

OpenAIRE's DOIBoost - Boosting CrossRef for Research

Sandro La Bruzzo¹, Paolo Manghi¹, Andrea Mannocci²

ISTI-CNR, Pisa, Italy

Knowledge Media Institute, Open University, UK



ISTITUTO DI SCIENZA E TECNOLOGIE
DELL'INFORMAZIONE "A. FAEDO"



IRCDL 2019
31 Jan-1 Feb, Pisa, Italy



Outline

- Research problem
- DOIBoost
 - Inputs
 - Schema
 - Methodology
 - Deployment
 - Results
- Conclusion



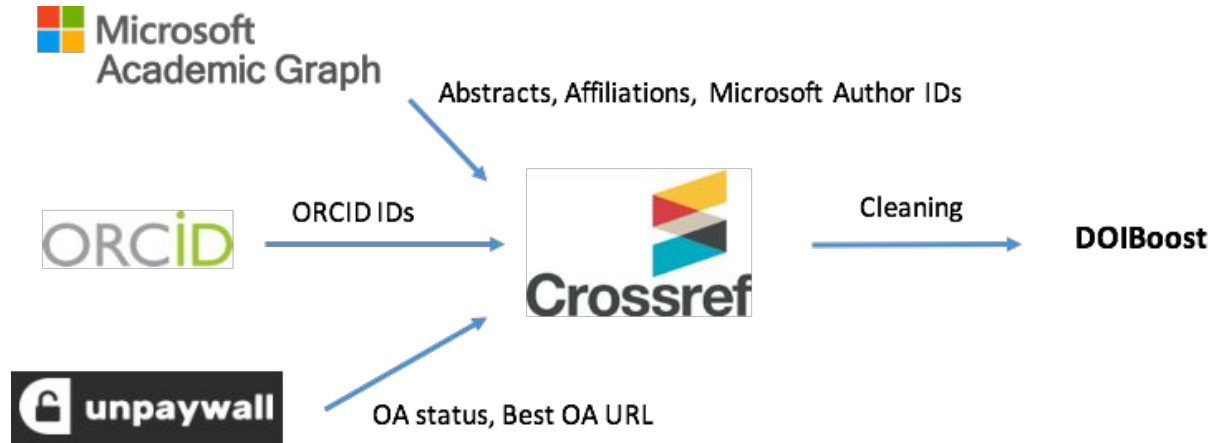
Research problem



- **Working with scholarly bibliographic (meta)data is generally a pain**
 - Paywalled/proprietary data (Elsevier, WoS, Springer, etc.)
 - In-house dataset construction
- **The latter has some typical issues**
 - information is often
 - scattered across diverse, freely accessible, online resources
 - (buried in PDFs)
 - integration problems to produce custom datasets
 - duplication of efforts, re-inventing the wheel day in day out
 - typically infringes open science best practices (transparency, reproducibility, documentability, etc.)



DOIBoost: the idea



Get the best of what's around and release it for people to use



Input: CrossRef

- Pivotal role in scholarly communication
- Mediator between publishers of scientific literature and consumers
- Publishers
 - publish scientific literature, mint a DOI from Crossref
 - push into the system a complete bibliographic record
- Consumers
 - retrieve bibliographic record from API Rest
 - retrieve entire metadata collection
- Size of Dump of CrossRef: 250GB (Nov '18)
- CC-BY 4.0 license





Input: UnpayWall

- Neat, tiny, little service (try the extension on your browser)
- Maps DOIs with the best Open Access URLs to papers (if available)
- Size: 6Gb Compressed Json
- CC-BY 4.0 license





Input: ORCID

- Assigns persistent identifiers to researchers
- De-facto standard worldwide
- Allows researcher to populate a publicly accessible curriculum, inclusive of article DOIs
- Gathers many more associations between articles in Crossref and ORCID IDs than Crossref is actually collecting from publishers.
- Size: 32 GB compressed XML
- CC0 license

ORCID



Input: Microsoft Academic Graph

- Uses AI-powered machine readers to process all documents discovered by Bing crawler
- Extract scholarly entities and their relationships to form a knowledge base
- MAG links to DOIs can enrich Crossref with extra information, e.g. author identifiers, affiliation identifiers, abstracts.
- Distributed through Microsoft Azure
- Size: 120GB (relevant DB tables dumped as TSV files)
- ODC-BY license



DOIBoost: Schema

- The schema includes a set of Crossref properties and a set of properties that can be integrated from other sources
 - Identifiers of authors
 - Affiliations of authors
 - Dates
 - Abstracts
 - Instances of the DOI work
 - Record quality report
- Each of these “inheritable” properties is equipped with a provenance field:
 - { Crossref | MAG | UnpayWall | Orcid }



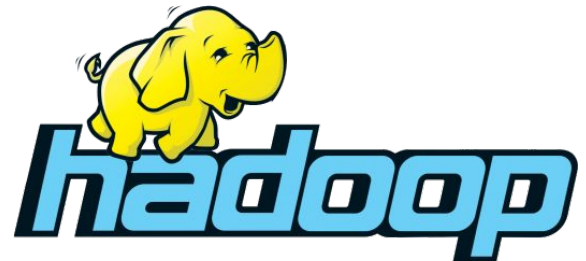
DOIBoost: Schema

```
1 {
2   "title": "My Title", ~
3   "authors": [ ~
4     { ~
5       "given": "Marco", ~
6       "family": "Rossi", ~
7       "fullname": "Marco Rossi", ~
8       "identifiers": [ ~
9         { ~
10          "schema": "ORCID", ~
11          "value": "https://.../0000-0002-3337-2025", ~
12          "provenance": "ORCID" ~
13        }, ~
14        { ~
15          "schema": "MAG ID", ~
16          "value": "https://.../1278293695", ~
17          "provenance": "MAG" ~
18        } ~
19      ], ~
20      "affiliations": [ ~
21        { ~
22          "value": "My Affiliation Name", ~
23          "official-page": "www.affiliation.org", ~
24          "identifiers": [ ~
25            { ~
26              "schema": "grid.ac", ~
27              "value": "https://.../grid.12345.a" ~
28            }, ~
29            { ~
30              "schema": "microsoftID", ~
31              "value": "https://.../4213412341" ~
32            }, ~
33            { ~
34              "schema": "wikipedia", ~
35              "value": "https://wiki/my_affiliation" ~
36            } ~
37          ], ~
38          "provenance": "MAG" ~
39        } ~
40      ], ~
41    }, ~
42    { ~
43      "given": "Giuseppe", ~
44      "family": "Trovato", ~
45      "fullname": "Giuseppe Trovato", ~
46      "identifiers": [ ~
47      ], ~
48    }, ~
49    "affiliations": [ ~
50    ], ~
51  ], ~
52 } ~
53 }
```

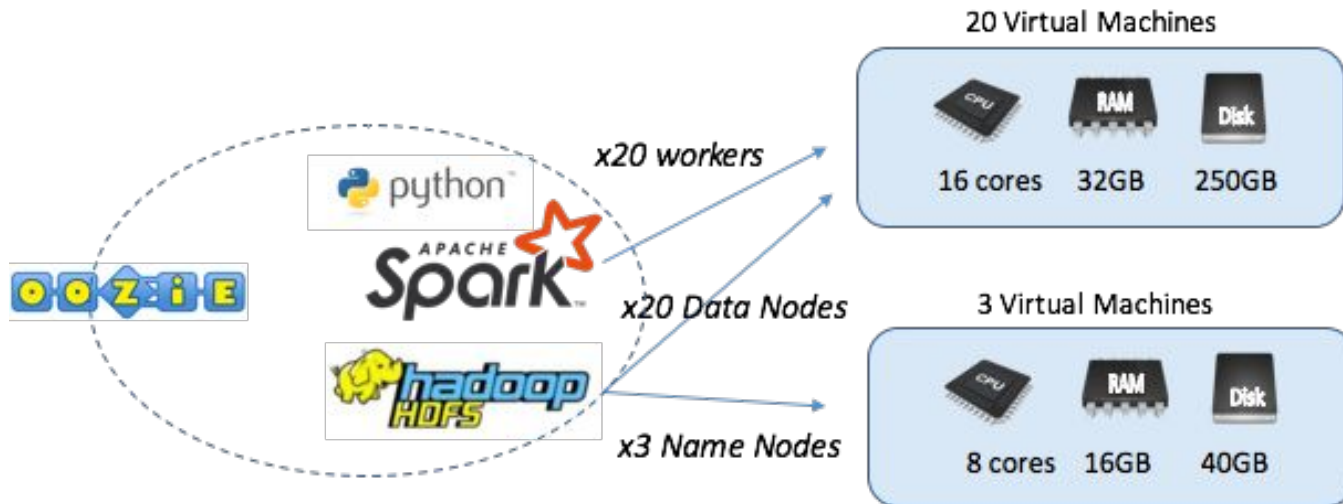
```
1 ~
2   "issued": "2016-07-01", ~
3   "abstract": [ ~
4     { ~
5       "value": "Abstract Text", ~
6       "provenance": "MAG" ~
7     }, ~
8     { ~
9       "value": "Abstract Text", ~
10      "provenance": "Crossref" ~
11    } ~
12  ], ~
13  "subject": [ ~
14    "Agronomy and Crop Science", ~
15    "Forestry" ~
16  ], ~
17  "type": "journal-article", ~
18  "license": [ ~
19    { ~
20      "url": "http://www.elsevier.com/tdm/userlicense/1.0/", ~
21      "date-time": "2011-07-01T00:00:00Z", ~
22      "content-version": "tdm", ~
23      "delay-in-days": 0 ~
24    } ~
25  ], ~
26  "instances": [ ~
27    { ~
28      "url": "http://unkonwonInstance.org", ~
29      "access-rights": "UNKNOWN", ~
30      "provenance": "Crossref" ~
31    }, ~
32    { ~
33      "url": "http://openAccessInstance.org", ~
34      "access-rights": "OPEN", ~
35      "provenance": "Unpaywall" ~
36    } ~
37  ], ~
38  "published-online": "2016-08-01", ~
39  "published-print": "2016-07-01", ~
40  "accepted": "2016-01-01", ~
41  "publisher": "Publisher Name", ~
42  "doi": "10.1016/j.ffhfhghf", ~
43  "doi-url": "http://dx.doi.org/10.1016/j.ffhfhghf", ~
44  "issn": [ ~
45    { ~
46      "type": "print", ~
47      "value": "01234-5678" ~
48    } ~
49  ], ~
50  "collected-from": [ ~
51    "Crossref", ~
52    "MAG", ~
53    "Unpaywall", ~
54    "ORCID" ~
55  ], ~
56  "record-quality-report": "complete" ~
57 }
```

DOIBoost: Methodology

- DOIBoost software toolkit (<https://zenodo.org/record/1441058#.XFQQFc9KgUs>)
- Apache Spark
 - Fast and general purpose cluster computing system
 - 10x (on disk) - 100X (in-Memory) faster
 - Provides high level APIS in
 - Java
 - **Python**
 - Scala
 - R
- Workflow in a nutshell
 - Load input datasets on HDFS
 - Prepare Spark DataFrames
 - Join them and produce DOIBoost



DOIBoost: Deployment





DOIBoost: Execution Time

Execution Step	Execution time
<i>generateCrossrefDataFrame.py</i>	6.1 m
<i>generateMAGDataFrame.py</i>	1.1 h
<i>generateORCIDDataFrame.py</i>	30 s
<i>generateUnpaywallDataFrame.py</i>	20 s
<i>createDOIBoost.py</i>	35 m



Results: Input datasets and contributing properties

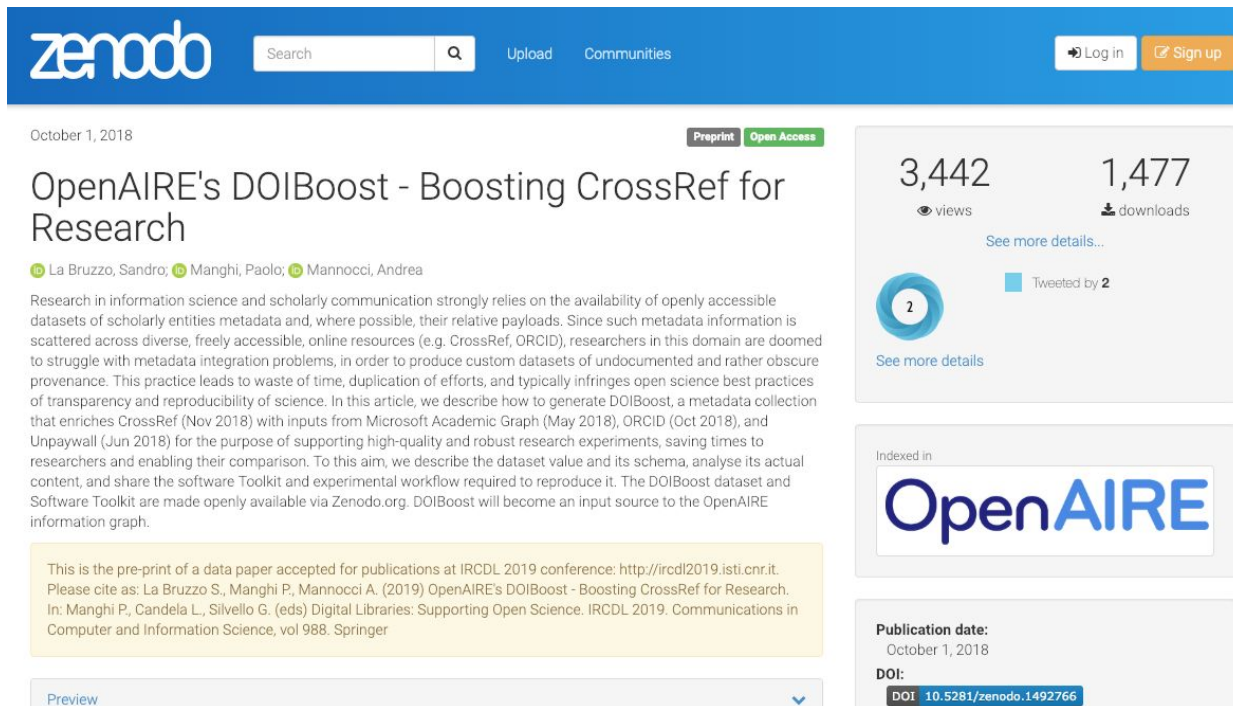
Source	Properties	# of Crossref DOIs enriched by the source with the property	Boost: # of Crossref DOIs enriched by the source with a missing property
ORCID	author IDs (ORCID)	11,345,996	9,666,098
MAG	DOIs	71,654,334	68,561,516
	affiliation (GRID.ac)	45,670,806	45,670,806
	affiliation (Microsoft)	51,630,810	47,528,221
	abstract	45,407,968	43,857,752
	author ID (Microsoft)	74,582,104	68,561,516
	date	71,654,334	2,542,773
Unpaywall	instances	97,751,914	22,328,223
Crossref	all fields	100,507,347	91,365,868



Results: Authorships enrichments since 2003

Indicator	Quantity	Boost
# authorships	263,869,225	–
# authorships in Crossref assigned an identifier	3,060,804 (1.15%)	212,291,232 (80.45%)
# authorships in Crossref assigned an affiliation	25,941,421 (9.83%)	165,271,110 (62.63%)

Find DOIBoost on Zenodo



The screenshot shows the Zenodo interface for the dataset 'OpenAIRE's DOIBoost - Boosting CrossRef for Research'. The page includes a search bar, navigation links for 'Upload' and 'Communities', and user options for 'Log in' and 'Sign up'. The dataset is dated October 1, 2018, and is marked as 'Preprint' and 'Open Access'. It has 3,442 views and 1,477 downloads. The authors listed are La Bruzzo, Sandro; Manghi, Paolo; and Mannocci, Andrea. A yellow box contains a pre-print notice. The dataset is indexed in OpenAIRE. A 'DOI' badge shows the identifier 10.5281/zenodo.1492766. A 'Preview' button is visible at the bottom left.

zenodo Search Upload Communities Log in Sign up

October 1, 2018 Preprint Open Access

OpenAIRE's DOIBoost - Boosting CrossRef for Research

La Bruzzo, Sandro; Manghi, Paolo; Mannocci, Andrea

Research in information science and scholarly communication strongly relies on the availability of openly accessible datasets of scholarly entities metadata and, where possible, their relative payloads. Since such metadata information is scattered across diverse, freely accessible, online resources (e.g. CrossRef, ORCID), researchers in this domain are doomed to struggle with metadata integration problems, in order to produce custom datasets of undocumented and rather obscure provenance. This practice leads to waste of time, duplication of efforts, and typically infringes open science best practices of transparency and reproducibility of science. In this article, we describe how to generate DOIBoost, a metadata collection that enriches CrossRef (Nov 2018) with inputs from Microsoft Academic Graph (May 2018), ORCID (Oct 2018), and Unpaywall (Jun 2018) for the purpose of supporting high-quality and robust research experiments, saving times to researchers and enabling their comparison. To this aim, we describe the dataset value and its schema, analyse its actual content, and share the software Toolkit and experimental workflow required to reproduce it. The DOIBoost dataset and Software Toolkit are made openly available via Zenodo.org. DOIBoost will become an input source to the OpenAIRE information graph.

This is the pre-print of a data paper accepted for publications at IRCDL 2019 conference: <http://ircdl2019.isti.cnr.it>. Please cite as: La Bruzzo S., Manghi P., Mannocci A. (2019) OpenAIRE's DOIBoost - Boosting CrossRef for Research. In: Manghi P., Candela L., Silvello G. (eds) Digital Libraries: Supporting Open Science. IRCDL 2019. Communications in Computer and Information Science, vol 988. Springer

3,442 views 1,477 downloads See more details...

2 Tweeted by 2 See more details

Indexed in OpenAIRE

Publication date: October 1, 2018 DOI: DOI 10.5281/zenodo.1492766

Preview



Paper



DOIBoost



Conclusion and future work

- Presented **DOIBoost**
 - Open scholarly dataset
 - Reproducible methodology
 - Suitable for diverse research applications (abstracts, OA resources/PDFs, authors and affiliations IDs, etc.)

Future work

- Integration with OpenAIRE (Done. Currently in beta)
- Maintain DOIBoost fresh and up-to-date
- Integrate other sources
- **Feedback is welcome!**

Thank you!

andrea.mannocci@open.ac.uk