Merits and Limits

Applying open data to monitor Open Access publications in bibliometric databases

Ali (Aliakbar Akbaritabar)
Stephan Stahlschmidt

DZHW - Berlin

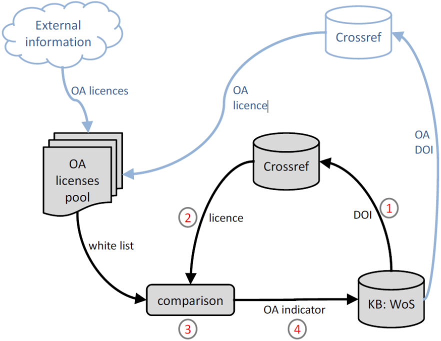
18/06/2019
Bielefeld
Workshop on Matching procedures

Introduction

- KB project to use Crossref and Unpaywall inforamtion and detect OA
  - including: Gold, HiddenGold, Hybrid and Delayed OA
- Updated report: **10.31235/osf.io/sdzft** (https://osf.io/preprints/socarxiv/sdzft/)

Our Workflow

- In-house WOS/Scopus of 2017
- In-house Crossref snapshot (April 2018)
- GESIS tables of Unpaywall on KB (April 2018)
- Schema for Crossref procedure



OA Identification Criteria on Crossref

- ISSN in DOAJ and/or ROAD $\Longrightarrow$ **Gold OA**
- All publications had *OA* licences, ISSN not in DOAJ or ROAD $\Longrightarrow$ **Hidden Gold OA**
- An issue had at least one *non-OA* publication + one or more *OA* pubs $\Longrightarrow$ **Hybrid OA**
- An issue did not have a *non-OA* publication + one or more *OA* pubs + pubs w/ no licence $\Longrightarrow$ **Probable Hybrid OA**
- If `delay-in-days` > 0 in all above cases $\Longrightarrow$ **Delayed OA**
- ISSN not in DOAJ and/or ROAD, number of pubs in issue = *non-OA* $\Longrightarrow$ **Closed Access**
- None of the above $\Longrightarrow$ **Not available (NA)**

1- What was the overall strategy you followed when matching OA-evidence sources with WoS?

In both cases DOI was the most straight forward option to match

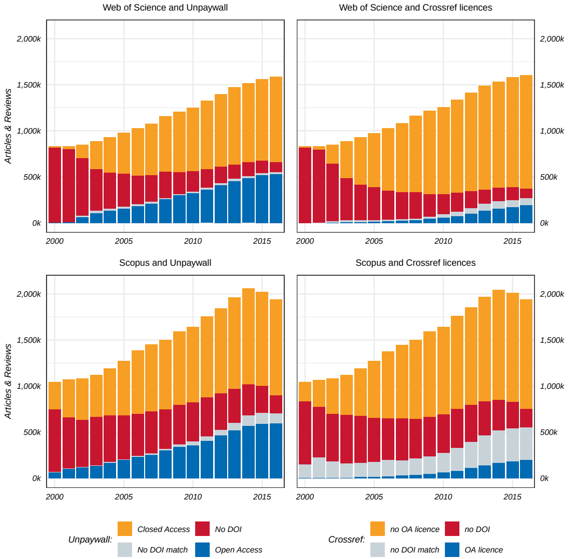
2- What was the major problem you faced during the matching?

Standardization of metadata?!

- *DOI* is *not* as standard as it is hoped
- Duplicate DOIs
- Non coherent *upper*/*lower* case usages
- An example:
  - *Exact matching AR/RE of WOS w/ Unpaywall* ⟹ *11,346,682* (57.03%)
  - *Lower case* DOI ⟹ *13,886,618* (69.8%)
  - Duplicated DOIs removed, increased matches even further
- WoS (Scopus), between 2000 and 2016:
  - 63 (913) journals **without an ISSN**
  - 23,000+ (140,000+) articles and reviews **without a unique DOI** (excluded from our results)
  - 5,700,000+ (7,400,000+) pubs **without any DOI**
  - 2,700,000+ (4,500,000+) pubs with a DOI which can't be resolved on Crossref

3- What kind of results could be achieved by matching the OA-evidence source with WoS?

OA from Crossref/Unpaywall matched WOS/Scopus



OA in Unpaywall (journal level)

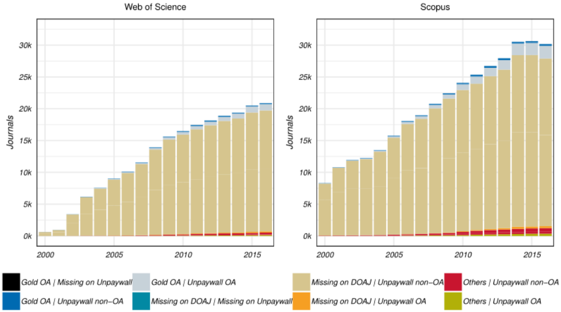
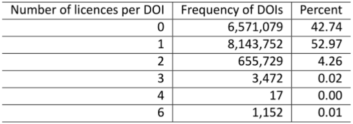


Figure 1: Journals indexed in WOS and Scopus matched with Unpaywall database and crosschecked the ISSNs with DOAJ (Gold OA) between 2000 and 2016

OA in Crossref (publication level)

Table 2: Number of licences per DOI found in Crossref for articles and reviews indexed in either WOS, Scopus or both between 2000 and 2016



Number of licences per DOI	Frequency of DOIs	Percent
0	6,571,079	42.74
1	8,143,752	52.97
2	655,729	4.26
3	3,472	0.02
4	17	0.00
6	1,152	0.01

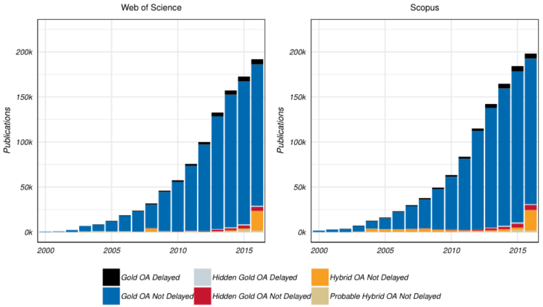
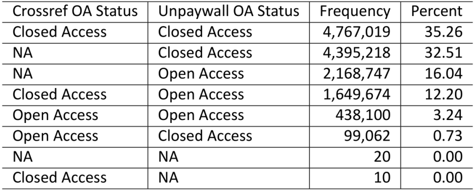


Figure 2: Count of gold and hybrid OA publications between 2000 and 2016 based on Crossref licence information, DOAJ and ROAD)

4- How did you attempt to evaluate the quality of the matching procedure? If no attempt was undertaken: How could an evaluation of the results look like?

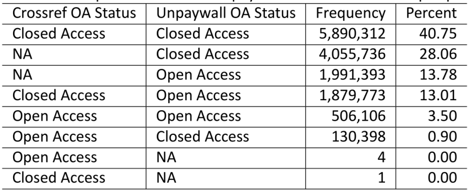
Check DOIs on both Crossref and Unpaywall

Table 3: OA status comparison between Unpaywall and Crossref on WOS publications



Crossref OA Status	Unpaywall OA Status	Frequency	Percent
Closed Access	Closed Access	4,767,019	35.26
NA	Closed Access	4,395,218	32.51
NA	Open Access	2,168,747	16.04
Closed Access	Open Access	1,649,674	12.20
Open Access	Open Access	438,100	3.24
Open Access	Closed Access	99,062	0.73
NA	NA	20	0.00
Closed Access	NA	10	0.00

Table 4: OA status comparison between Unpaywall and Crossref on Scopus publications



Crossref OA Status	Unpaywall OA Status	Frequency	Percent
Closed Access	Closed Access	5,890,312	40.75
NA	Closed Access	4,055,736	28.06
NA	Open Access	1,991,393	13.78
Closed Access	Open Access	1,879,773	13.01
Open Access	Open Access	506,106	3.50
Open Access	Closed Access	130,398	0.90
Open Access	NA	4	0.00
Closed Access	NA	1	0.00

Manual Check of Random Samples

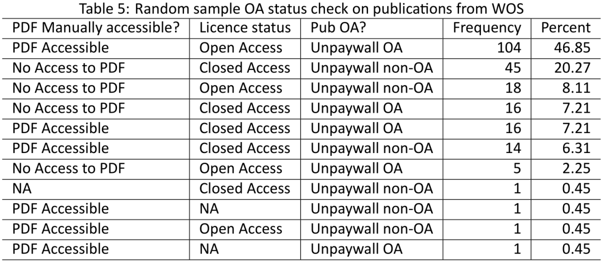


Table 5: Random sample OA status check on publications from WOS

PDF Manually accessible?	Licence status	Pub OA?	Frequency	Percent
PDF Accessible	Open Access	Unpaywall OA	104	46.85
No Access to PDF	Closed Access	Unpaywall non-OA	45	20.27
No Access to PDF	Open Access	Unpaywall non-OA	18	8.11
No Access to PDF	Closed Access	Unpaywall OA	16	7.21
PDF Accessible	Closed Access	Unpaywall OA	16	7.21
PDF Accessible	Closed Access	Unpaywall non-OA	14	6.31
No Access to PDF	Open Access	Unpaywall OA	5	2.25
NA	Closed Access	Unpaywall non-OA	1	0.45
PDF Accessible	NA	Unpaywall non-OA	1	0.45
PDF Accessible	Open Access	Unpaywall non-OA	1	0.45
PDF Accessible	NA	Unpaywall OA	1	0.45

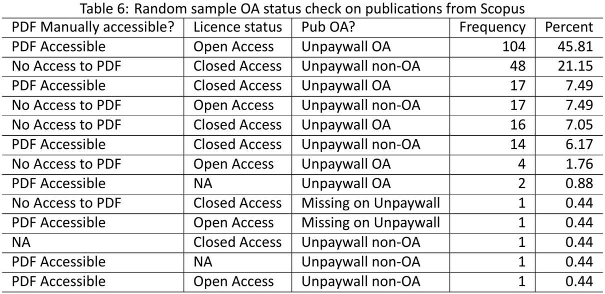


Table 6: Random sample OA status check on publications from Scopus

PDF Manually accessible?	Licence status	Pub OA?	Frequency	Percent
PDF Accessible	Open Access	Unpaywall OA	104	45.81
No Access to PDF	Closed Access	Unpaywall non-OA	48	21.15
PDF Accessible	Closed Access	Unpaywall OA	17	7.49
No Access to PDF	Open Access	Unpaywall non-OA	17	7.49
No Access to PDF	Closed Access	Unpaywall OA	16	7.05
PDF Accessible	Closed Access	Unpaywall non-OA	14	6.17
No Access to PDF	Open Access	Unpaywall OA	4	1.76
PDF Accessible	NA	Unpaywall OA	2	0.88
No Access to PDF	Closed Access	Missing on Unpaywall	1	0.44
PDF Accessible	Open Access	Missing on Unpaywall	1	0.44
NA	Closed Access	Unpaywall non-OA	1	0.44
PDF Accessible	NA	Unpaywall non-OA	1	0.44
PDF Accessible	Open Access	Unpaywall non-OA	1	0.44

Conclusions

- *WOS/Scopus*: Incoherent metadata in each database (duplicate DOIs, lower/upper case, no DOI)
- *Unpaywall/Crossref*: Incoherent metadata over databases (same DOI, different OA status)
- Good opportunity of mixing/matching/merging different databases
- Unpaywall has a higher coverage that goes beyong Crossref

Thanks for your attention!