2018 IEEE 14th International Conference on e-Science

De-duplicating the OpenAIRE Scholarly Communication Big Graph

Claudio Atzori, Paolo Manghi, Alessia Bardi
Istituto di Scienza e Tecnologie dell'Informazione "A. Faedo" - CNR
Via Moruzzi 1, Pisa, Italy
Email: {name.surname}@isti.cnr.it

***Abstract*—The OpenAIRE infrastructure populates a scholarly communication big graph interlinking metadata objects of publications, datasets, software, organizations, funders, and projects. In order to de-duplicate this graph, OpenAIRE has developed GDup , an integrated, scalable, general-purpose system for entity deduplication over big information graphs. GDup offers functionalities to realize a fully-fledged entity deduplication workflow over a generic input graph, inclusive of Ground Truth support, end-user feedback, and strategies for identifying and merging duplicates to obtain an output disambiguated graph.**

***Keywords*—deduplication, graph, big data, scholarly communication, OpenAIRE**

I. Introduction

A large number of online services offer today access to very large information graphs obtained as integration of distributed data sources. Their intent is typically that of providing integrated and enhanced access to such sources by applying a set of harmonisation and enrichment processes to the original data, e.g. harmonization of properties, identification of links between objects, named entity recognition (e.g. author identifiers). Examples in the scholarly communication domain are the Google Scholar graph[1], the Microsoft Academic graph[2], and the OpenAIRE scholarly information graph.[3]

This work focuses on one of such graphs, maintained and publicly provided by the technical services of the OpenAIRE infrastructure [1]. OpenAIRE is funded by the European Commission (and soon to become a Legal Entity) and its purpose is to facilitate, foster, support, and monitor Open Science in Europe. The infrastructure has been operational for almost a decade and successful in linking people, ideas and resources for the free flow, access, sharing, and re-use of research outcomes. On the one hand, OpenAIRE manages and enables an open and participatory network of people willing to identify the commons and forums required to foster and implement Open Science policies and practices in Europe and globally. On the other hand, it supports the technical services required to facilitate and monitor Open Science publishing trends and research impact across geographic and discipline boundaries. Such services collect bibliographic citation metadata from various Internet sources (e.g. libraries, publication repositories, publishers, author directories) and populate a "big graph" whose main entities are organizations, results (publications, datasets, software, other products), funders, projects, and data sources [2]. Due to the heterogeneity and overlap of the original sources, which often keep publications, datasets, and other information relative to the same authors and organizations, this graph suffers from disruptive duplication rates, which jeopardize Open Science monitoring, and therefore require adequate countermeasures.

In order to tackle this issue, starting from the experiences and solutions for duplicate identification in big data collections, the OpenAIRE infrastructure has designed and developed production services specifically addressing the problem of *entity deduplication in big information graphs*. By *big* we mean that duplicate identification over the objects of such entity types require parallel-oriented approaches to scale up to arbitrary numbers and still perform in reasonable time. By *entity deduplication* we mean the combined process of *duplicate identification* and *graph disambiguation*: duplicate identification has the aim of efficiently identifying pairs of equivalent objects of the same entity type; graph disambiguation has the goal of removing the duplication anomaly from the graph, while semantically preserving the topology of the graph, i.e. a combination of merging equivalent objects and re-distributing their relationships.

This poster presents GDup, an integrated, scalable, general-purpose system for entity deduplication over big graphs (GDup software [3]) developed in the context of the OpenAIRE service infrastructure.

II. GDup: deduplicating scholarly communication graphs

GDup is intended to support data curators with the out-of-the-box functionalities they require to support a fully-fledged entity deduplication workflow over an generic input graph, inclusive of "ground truth" support, end-user feedback, and strategies for identifying and merging duplicates to obtain an output disambiguated graph. Its aim is to address the general lack of tools capable of addressing a complete graph deduplication workflow, from the graph input phase to the materialisation of the disambiguated graph, enhanced by end user feedback and supported by ground truth. As such, GDup is not about better recall/precision for given deduplication problems, but rather about provision of tools enabling data curators to concentrate on modeling and customizing their deduplication solutions without bothering about the extra conceptual and technical challenges that such task necessarily imply. The architecture of system is depicted in Figure 1, whose main functional areas support an end-to-end workflow enabling data curators at:

[1]Google Scholar: http://scholar.google.com
[2]Microsoft Academic Graph https://academic.microsoft.com
[3]OpenAIRE Graph http://api.openaire.com

978-1-5386-9156-4/18/$31.00 ©2018 IEEE
DOI 10.1109/eScience.2018.00104

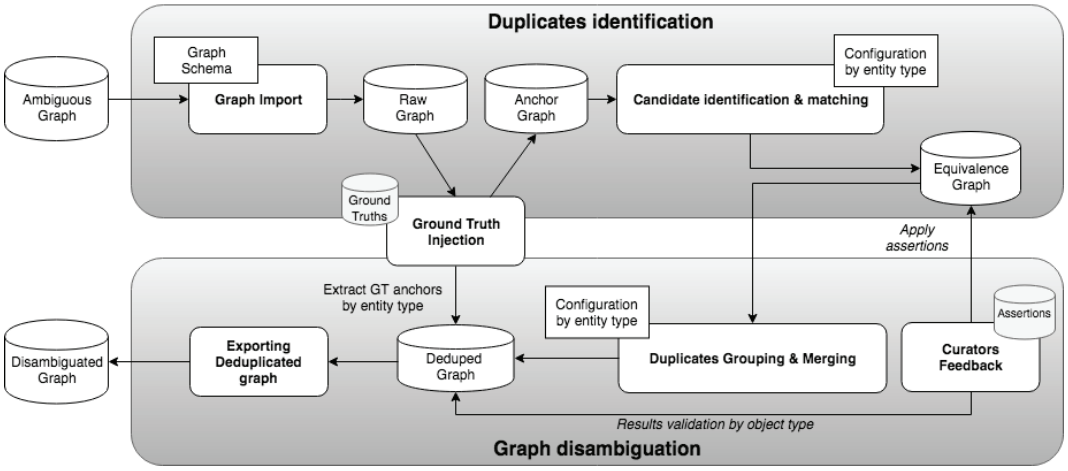

Fig. 1. GDup : architecture of de-duplication workflow

1) Importing their graph in the system;
2) Configuring for each entity type the relative duplicate identification "configurations";
3) Managing Ground Truth generation and injection;
4) Configuring graph disambiguation strategies;
5) Supporting data curators at manually fixing the results of deduplication;
6) Exporting a disambiguated graph, i.e. devoid of duplicate nodes.

GDup implements these functionalities by assembling known Open Source technologies (GDup Release 1.0 [3]) to support scalability of size, parallel processing, flexibility of models, and efficient bulk read and write operations. Such conditions excluded the adoption of traditional graph databases, oriented to efficient graph traversal functionalities [4] and pushed OpenAIRE services towards an Apache HBase [5]. [4]. HBase is based on the Hadoop framework [6], enabling large scale distributed data processing and analytics based on the Map Reduce programming model [7], [8].

GDup is today a production-ready service of the OpenAIRE infrastructure [9].[5] The infrastructure populates a scholarly communication big graph whose goal is to support monitoring of Open Science trends and research impact for funders, institutions, and researchers in specific disciplines. GDup is used to deduplicate publications, datasets, software entities, and organization entities to ensure sensible statistics are delivered. The graph counts ∼150Mi publications, datasets, and software entities, with ∼40Mi links between them.

III. Ongoing and Future work

GDup operates in the production system of the OpenAIRE infrastructure. However, its developments are ongoing to (i) make it a fully user-friendly product, i.e. completion of data curators GUI, and (ii) address further functional scenarios, e.g. deduplication by crowd-sourcing by delegating to a set of experts the addition of assertions to clean deduplication results and build ground truth. Among other developments, the team is devising a version of GDup resting on Apache Spark GraphX.[6]

[4]HBase - https://hbase.apache.org

[5]OpenAIRE infrastructure, http://www.openaire.eu

[6]Apache Spark GraphX, https://spark.apache.org/graphx/

Acknowledgment

This research was co-funded by the EC OpenAIRE2020 project (grant 643410, call H2020-EINFRA-2014-1) and EC OpenAIRE-Advance project (grant 777541, call H2020-EINFRA-2017-1).

References

[1] P. Manghi, N. Manola, W. Horstmann, and D. Peters, "An infrastructure for managing ec funded research output-the openaire project," *The Grey Journal (TGJ): An International Journal on Grey Literature*, vol. 6, no. 1, 2010.

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

[3] C. Atzori and P. Manghi, "gdup: a big graph entity deduplication system - Release 1.0," Feb. 2017. [Online]. Available: https://doi.org/10.5281/zenodo.292980

[4] M. A. Rodriguez and P. Neubauer, "The graph traversal pattern," *arXiv preprint arXiv:1004.1001*, 2010.

[5] L. George, *HBase: the definitive guide*. " O'Reilly Media, Inc.", 2011.

[6] T. White, *Hadoop: The definitive guide*. " O'Reilly Media, Inc.", 2012.

[7] J. Dean and S. Ghemawat, "Mapreduce: simplified data processing on large clusters," *Communications of the ACM*, vol. 51, no. 1, pp. 107–113, 2008.

[8] K.-H. Lee, Y.-J. Lee, H. Choi, Y. D. Chung, and B. Moon, "Parallel data processing with mapreduce: a survey," *AcM sIGMoD Record*, vol. 40, no. 4, pp. 11–20, 2012.

[9] P. Manghi, L. Bolikowski, N. Manold, J. Schirrwagen, and T. Smith, "OpenAIREplus: the European Scholarly Communication Data Infrastructure," *D-Lib Magazine*, vol. 18, no. 9/10, sep 2012. [Online]. Available: http://dx.doi.org/10.1045/september2012-manghi