De-duplicating the OpenAIRE Scholarly Communication Big Graph

Claudio Atzori, Paolo Manghi, Alessia Bardi

Institute of Information Science and Technologies, Italian National Research Council (ISTI-CNR), Pisa
{name.surname}@isti.cnr.it

An academic graph aggregating all information required to deliver monitoring tools

The scholarly graph is obtained as continuous aggregation of bibliographic metadata records originating from a variable set of information systems (repositories, publishers, funder databases) with heterogeneous and duplicated content. Main entities of the graph are organizations, results (literature, datasets, software, other products), funders, projects, and data sources.
The graph counts ~26Mi result entities, 2,5Mi projects, with~40Mi links between them.

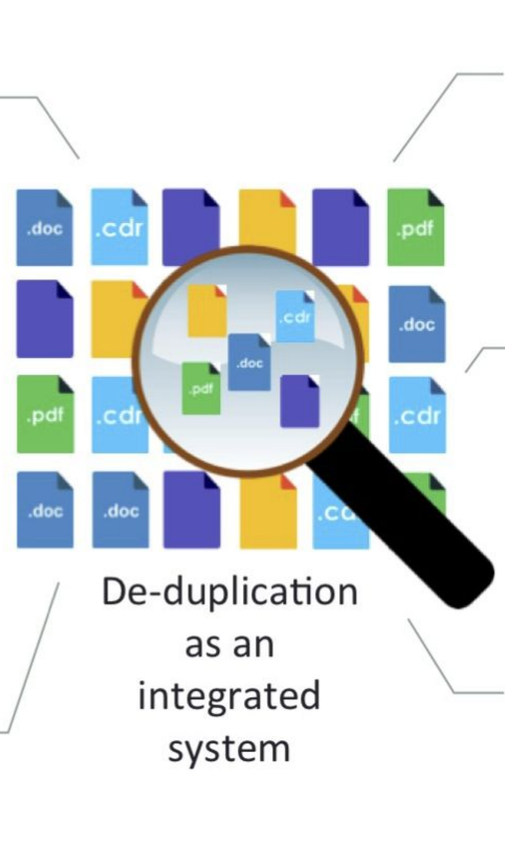
Title	Authors	Date
OpenAIREplus - OpenAIRE APIs for third party services. D8.6	Manghi et all	2012-06-12
OpenAIREplus - OpenAIRE APIs for third party services. D8.4	Manghi, P.;	2012-12-06
OpenAIREplus - OpenAIRE APIs for third party services. D8.6	Manghi Paolo	NA

Graph de-duplication: challenges & requirements

Graph orientation
Handling consequences of duplicates identification: **merge** and **redistribution of relationships.**

Ground Truth optimization
Reuse "certified" de-duplication outputs to reduce computational cost.

Scalability
Processing large graphs shouldn't be an issue.



Experiment-driven
Tooling should support experiment-driven processing: repeating and refining results

General-purposeness
Tooling should be customizable and configurable

Data curators feedback
Automation is correct up to a given degree of approximation. Domain experts might be able to refine the results.

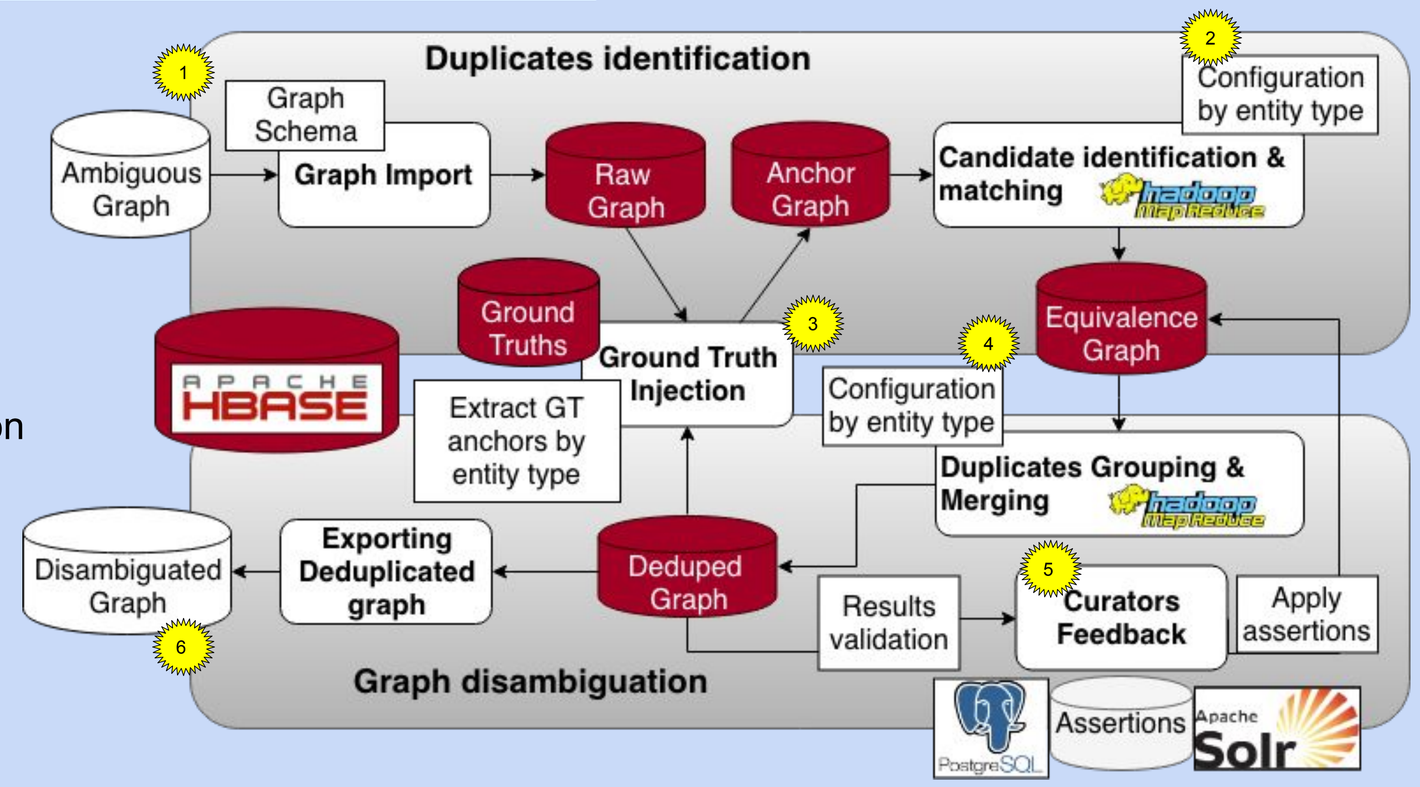
Solution proposed: GDup

GDup is an integrated, scalable, general-purpose system for entity de-duplication over big graphs.
GDup supports data curators with out of the box functionalities they require to support an end-to-end entity deduplication workflow over a generic input graph.

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 implies.

End-to-end workflow enabling data curators at:

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



Results

- GDup Open Source Software: https: //doi.org/10.5281/zenodo.292980
- GDup is today a production service (TRL9) of the OpenAIRE infrastructure
- GDup is used to de-duplicate literature, datasets, software and organisation entities to ensure sensible statistics are delivered by the OpenAIRE infrastructure.

Ongoing & Future work

- Make it a fully user-friendly product, i.e. completion of data curators GUI
- Address further functional scenarios, e.g. crowd-sourcing deduplication by delegating to a set of experts the addition of assertions to clean deduplication results and build ground truth
- Apache **Spark** to implement candidate identification and matching phases
- Apache **GraphX** to implement the graph disambiguation phase

Acknowledgments

This work is partially supported by the European Commission as part of the projects
- OpenAIRE2020 (H2020-EINFRA-2014-1, Grant Agreement 643410)
- OpenAIRE-Advance (H2020-EINFRA-2017-1, Grant Agreement 777541)