogeneous data sources. In: Database Systems for Advanced Applications: 21st International Conference, DASFAA 2016, Dallas, TX, USA, April 16-19, 2016, Proceedings, Part I 21. pp. 133–146. Springer (2016)
11. Kong, X., Shi, Y., Yu, S., Liu, J., Xia, F.: Academic social networks: Modeling, analysis, mining and applications. Journal of Network and Computer Applications **132**, 86–103 (2019)
12. 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)
13. Nasar, Z., Jaffry, S.W., Malik, M.K.: Information extraction from scientific articles: a survey. Scientometrics **117**(3), 1931–1990 (2018)
14. Priem, J., Piwowar, H., Orr, R.: Openalex: A fully-open index of scholarly works, authors, venues, institutions, and concepts. (2022)
15. Roozbahani, Z., Rezaeenour, J., Shahrooei, R., Emamgholizadeh, H.: Presenting a dataset for collaborator recommending systems in academic social network: A case study on reseachgate. Journal of Data, Information and Management **3** (03 2021)
16. Saier, T., Faerber, M.: unarxive: a large scholarly data set with publications' full-text, annotated in-text citations, and links to metadata. Scientometrics **125** (03 2020)
17. Sefid, A., Wu, J., Allen, C.G., Zhao, J., Liu, L., Caragea, C., Mitra, P., Giles, C.L.: Cleaning noisy and heterogeneous metadata for record linking across scholarly big datasets. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 33, pp. 9601–9606 (2019)
18. Shlomo, N.: Overview of Data Linkage Methods for Policy Design and Evaluation, pp. 47–65. Springer International Publishing (2019)
19. Wang, L.: Heterogeneous data and big data analytics. Automatic Control and Information Sciences **3**(1), 8–15 (2017)
20. Wolski, M., Martyn, K., Walter, B.: A recommender system for eosc. challenges and possible solutions. In: International Conference on Research Challenges in Information Science. pp. 70–87 (2022)
21. Wu, J., Sefid, A., Ge, A.C., Giles, C.L.: A supervised learning approach to entity matching between scholarly big datasets. In: Proceedings of the Knowledge Capture Conference. pp. 1–4 (2017)
22. Wu, Z., Wu, J., Khabsa, M., Williams, K., Chen, H.H., Huang, W., Tuarob, S., Choudhury, S.R., Ororbia, A., Mitra, P., et al.: Towards building a scholarly big data platform: Challenges, lessons and opportunities. In: IEEE/ACM Joint Conference on Digital Libraries. pp. 117–126 (2014)
23. Xia, F., Wang, W., Bekele, T.M., Liu, H.: Big scholarly data: A survey. IEEE Transactions on Big Data **3**(1), 18–35 (2017)
24. Zhang, F., Liu, X., Tang, J., Dong, Y., Yao, P., Zhang, J., Gu, X., Wang, Y., Shao, B., Li, R., Wang, K.: Oag: Toward linking large-scale heterogeneous entity graphs. In: Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. pp. 2585—-2595 (2019)