A Preliminary Assessment of the Article Deduplication Algorithm Used for the OpenAIRE Research Graph

Kleanthis Vichos[1], Michele De Bonis[2], Ilias Kanellos[1], Serafeim Chatzopoulos[1], Claudio Atzori[2], Natalia Manola[3], Paolo Manghi[2,3] and Thanasis Vergoulis[1]

[1] *IMSI, ATHENA RC, 6 Artemidos St, Marousi, 15125, Greece*
[2] *Istituto di Scienza e Tecnologie dell'Informazione, National Research Council, Pisa, Italy*
[3] *OpenAIRE, Athens, Greece*

Abstract

In recent years, a large number of Scholarly Knowledge Graphs (SKGs) have been introduced in the literature. The communities behind these graphs strive to gather, clean, and integrate scholarly metadata from various sources to produce clean and easy-to-process knowledge graphs. In this context, a very important task of the respective cleaning and integration workflows is deduplication. In this paper, we briefly describe and evaluate the accuracy of the deduplication algorithm used for the OpenAIRE Research Graph. Our experiments show that the algorithm has an adequate performance producing a small number of false positives and an even smaller number of false negatives.

Keywords

deduplication, open science, scholarly data, knowledge graphs

1. Introduction

In recent years, large amounts of scholarly data have become openly available due to the increased popularity of the Open Science [1] initiatives. This abundance of scholarly content is really important since it catalyzes the creation and provision of several added value services that can facilitate scientific knowledge discovery, as well as research assessment and monitoring. In most cases, the scholarly content is published in the form of *Scholarly Knowledge Graphs (SKGs)*. Knowledge graphs are heterogeneous graphs (i.e., having multiple node and edge types) capable of representing the semantics of complex knowledge spaces; this makes them attractive for the case of scholarly data, since this domain consists of many entities (e.g., articles, researchers, venues, software, datasets) which are highly interconnected with different types of relationships.

Several SKGs have been produced in recent years either from the academic community (e.g., the OpenAIRE Research Graph [2], the Open Research Knowledge Graph [3]) or industry-driven

IRCDL 2022: 18th Italian Research Conference on Digital Libraries, February 24–25, 2022, Padova, Italy

kvichos@athenarc.gr (K. Vichos); michele.debonis@isti.cnr.it (M. De Bonis); ilias.kanellos@athenarc.gr (I. Kanellos); schatz@athenarc.gr (S. Chatzopoulos); claudio.atzori@isti.cnr.it (C. Atzori); natalia.manola@openaire.eu (N. Manola); paolo.manghi@openaire.eu (P. Manghi); vergoulis@athenarc.gr (T. Vergoulis)
https://schatzopoulos.github.io/ (S. Chatzopoulos); http://thanasis-vergoulis.com/ (T. Vergoulis)
0000-0002-8955-9489 (K. Vichos); 0000-0003-2347-6012 (M. De Bonis); 0000-0003-1714-5225 (S. Chatzopoulos); 0000-0001-9613-6639 (C. Atzori); 0000-0001-7291-3210 (P. Manghi); 0000-0003-0555-4128 (T. Vergoulis)

ones (e.g., the Microsoft Academic Graph [4]). Such initiatives strive to gather, clean, and integrate content from different and diverse data sources (e.g., libraries, publication repositories, publishers, etc) and assemble graphs whose nodes represent articles, datasets, researchers, etc. At the same time, scholarly content is inherently heterogeneous, comprising a variety of research object types and (meta-) data in diverse formats, curation levels, and even languages. In addition, best practices and standard procedures in research vary across disciplines, while the entities of interest are usually domain-specific. This heterogeneity in scholarly content is a major impediment to the acquisition, integration, and interlinking of content from different sources leading to disruptive duplication rates. Consequently, the developing teams of SKGs have implemented fully-fledged entity deduplication workflows for their needs.

In this work, we conduct a preliminary evaluation of the effectiveness of the deduplication process currently used for the creation and update of the OpenAIRE Research Graph [2], one of the most widely known community-driven SKGs. Although the current process (to which we refer as `fDup-2021`) is based on gDup, a framework that has been introduced in a previous work [5], there are no hitherto experiments to assess its accuracy (or the accuracy of any other instance of gDup). Apart from the assessment of the particular gDup instance, another contribution of our work is the creation of a new curated dataset that contains expert judgements regarding the equivalence (or not) of research objects. This dataset can be useful for assessing the accuracy of other instances of the gDup framework, but also as a set of expert validated equivalent objects, each having its unique digital object identifier (DOI).

2. Background & Related Work

2.1. Scholarly Knowledge Graphs (SKGs)

One of the most popular approaches for scientific knowledge representation is that of Scientific/Scholarly Knowledge Graphs (SKGs), many of which have been developed as industry-driven initiatives, such as the Web of Science (WoS) [6], Microsoft Academic Graph (MAG) [4], and Dimensions [7]. Among academic or non-profit initiatives, Crossref [8] is probably the largest source of scholarly metadata supporting 13 major content types (e.g., articles, datasets, peer reviews). The OpenAIRE Research Graph [2] encompasses scholarly metadata of a large variety and empowers the EOSC resource catalogue. Moreover, the Open Research Knowledge Graph (ORKG) [3] describes research papers in a structured manner. Finally, OurResearch has lately developed and released OpenAlex, a large scholarly dataset that attempts to cover the gap created by the discontinuation of MAG by the end of 2021.

2.2. The OpenAIRE Research Graph

The OpenAIRE infrastructure[1] is an initiative and Legal Entity whose purpose is to facilitate, foster and support Open Science in Europe. Among others, OpenAIRE supports the technical services that facilitate and monitor Open Science publishing trends. To this end, the OpenAIRE service infrastructure consists of metadata aggregation services and information inference services whose purpose is to populate the OpenAIRE Research Graph [2].

[1] https://www.openaire.eu/

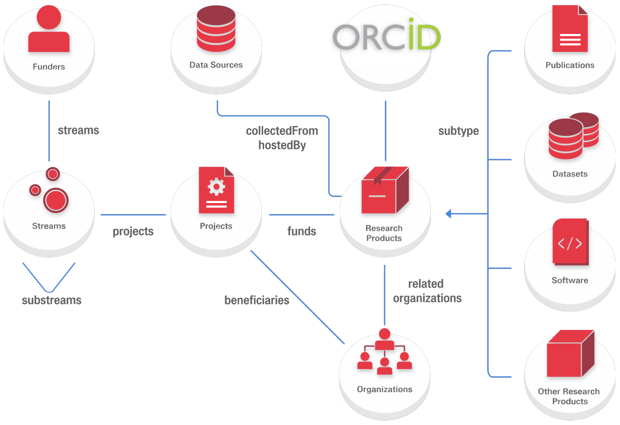


Figure 1: The OpenAIRE Research graph model.

The graph's data model is depicted in Figure 1 and its main entities are described below:

- *Research Products* represent the outcomes of research activities.
- *Organizations* correspond to companies or research institutions involved in projects, responsible for operating data sources or consisting the affiliations of Product creators.
- *Funders* (e.g. EC, Wellcome Trust) are agencies responsible for a list of Funding Streams.
- *Funding Streams* represent investments (funding actions) from Funders (e.g. FP7 or H2020).
- *Projects* are research projects funded by a Funding Stream of a Funder.
- *Data Sources* are the resources used to collect metadata for the graph objects.

On top of the graph, OpenAIRE offers various services, such as a search and exploration portal and a number of dashboards (the Research Community Dashboard [9], the Funder Dashboard, etc.). Deduplication of products and organizations is therefore crucial to deliver meaningful statistics to the users. In addition, since all data are open by design, it is crucial for any added value services built on top of OpenAIRE's data, as well.

2.3. OpenAIRE's deduplication framework

The entire deduplication process used to materialize the final version of the OpenAIRE Research Graph is managed by the gDup framework [5, 10]. gDup is an integrated, scalable, general-purpose system for entity deduplication over big SKGs. It supports practitioners with the typical functionalities needed to realize a full entity deduplication workflow over a generic input graph. The deduplication workflow of gDup (Figure 2) consists of the following main phases:

- *Collection import*: it loads the collection to be processed, by defining a set of labels (custom names) and values (extracted from the original entity).
- *Candidate identification*: a preliminary grouping stage to divide the input space into smaller clusters, leveraging the object's DOI and title.
- *Duplicates identification*: it involves intra-cluster pair-wise comparisons between entities; the number of comparisons is reduced using a sliding window mechanism after ordering the entities so that potentially equivalent entities will be in the same window.

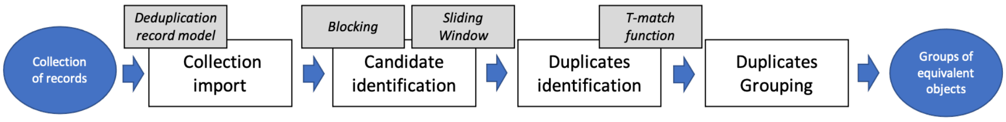


Figure 2: OpenAIRE's deduplication workflow.

- *Duplicates grouping*: the final operation that creates representative objects and persistent identifiers for the newly created records.

Of course, the similarity function used to compare pairs of entities should be able to capture record equivalence and should be flexible and configurable for every different entity type. In gDup, the similarity function was defined by a weighted sum of the similarity scores between entity attributes, while a set of conditions that implement early exits in the comparison have been defined. To make this task smarter, gDup was extended to a new framework, called fDup, which introduces a decision tree mechanism which enables early exits and different similarity match strategies based on intermediate results of the comparisons between entity attributes. The mechanism considers the Levenshtein distance (normalized to obtain a value between 0 - very different - and 1 - identical) of the entity titles to determine if two entities are equivalent or not. A threshold depending on the number of common IDs is applied to the similarity score: in case the entities have common IDs, the threshold on the score is lower than the other case (0.9 vs. 0.99). This means that a higher similarity score of the title is needed if two entities do not share IDs. In the last case, a further comparison on the title version (i.e. numbers in the title string) and the author lists is performed to guarantee the correct result. If the entities have different versions in the title and different sizes of author lists, the early exit tells that there is no need to compute the Levenshtein distance as the two entities are considered to be different. This specific deduplication configuration is currently used as the OpenAIRE's deduplication algorithm and it will be referred in the following as `fDup-2021`.

3. Evaluation

In this section we elaborate on our assessment process to evaluate the accuracy of the `fDup-2021` algorithm in identifying DOIs that correspond to equivalent objects (i.e., closely related entities). Our experiments can be divided into two groups: first we compare `fDup-2021`'s output to sets of known DOI aliases (Section 3.1) and, then, we further investigate those `fDup-2021`'s equivalent objects that do not correspond to known aliases (Section 3.2).

3.1. Quantifying `fDup-2021`'s false negatives using DOI aliases

To perform a preliminary analysis on the quality of the output of `fDup-2021`, in our first experiment, we leveraged information from doi.org's REST API[2] regarding *DOI aliases*. Reporting DOI aliases is the default mechanism for registrants of DOIs to report duplicate DOIs[3]. Since not

[2] In particular, we gathered data from the `HS_ALIAS` field provided by the API.

[3] DOI aliases: https://www.crossref.org/documentation/reports/conflict-report/#00243 (accessed Dec 20th 2021)

all duplicates are reported by the respective registrants, DOI aliases cannot be used to quantify the false positives that DOI deduplication algorithms produce. However, any DOIs that have been reported as aliases are guaranteed to refer to equivalent objects, hence they can be used as a ground truth to quantify false negatives and this is how we leveraged them in this experiment.

Since gathering the aliases for all distinct DOIs in the OpenAIRE Research Graph (>120M) is a time-consuming process (especially, if the implemented process makes responsible usage of the API respecting request limits), we decided to restrict our analysis only to those DOIs that are reported to have at least one equivalent DOI according to the fDup-2021 algorithm (a more complete evaluation is planned for an extension of the current work). Our snapshot of the graph (produced on October 26th, 2021) contained 112 216 333 distinct deduplicated entries (i.e., distinct OpenAIRE IDs) in total, 5 885 861 of which contained at least two equivalent DOIs. Using doi.org's REST API we gathered all the aliases of the respective distinct DOIs (14 427 982 in total) and generated the corresponding groups of DOI aliases. It should be noted that 6 185 of the DOIs of the graph were problematic, i.e., *unresolvable* at the time of data gathering.[4]

We, then, compared these sets of aliases with the sets of equivalent entries provided by fDup-2021. During this comparison, we ignored all unresolvable DOIs (i.e., the analysis was performed using the rest). A summarisation of the results is presented in Table 1.

Table 1
Statistics for the various types of fDup-2021's deduplicated entries.

Types of deduplicated entries	# of entries
Completely matching sets of aliases (true positives)	32 476
Involving unreported aliases (false negatives)	1 100
Not compliant to aliases (false positives or missing aliases)	5 852 174
Containing only unresolvable DOIs	111

In particular, we found that a lot of fDup-2021's deduplicated entries (32 476) were completely compliant with the list of known aliases (i.e., were confirmed true positives). Also 1 100 of the entries could be considered as false negatives, since they did not contain even one known alias. However, the vast majority of the deduplicated entries were containing groups of known aliases (hence, implying missing aliases or false positives). Finally, only a negligible number of the deduplicated entries contained only unresolvable DOIs.

It is evident that fDup-2021 produces a very small number of confirmed false negatives (they account for less than 0.02% of the examined entries). In addition, it seems that fDup-2021 identifies a very large number of equivalent DOIs which are not reported as aliases in doi.org. In the next section, we attempt to determine whether this can be mainly attributed to a large number of false positives, or if a huge number of equivalent DOIs are not reported as aliases.

3.2. Investigating reported equivalent objects with no DOI aliases

In this experiment, we further investigated fDup-2021's deduplicated entries that involve sets of DOIs that have not been reported as aliases (line 3 in Table 1). Our main objective was to get

[4] It should be noted that although OpenAIRE aggregates data from multiple resources (as discussed in Section 2.2), it is not responsible to guarantee that all collected DOIs are resolvable.

insights about the scale of false positives in fDup-2021's output. The only way to fulfil this objective is to have expert judgements on the sets of equivalent objects that the deduplication algorithm produces. The experts use DOI-related metadata and the corresponding content (e.g., the manuscript in case of publications) and provide judgements regarding the correctness of the algorithm output (i.e., if the reported DOIs correspond to equivalent objects or not). We followed this approach assigning the respective task to 4 experts (computer engineers, two of them PhDs). However, since the manual inspection is time consuming and the data to be examined is immense (more than 5.8M entries), we opted to assign a sample of 300 randomly selected entries per expert, resulting in a dataset of 1 200 entries. Each expert was given the task of assigning each set of equivalent DOIs with one of 8 predetermined class labels (Table 2).

Table 2
Annotation classes.

Name	Interpretation	Judgement
AMBIGUOUS	At least one DOIs is invalid (no metadata are available).	N/A
DELETED-DUPLICATES	DOIs once pointing to the same research object, currently deleted.	TRUE
MULTI-PUBLISHED	Article published in more than one locations (full or abstract).	
VERSIONS	Multiple versions of the same research object (e.g. pre-prints, post-prints etc).	
ERRONEOUS	Unrelated set of objects.	FALSE
PAPER-EXTENSIONS	Extended version of a conference paper in a journal.	
PART-OF-A-GROUP	Multiple parts of the same research object (e.g., photos of the same collection).	
SUPPLEMENTARY	Article and its supplementary material (including errata).	

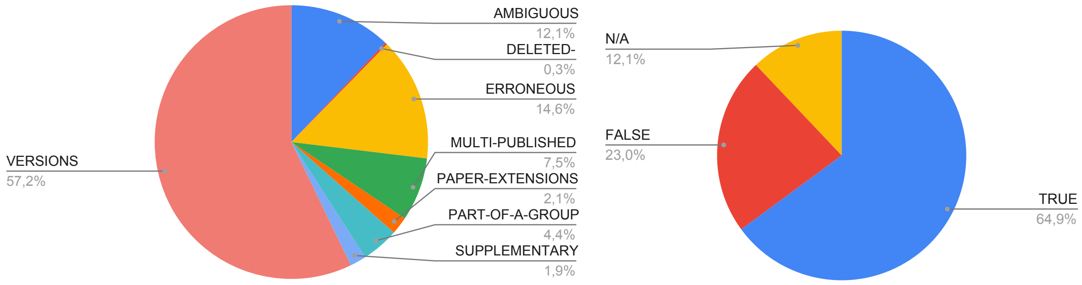
Each of the classes has particular semantics, explained in the 'Interpretation' column. These semantics determine whether the objects in the respective group are equivalent or not ('Judgement' column). The dataset that has been generated by the aforementioned process, was made openly available on Zenodo[5] under CC-BY license.

Figure 3a illustrates the proportion of deduplicated entries that have been annotated with each of the classes, while Figure 3b summarises the proportion of true and false positives; due to the existence of the AMBIGUOUS class, there were also a lot of entries for which it was not possible to provide a judgement (denoted by 'N/A' in Figure 3). It is evident that the majority of deduplicated entries (64.9%) produced by fDup-2021 are correct; most of them contain different versions of the same object and extensions of older works. The false positives, on the other hand, correspond to a significantly smaller percentage (23%).

4. Discussion

Our main findings can be summarized as follows: deduplication algorthims are useful and bring significant added-value; this is highlighted by the fact that manually curated collections (like the DOI aliases) fail to report a large number of true positive equivalent objects. Specifically, for more than 5.8M sets of equivalent objects (according to fDup-2021) there is no reported DOI alias. Furthermore, fDup-2021 has adequate results, producing a lot of useful true positives; however, there is room for improvements since the proportion of false positives is relatively large. This is

[5] https://doi.org/10.5281/zenodo.5794406



(a) Deduplicated entries annotated with the classes of Table 2.

(b) True and false positives.

Figure 3: Expert evaluation results.

expected since fDup-2021 is fairly inclusive (favoring false positives instead of false negatives). Some common errors were related to the fact that the algorithm is oblivious of the author lists, grouping together articles of different authors having the same title. Another common mistake was that it could not distinguish between main articles and their supplementary materials.

It is worth noting that this work contains a preliminary analysis on this subject. Our analysis has important limitations. For instance, for efficiency reasons, we used only deduplicated entries with at least two equivalent objects for our analysis. To alleviate this issue, more time is required to collect the DOI alias info from doi.org's REST API; we plan it as an extension of the current work along with extending our ground truth that currently consists of 1 200 expert judgements.

5. Conclusions & Future Work

In this work, we conducted a preliminary assessment on the effectiveness of fDup-2021, the deduplication process used for the creation and update of the OpenAIRE Research Graph. The main contributions of our work were the following: we explain why DOI deduplication algorithms are important; we introduce a ground truth dataset that can be used for the assessment of deduplication processes for Scholarly Knowledge Graphs (SKGs) and leveraged it to perform a first assessment of fDup-2021, providing insights on its major weaknesses. In the future we plan to perform more thorough experiments to confirm the results of the current study and we aim to design an improved instance of the gDup framework that alleviates all identified issues.

Acknowledgments

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 101017452.

References

[1] B. Fecher, S. Friesike, Open science: one term, five schools of thought, in: Opening science, Springer, Cham, 2014, pp. 17–47.

[2] P. Manghi, N. Houssos, M. Mikulicic, B. Jörg, The data model of the openaire scientific communication e-infrastructure, in: Research Conference on Metadata and Semantic Research, Springer, 2012, pp. 168–180.

[3] M. Y. Jaradeh, A. Oelen, K. E. Farfar, M. Prinz, J. D'Souza, G. Kismihók, M. Stocker, S. Auer, Open research knowledge graph: next generation infrastructure for semantic scholarly knowledge, in: Proceedings of the 10th International Conference on Knowledge Capture, 2019, pp. 243–246.

[4] K. Wang, Z. Shen, C. Huang, C.-H. Wu, Y. Dong, A. Kanakia, Microsoft Academic Graph: When experts are not enough, QSS 1 (2020) 396–413.

[5] P. Manghi, C. Atzori, M. De Bonis, A. Bardi, Entity deduplication in big data graphs for scholarly communication, Data Technologies and Applications (2020).

[6] P. Mongeon, A. Paul-Hus, The journal coverage of web of science and scopus: a comparative analysis, Scientometrics 106 (2016) 213–228.

[7] D. W. Hook, S. J. Porter, C. Herzog, Dimensions: building context for search and evaluation, Frontiers in Research Metrics and Analytics 3 (2018) 23.

[8] G. Hendricks, D. Tkaczyk, J. Lin, P. Feeney, Crossref: The sustainable source of community-owned scholarly metadata 1 (2020) 414–427. doi:10.1162/qss_a_00022.

[9] M. Baglioni, A. Bardi, A. Kokogiannaki, P. Manghi, K. Iatropoulou, P. Principe, A. Vieira, L. H. Nielsen, H. Dimitropoulos, I. Foufoulas, et al., The openaire research community dashboard: on blending scientific workflows and scientific publishing, in: International Conference on Theory and Practice of Digital Libraries, Springer, 2019, pp. 56–69.

[10] C. Atzori, P. Manghi, A. Bardi, Gdup: De-duplication of scholarly communication big graphs, in: 2018 IEEE/ACM 5th International Conference on Big Data Computing Applications and Technologies (BDCAT), IEEE, 2018, pp. 142–151.