SONAR Deliverable 5.1

Evaluation of acquisition accuracy

Gobeill Julien[1,2], Liyanapathirana Jeevanthi[2], Santos Anouk[2], and Ruch Patrick[1,2]

[1]SIB Text Mining group, Swiss Institute of Bioinformatics, Geneva, Switzerland
[2]HES-SO / HEG, Information Sciences, Geneva, Switzerland

1. Executive summary

Purpose. This study aims at assessing the feasibility of a pipeline for automatically retrieving publications of researchers affiliated to Swiss public institutions, and at evaluating its accuracy and coverage. The scope of this study is limited to publications in which a Swiss institution is mentioned as authors' affiliation in the publication. In all resources, we only considered entries provided with a Digital Object Identifier (DOI).

Methods. Not all bibliographic databases propose an affiliation metadata for publications. We exploited four significant international archiving resources, namely MEDLINE, PubMedCentral, CrossRef and Unpaywall. We tested and developed specific methods for harvesting targeted bibliographic records and Open Access full-texts (PDF files) from these resources. The affiliation metadata of publications were compared to a manually designed authority list of 75 Swiss institutions, with 523 naming standards and synonyms. Additionally, for retrieving publications without completed affiliation metadata in investigated databases, we explored an additional brute force pipeline: it consists in downloading all available Open Access full-texts, then applying a Text Mining solution (Grobid) in order to extract the affiliation from the PDF files.

Results. Three benchmarks have been used to evaluate the coverage of the pipeline. The first two are based on publications having the Swiss National Science Foundation (SNF) in funding metadata in CrossRef (24,607), and on publications present in the SNF P3 database (63,747 with a DOI). For recent publications (since 2015), we are able to recover a bibliographic record for respectively 60% and 51% of the targeted publications, and a full-text for respectively 26% and 22%. The analysis of the not-retrieved publications shows that, while some are really unreachable (e.g. no available affiliation metadata, nowhere), many are not in the perimeter of this study, because the institution is not considered (e.g. CERN) or the SNF is mentioned without Swiss affiliations (e.g. mobility grants, or proceedings of conferences funded by the SNF). The last benchmark is based on Swiss institutional repositories (Universities of Zurich, Bern, and Geneva). For recent publications (since 2015), we are able to recover a bibliographic record for between 60% and 63% of the targeted publications, and a full-text for between 32% and 36%. A promising result is that, if they incorporated our retrieved dataset into their collection, these institutional repositories potentially could see their coverage gain between +30% and +54%, including numerous Open Archives full-texts (for ~50% of the additional records). Finally, the brute force pipeline with Grobid could bring an additional portion of 10% to 15% of records; but the

estimated downloading and processing time, estimated to 20 hours per day for a single core, is a serious hurdle.

Conclusion. Thanks to our pipeline, 190,000 bibliographic records with at least one of our Swiss affiliations were retrieved, including 53,100 (28%) Open Access full-texts. We thus retrieve records for between 60% and 63% of the Swiss publications contained in our benchmarks, and between 26% and 36% of full-texts. This retrieval rate is quite stable since 2015, regarding the publication year. If Swiss institutions enriched their repositories with our retrieved dataset, they could potentially gain between +30 and +54% of bibliographic records, including numerous full-texts. An additional proportion could be provided by the additional brute force pipeline, but the computing price is high.

Table of contents

1. Executive summary 1
2. Purpose and scope 3
3. Overview 4
   - 3.1. The curse of the affiliation metadata in databases 4
   - 3.2. Exploited databases 4
   - 3.3. The landscape of Swiss publications 5
   - 3.4. Investigated pipeline for Swiss publications retrieval 6
4. Methods 7
   - 4.1. The authority list of Swiss affiliations 7
   - 4.2. Querying MEDLINE & PubMed Central 8
   - 4.3. Querying CrossRef & Unpaywall 9
   - 4.4. Designing benchmarks for evaluation 10
5. Evaluation 12
   - 5.1. Statistics on retrieved sets 12
   - 5.2. Benchmark #1: SNF_Xref 12
   - 5.3. Benchmark #2: SNF_P3 14
   - 5.4. Benchmark #3: Swiss institutional repositories 17
6. Error analysis: the unreachable publications 17
7. Additional brute force pipeline 18
   - 7.1. Grobid 18
   - 7.2. Extrapolated gains and efforts 20
8. References 21

2. Purpose and scope

In order to “retrieve undisclosed Swiss publications from external sources” (according to the proposal), the Work Package 5 of the SONAR project – Recovering of full-text from 3rd-party OA (Feasibility study) – focuses on harvesting international open archives.

Swiss research institutions. In this study, an authority list of Swiss research institutions considered by the Swiss National Science Foundation was extracted from http://p3.snf.ch/. This authority list gathers 75 different institutions, including universities (such as UNIGE), ETH schools (such as EPFL), schools of higher education (such as HES-SO), and other research institutes (such as SIB). This authority list contains 523 naming standards and synonyms manually designed. See section 4.1 for more information.

Swiss publications. Consequently, in this study, targeted “Swiss publications” refers to research publications with at least one author’s affiliation, in the publication metadata, belonging to this authority list. This definition especially excludes: (1) publications from institutions outside the scope of the study, e.g. the CERN ; (2) publications for which the author has fulfilled his affiliation in an uncomplete or confusing way, thus not contained in our authority list, e.g. mentioning “faculty” instead of “university” ; (3) publications claimed today by a researcher in a Swiss institution, but written while he belonged to a foreign institution, e.g. French or American universities. We did not investigate any approach based on authors names, but only focused on affiliations.

Investigated resources. In our study, we investigated open and sustainable approaches. Several reference resources were considered: CrossRef, Unpaywall, MEDLINE, PubMed Central, arXiv, and SemanticScholar. We only interacted with these resources via dedicated Application Programming Interfaces (defined and authorized Machine-to-Machine communication). This especially excludes Web scraping, which is downloading and extracting information from web pages initially intended for humans (Machine-to-Human communication). Web scraping solutions are generally highly perishable, as publishers can decide at any time to change their webpages, or to ban the IP of the scrapers.

DOI as unique identifier & publication year. In both investigated resources, benchmarks for evaluation, and outputted datasets, publications were identified thanks to Digital Object Identifiers (DOIs). Consequently, we limited our study to publications with a DOI. Furthermore, in order to have consistent results across all successive experiments, we limited our study to publications published in 2018 and before.

3. Overview

3.1. The curse of the affiliation metadata in databases

By definition, the scope of this study is to retrieve publications for which the authors' affiliations metadata contains a Swiss institution. Unfortunately, the affiliation metadata is just absent in many of the databases we investigated :

- arXiv (https://arxiv.org/), owned and operated by Cornell University, provides in Nov. 2019 Open access to 1,600,000 e-prints in the fields of physics, mathematics, or computer science... Unfortunately, the bibliographic records in arXiv have an authors metadata, but no affiliations metadata.
- Semantic Scholar (https://www.semanticscholar.org/), a free, nonprofit, academic search engine from Allen Institute for AI, claims in Nov. 2019 180,000,000 bibliographic records for publications from all fields of science. Indeed, semantic scholar is promising in terms of coverage (most of our publications were present, competitive with commercial databases), but, like arXiv, the affiliations metadata is simply missing.
- some commercial databases were explored, such as Google Scholar (https://scholar.google.fr/, free), and Scopus (https://www.scopus.com, accessed via EPFL subscription). Once again, the affiliations metadata is missing, and it is not possible to retrieve publications based on an affiliation (such as University of Geneva).

We finally exploited four databases which, at least, contain an affiliation metadata in their bibliographic records.

3.2. Exploited databases

Here are the four databases we exploited:

- MEDLINE (https://www.ncbi.nlm.nih.gov/pubmed/), compiled by the United States National Library of Medicine, comprises in Nov. 2019 more than 30,000,000 bibliographic records of journal articles in life sciences with a concentration on biomedicine. Citations may include links to full-text content from PubMed Central and publisher web sites.
- PubMed Central (PMC, https://www.ncbi.nlm.nih.gov/pmc/), developed by the United States National Center for Biotechnology Information (NCBI), is a free digital repository that archives in Nov. 2019 2,600,000 publicly accessible full-texts. Almost 100% of Open Access publications found in PMC have a corresponding record in MEDLINE.
- CrossRef (https://www.crossref.org/), developed by the official DOI Registration Agency, claims in Nov. 2019 more than 80,000,000 bibliographic records of journal articles from all scientific domains. The non-profit goal of CrossRef is to structure, process, and share metadata to reveal relationships between distributed contents hosted at other sites (publishers, open databases…).
- Unpaywall (https://unpaywall.org/) is a free service which locates open-access articles and presents paywalled papers that have been legally archived and are freely available on other websites. It claims 24,800,000 links to free scholarly articles (in Nov. 2019). This service can be accessed via an APi or by downloading an entire database snapshot.

Here are some characteristics of exploited databases revealed by some preliminary analyses:

Resource	MEDLINE	PMC	CrossRef	Unpaywall
Content	Bibliographic records	Full-texts	Bibliographic records	Links to full-texts
Quantity	30M	3.6M	80M	20M
Scientific domains	Biomedicine	Biomedicine	All	All
Affiliation metadata completion	100% of records	100%	~33%	NA
Programming Interface / database download	yes / yes	yes / yes	yes / no	yes / yes

Tab. 1 : statistics on investigated databases

3.3. The landscape of Swiss publications

The next figure presents a landscape of Swiss publications, in the perspective of our selected databases.

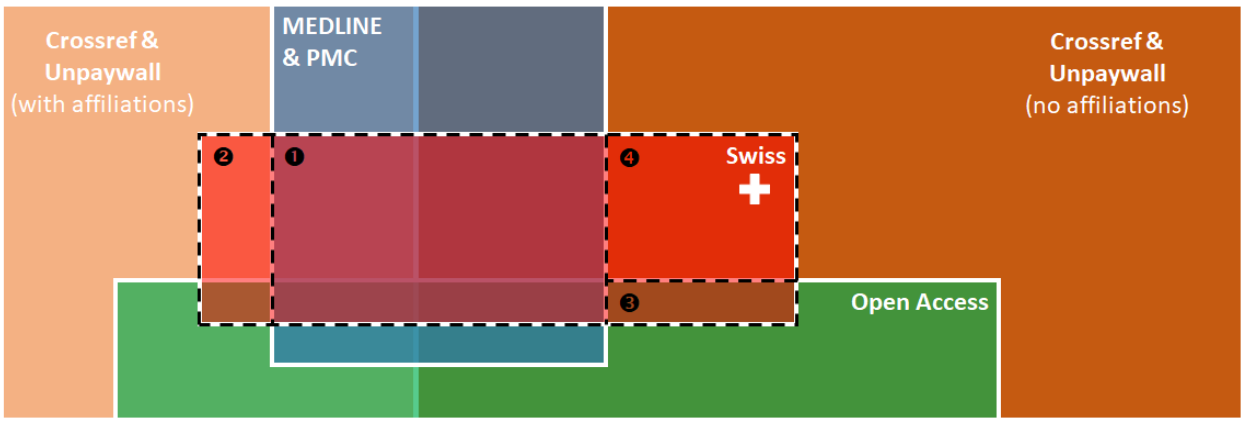

Fig. 1: landscape of Swiss publications

In orange stand publications covered by CrossRef and Unpaywall (both resources are grouped in this figure), divided into two sub-parts: for around 37% of records (light orange, on the left, see section 4.3 for details about this 37% estimation), the affiliation metadata is completed, for the rest (dark orange, on the right) it is not. In green stands the Open Access subset, which is a priori distributed equally among the light and dark orange subsets. In this figure, we hypothesize that this Open Access subset is approximately 25~30% of the landscape (Piwowar et al., 2018). On top of this, in blue, stands MEDLINE (biomedical publications). According to previous studies (Putallaz et al., 2018), we hypothesize that biomedical publications represent 50% of Swiss publications. The intersection of blue and green subsets is PubMed Central (biomedical Open Access publications). Finally, on top of this are the Swiss publications (in red).

In this figure, Swiss publications are divided into four numbered subsets:

(1) blue and red subset: Swiss biomedical publications (i.e. indexed in MEDLINE). The approach for retrieving this subset is the so-called blue way in the next pipeline schema.
(2) light orange and red: Swiss non-biomedical publications (i.e. not indexed in MEDLINE), but with the complete affiliation metadata in CrossRef. The approach for retrieving this subset is the so-called orange way in the next pipeline schema.
(3) dark orange, green and red: Swiss non-biomedical publications (i.e. not indexed in MEDLINE), without affiliation in CrossRef, but Open Access (i.e. retrievable PDF file). The approach for retrieving this subset is the so-called green way in the next pipeline schema.
(4) dark orange and red: Swiss non-biomedical publications (i.e. not indexed in MEDLINE), without complete affiliation in CrossRef, and not Open Access. This subset is theoretically unreachable by our approach.

3.4. Investigated pipeline for Swiss publications retrieval

The next figure presents the global pipeline of our system. The input is the name of a Swiss institution; the output are sets of bibliographic records and full-text PDFs for which this institution is mentioned as affiliation in metadata. Thanks to the authority list, the query is expanded to a set of potential synonyms. Then, three colors stand for three different approaches :

- the “blue way”: MEDLINE is queried with the expanded query for metadata affiliation. Corresponding records populate the bibliographic records output. Corresponding records present in PMC are downloaded then populate the full-texts output.
- the “orange way”: CrossRef is queried with the expanded query for metadata affiliation. Corresponding records populate the bibliographic records output. Corresponding records with an OA link in Unpaywall are downloaded then populate the full-texts output.
- the “green way” (exploratory): an additional brute force pipeline. All OA PDF links in Unpaywall are downloaded (possibly millions per year). A Text Mining system is applied to PDF files in order to extract the affiliations. Full-texts with corresponding affiliations populate the full-texts output, and corresponding records in CrossRef populates the bibliographic records output. This approach is only explored, with a random samples, in our study (see part 6).

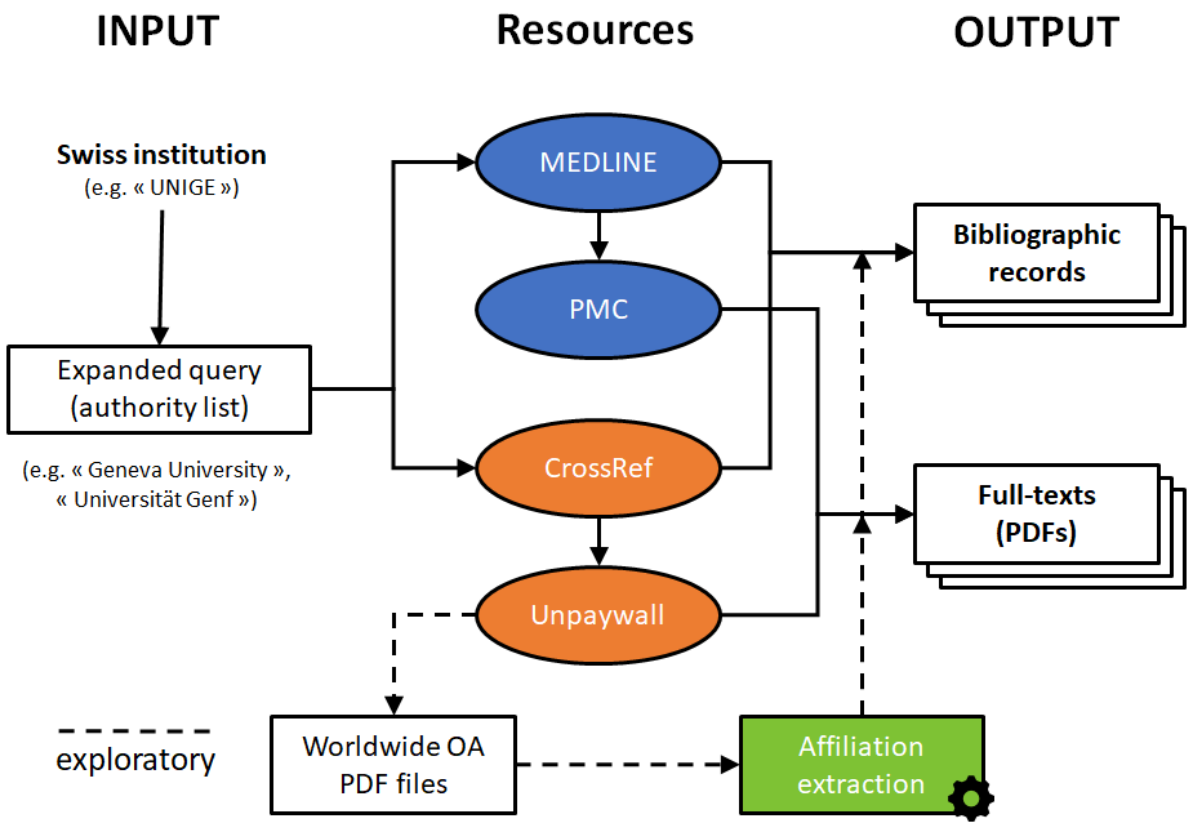

Fig. 2 : investigated pipeline for Swiss publications retrieval. Dotted lines stand for the additional brute force pipeline, which is exploratory in this study.

4. Methods

4.1. The authority list of Swiss affiliations

Manual design of the list. In order to have a complete list of Swiss research institutions we considered all those mentioned by the Swiss National Science Foundation in http://p3.snf.ch/ (universities, EPF schools, schools of higher education, universities of teacher education and other research institutes). To design the list, we searched in Crossref and Medline via API in the affiliation field. The purpose was to try to figure out the right combination of words for an affiliation, that allows to find multiple variants of that affiliation. For exemple, when typing "Lausanne University", we found :

- Lausanne University Hospital
- Lausanne University Hospital and University of Lausanne (CHUV-UNIL)
- Lausanne University hospital - CHUV
- Lausanne University Hospital Chuv
- Lausanne University Hospital (CHUV)
- Lausanne University Hospital CHUV
- Lausanne University Hospital and University of Lausanne
- Lausanne University Hospital Medical Center

So, we kept the formulation “Lausanne University” in the affiliation list, because that formulation allows us to find a lot more affiliations used by the searchers of that university. We also searched for affiliations used by all those Swiss research institutions in the work of the consortium, which is making lists of secondary publications (green OA) per institution : https://consortium.ch/open-access/?lang=en. Finally, the list was completed by affiliations mentioned in http://viaf.org/.

Explicit “Switzerland” affiliations. Preliminary experiments showed that our authority list retrieved numerous false positives (i.e. non Swiss affiliations). For instance, the “Institute for Word and Health” affiliation, used by researchers from the Swiss IST (Institut universitaire romand de Santé au Travail), is also used by researchers from Toronto. Or the SIB acronym (for Swiss Institute of Bioinformatics) also refers in databases to the Science Innovation Business labs from Finland. We thus decided to discard all affiliations which did not contain “Schweiz” OR “Suisse” OR “Switzerland” OR “Svizzera”. This strong assumption probably costs some false negatives (i.e. Swiss affiliations without country name, which we thus considered non-standard), but is the only way for discarding false positives.

4.2. Querying MEDLINE & PubMed Central

Database access. MEDLINE and PubMed Central are accessible to programs via APIs: the e-utils (https://www.ncbi.nlm.nih.gov/home/tools/). In particular, advanced queries allow to search for a string (e.g. “EPFL”) only in the affiliation metadata. Yet, MEDLINE and PMC also are accessible via bulk download: baseline and updates files covering the entire collections are provided by FTP (https://ftp.ncbi.nlm.nih.gov/). Our group maintains local and daily-updated mirrors of both databases, for providing the SIBiLS services (http://candy.hesge.ch/SIBiLS/). Technically, we load data in MongoDB collections; this Data Management System allows efficient queries in metadata.

Completion of the DOI metadata in MEDLINE. For identifying records in MEDLINE, the NLM uses a PMID (PubMed IDentifier, a unique identifier). DOI is a non-mandatory metadata of a record. The following figure shows percentages of MEDLINE records with a provided DOI across the publication year, when the affiliation contains “Switzerland”. We see that, for publications between 1983 and 2002, there is a serious issue: less than 70% (only 50% for 1986) of records have a DOI. Our retrieval rate could suffer from this rate. Yet, for recent publications, this issue seems to be solved, at least 99% completion after 2014.

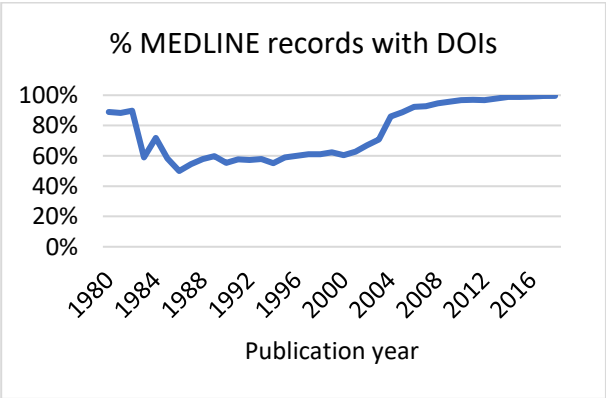

Fig. 3: completion of DOI metadata in MEDLINE across the time

Completion of the affiliation metadata in MEDLINE. Preliminary analyses showed that 100% of MEDLINE records had the affiliation metadata completed. Yet, we discovered examples of MEDLINE records where only the first affiliation was present. For example, this article (https://www.ncbi.nlm.nih.gov/pubmed/22308148) has only 1 given affiliation (Roma) for 5 authors, when the full-text actually gives 4 affiliations, including 2 Swiss (from Basel). This lack of information (missing affiliations) can impact our retrieval rate. We cross-analyzed the percentages of these missing affiliations in MEDLINE for SIB and Uni Zurich retrieved publications, as shown in the following figure. Once again, this technical issue culminates for past publications, but is completely solved for recent ones (less than 5% from 2015). Interestingly, these missing affiliations are most present for SIB publications than for Uni Zurich (max 78% versus max 21%): this can be explained by the higher percentage of multi affiliated researches in SIB, thus this affiliation is more likely to be lost by MEDLINE.

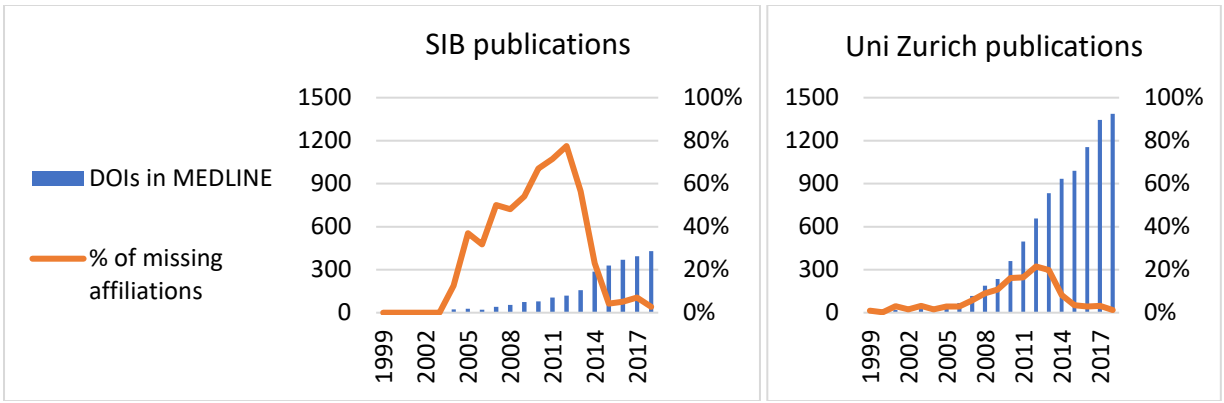

Fig. 4: evolution of the numbers of DOIs in MEDLINE (blue bars) and percentages of missing affiliations (orange curve) for two targeted institutions: SIB and Uni Zurich.

4.3. Querying CrossRef & Unpaywall

Database access. We interacted with CrossRef via the API, and different types of queries.

- URL examples for searching for the information by DOI:
  - https://api.crossref.org/v1/works/10.1093/database/bay129
  - https://api.crossref.org/v1/works/10.1111/evo.1304
- URL examples for searching based on the funding information :
  - http://api.crossref.org/funders/10.13039/501100001711 for SNF (providing synonyms)
- URL examples for searching entries between two dates:
  - filter=from-pub-date:1990-01,until pub-date:1995-01

The querying limit for one query is only 1000 entries, so the data had to be collected using offsets when the number of entries for a certain time period exceeds 1000 entries. The online querying via an API currently supports 10,000 queries for free, which serves the necessity of this research at the moment.

Then, there is a necessity of obtaining the full text pertaining to each DOI. For this purpose, Unpaywall database was used, which will be described below.

In each entry in Unpaywall, there are several types of metadata that were used in our analysis:

- <is_oa>: if the entry belonging to the DOI is open, is_oa will be true, else false.
- <best_oa_location>: if is_oa is true for an entry, there might be a value for the best location you can find the open access pdf for that doi.

- <url_for_pdf>: if is_oa is true for an entry, there might be values for multiple locations for retrieving the open access pdf file.

Completion of the DOI metadata in CrossRef. All entries found in CrossRef do contain a DOI.

Completion of the affiliation metadata in CrossRef. One subset we extracted from CrossRef contained 25,170 entries with SNF as funder (see next section, benchmark #1 SNF_Xref). Out of those entries, 9,313 did have affiliation metadata completed, and the remaining 15,857 did not. Thus, we could estimate that only approximately 37% of the entries of CrossRef contain affiliation details. The figure below depicts the rate of affiliation completion within Crossref, for years 2001 to 2017, in which we can see a positive trend in affiliation existence in the recent past.

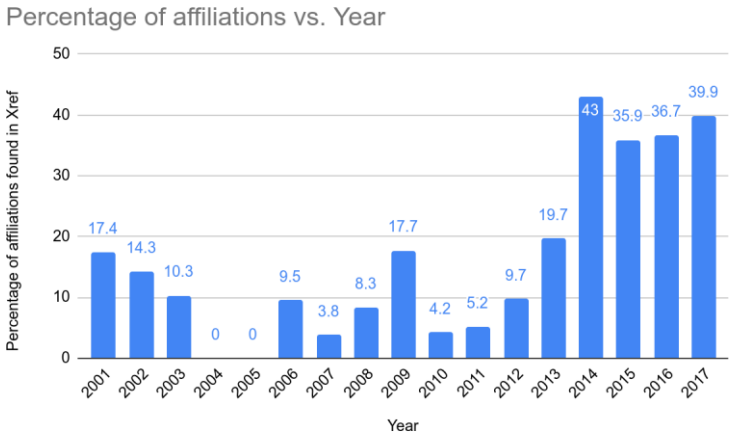

Fig. 5: Percentage of affiliation existence per year, in CrossRef entries, for benchmark #1.

4.4. Designing benchmarks for evaluation

In order to evaluate the accuracy of our retrieval, we designed three benchmarks. The first two are based on Swiss National Science Foundation (SNF) funding, the third on Swiss institutional repositories.

Benchmark #1: SNF_Xref. CrossRef provides a funding metadata in its records, with standardized names of funding organisations. Funding information are uploaded by publishers, possibly with the help of CrossRef upload interface. We thus harvested all the records having the SNF as funder, with a publication date posterior to 2000. Note that not all publications acknowledging a SNF grant have a completed funding metadata in CrossRef. This represents a set of 24,607 records.

Benchmark #2: SNF_P3. Parallel to benchmark #1, we also downloaded publication records directly on the SNF website. These records are generally uploaded by researchers. All publications can be accessed at http://p3.snf.ch/Pages/DataAndDocumentation.aspx. Yet, out of 125,000 records, only 48% have a completed DOI metadata. We thus relied on the work done by Christian Gutknecht at the Co-ordination of information systems in research support (CoSi) in SNF. With his team, he curates the original CSV in order to automatically add DOIs when it is possible. For evaluation, we thus exploited their curated file of 76,293 records with DOI (61%). We discarded publication years prior to 2001 and obtain a final set of 63,747 records. In this benchmark, it was important to preprocess the DOIs. Indeed, 5% of DOIs in P3 are in uppercase, thus could not

match with DOIs from reference databases. An additional 1% of DOIs in P3 have a unnecessary final dot, which was removed.

On all records contained in the first two benchmarks, 68% are only contained in SNF_P3, 16% only in SNF_Xref, and 16% (overlap) in both benchmarks.

Benchmark #3: Swiss institutional repositories. Beyond the SNF, the third benchmark is based on publication records contained in Swiss institutional repositories: Repositorium PHZH for the University and Hospitals of Zurich, ARODES for HESSO, BORIS for the University and Hospitals of Bern, and Archive ouverte UNIGE for the University and Hospitals of Geneva. Records were harvested via the Open Archives Initiative Protocol for Metadata Harvesting. We limited the records selection to the document type “scientific article”, with a publication date posterior to 2015. Note that, in ARODES, publication types are not available (the only value is “text”), thus we selected all records. Moreover, only 90% of the harvested records have a DOI (27% for ARODES). This finally represents a set of 56,549 records.

Our retrieval pipeline was evaluated on its capacity to retrieve records contained in these benchmarks.

5. Evaluation

5.1. Statistics on retrieved sets

Thanks to our pipeline, 173,000 bibliographic records with at least one of our Swiss affiliations were retrieved via MEDLINE, and 65,000 via CrossRef. The overlap between both retrieved sets is only 11%. The percentage of Open Access publications in this dataset is 28%. The following figure shows numbers of retrieved records and percentages of Open Access for the top 15 publishing Swiss affiliations for publication years posterior to 2015.

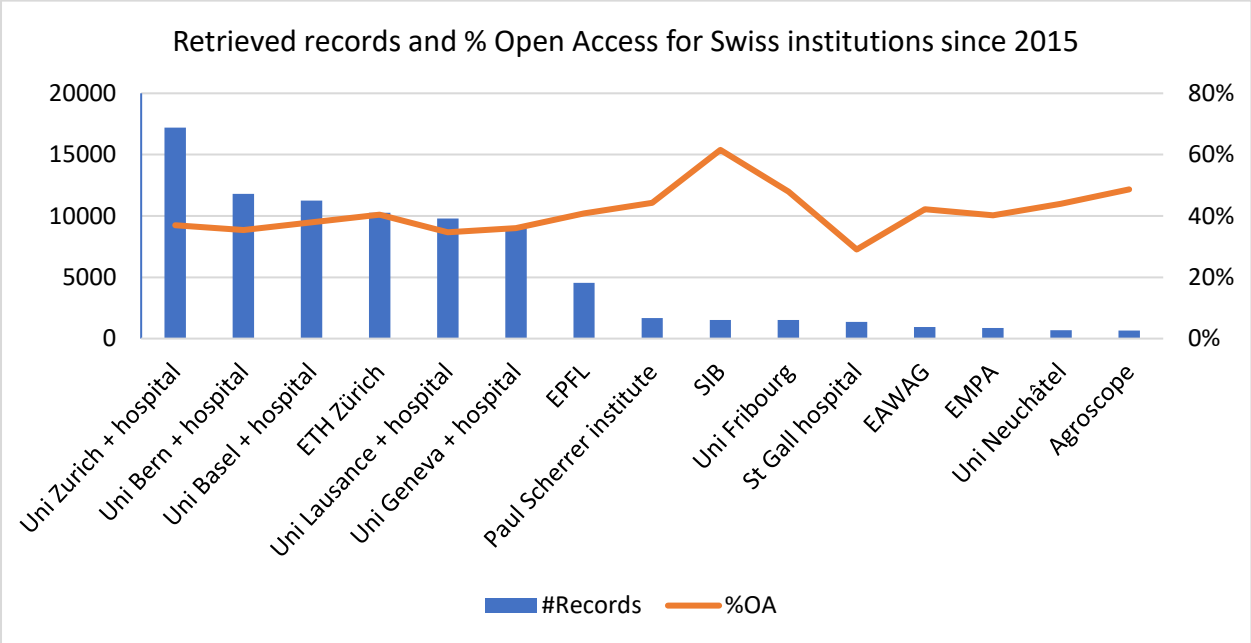

Fig. 6: numbers of bibliographic records, and percentages of Open Access, for top 15 publishing Swiss institutions between 2015 and 2018.

Zurich university is the most publishing institution in our retrieved dataset with 17,200 records between 2015 and 2018, followed by Bern (11,800) and Basel (11,300). In another perspective, the Swiss Institute of Bioinformatics is the leader for Open Access (62%).

5.2. Benchmark #1: SNF_Xref

The following figure shows the evolution of records contained in the benchmark SNF_Xref, since 2010, retrieved by MEDLINE (blue way), by CrossRef (orange way), by both, or by none.

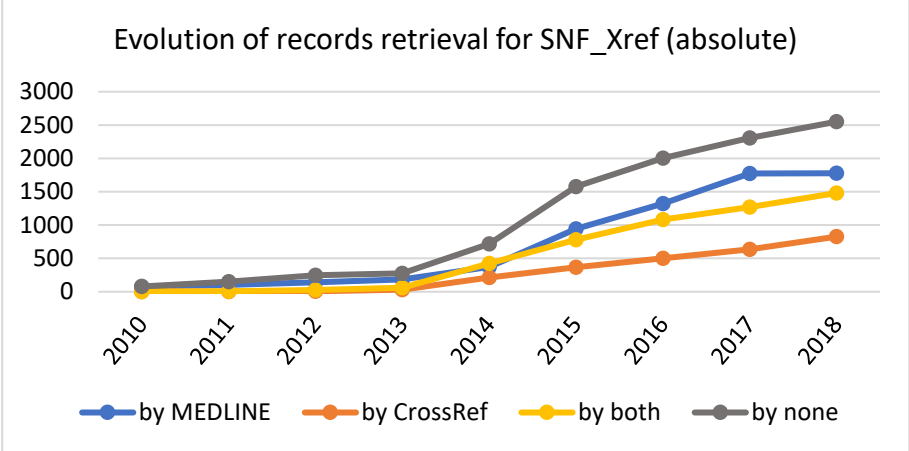

Fig. 7: evolution of records contained in the benchmark #1, whether they were retrieved by blue way, orange way, both, or none.

The number of publications is very low for years previous to 2015. As the funding metadata is provided by publishers, we hypothesize that this practice was not usual before this date. Consequently, 86% of the benchmark's publications are recent (published after 2015).

It is not easy to appreciate the supply for each way with the previous absolute numbers. The following figure shows the same statistics in a relative perspective.

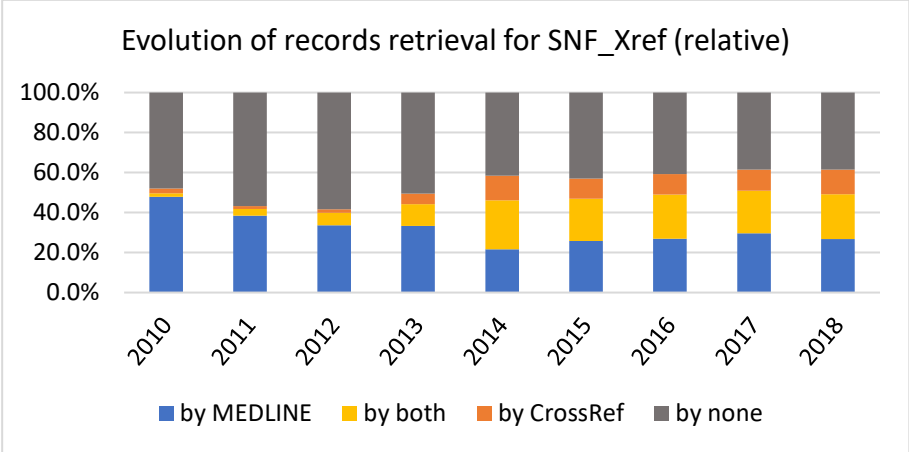

Fig. 8: evolution of the proportions of retrieved records from the benchmark #1, regarding each way.

The records retrieval is quite stable since 2014 (around 60%). If the blue way supply (MEDLINE) is quite stable, the supply of the orange way (CrossRef) steadily increases.

The following figure also shows statistics in a relative perspective, but for the full-texts retrieved.

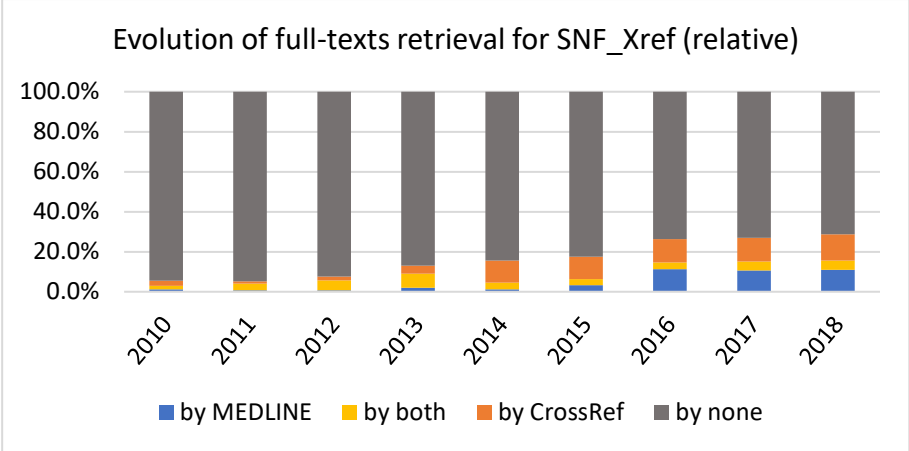

Fig. 9: evolution of the proportions of retrieved full-texts from the benchmark #1, regarding each way.

For full-texts, the orange way (CrossRef + Unpaywall) seems to be more useful than the blue way (PubMedCentral), especially before 2015.

At last, the following figure shows the percentage of records (left) and full-texts (right) retrieved by both ways, for publications since 2015.

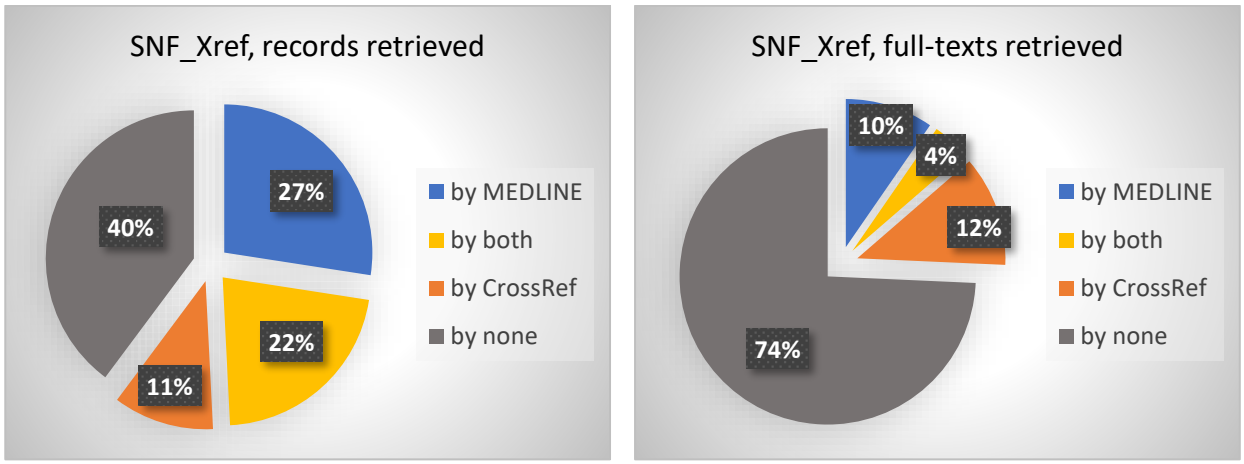

Fig. 10: proportions of recent (published since 2015) records and full-texts, retrieved for the benchmark #1.

For recent publications (since 2015), 60% of records contained in the benchmark SNF_Xref are retrieved by MEDLINE, CrossRef, or both. 40% of records in the benchmark are not retrieved. Dealing with Open Access, 26% of publications are detected as OA, with an accessible PDF file.

5.3. Benchmark #2: SNF_P3

Similar graphics were computed with benchmark #2.
The following figure shows the evolution of records contained in the benchmark SNF_P3, since 2010, retrieved by MEDLINE (blue way), by CrossRef (orange way), by both, or by none.

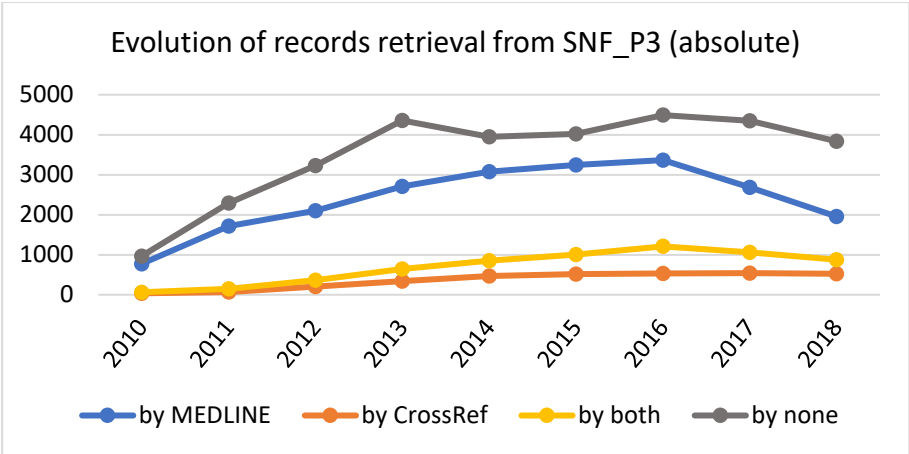

Fig. 11: evolution of records contained in the benchmark #2, whether they were retrieved by blue way, orange way, both, or none.

We observe stable numbers since 2013, and also decreasing numbers for very recent years (2017 – 2018). As P3 contains publications uploaded by researchers, we hypothesize that this is due to a latency period.

The following figure shows the same statistics in a relative perspective.

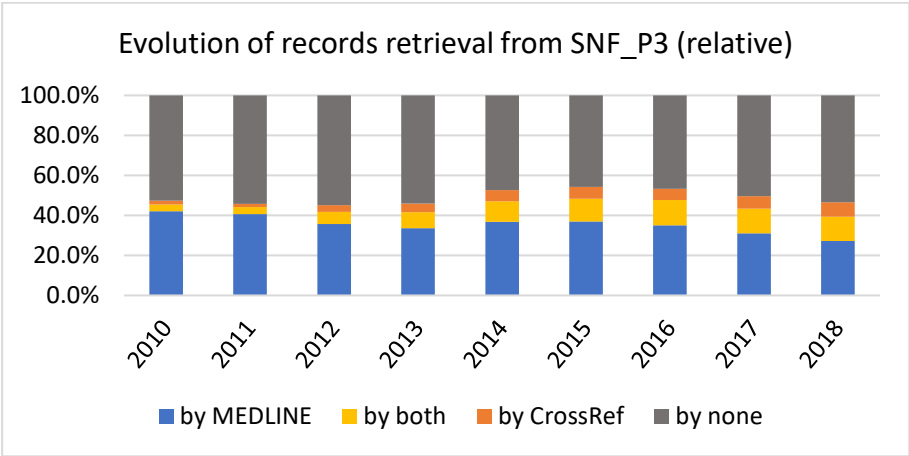

Fig. 12: evolution of the proportions of retrieved records from the benchmark #2, regarding each way.

As for the benchmark #1, the blue way supply is quite stable, while the orange one steadily increases since 2012.

The following figure also shows statistics in a relative perspective, but for the full-texts retrieved.

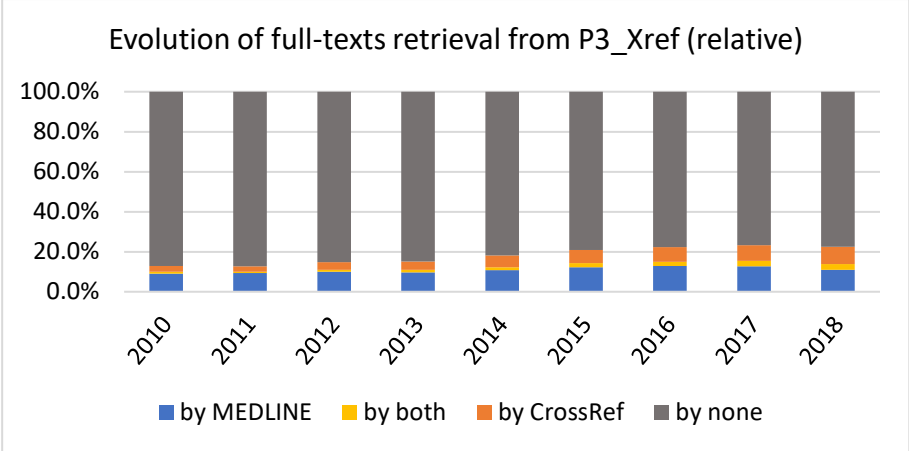

Fig. 13: evolution of the proportions of retrieved full-texts from the benchmark #1, regarding each way.

Once again, the blue way supply is quite stable since 2010, and the orange way one increases.

At last, the following figures show the percentage of records (left) and full-texts (right) retrieved by both ways, for publications since 2015.

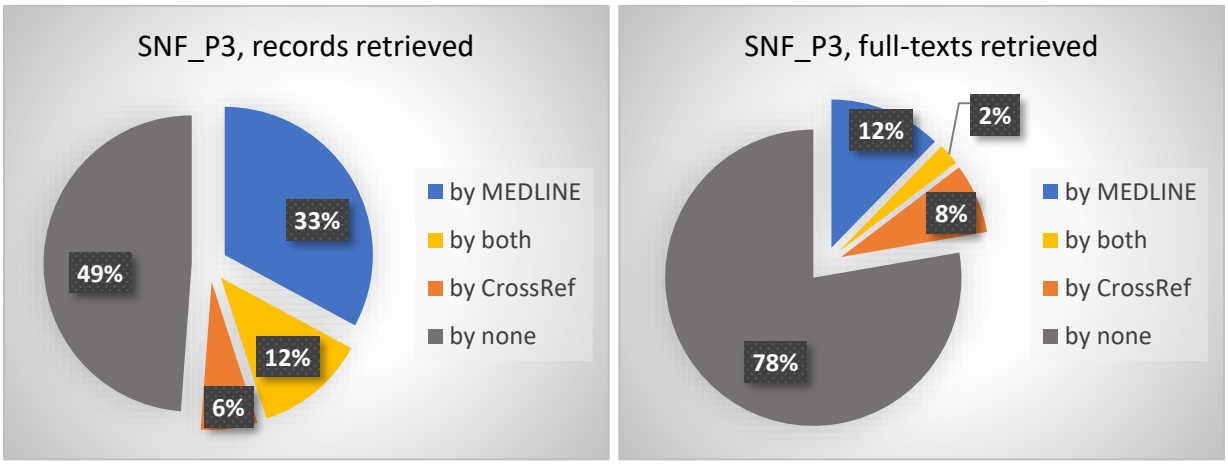

Fig. 14: proportions of recent (published since 2015) records and full-texts, retrieved for the benchmark #2.

For recent publications (since 2015), 51% of records contained in the SNF_P3 benchmark are retrieved by MEDLINE, CrossRef, or both. 49% of records in the benchmark are not retrieved. Dealing with Open Access, 22% of publications are detected as OA, with an accessible PDF file. Compared to the SNF_Xref benchmarks, lower proportions of records and full-texts are retrieved.

However, before drawing conclusions, we must deeper analyze not retrieved records, in order to appreciate what our pipeline misses, and why. Such analyses are made in section 6.

5.4. Benchmark #3: Swiss institutional repositories

The following table contains the results of our retrieval for each institution. The first column is the number of records harvested in this institution. Columns 2 and 3 contain the percentages of records and full-texts retrieved by our pipeline (blue + orange ways) for each repository. Finally, the last three columns contain the retrieved records not contained in the institutional repository, thus the potential improvement for it if our retrieved dataset was added. Note that this improvement is potential as, due to absence of DOIs, some publications may be counted as not retrieved while present in the institutional repository.

	records	records retrieved	full-texts retrieved	retrieved records not in IR	potential improvement	including full-text
Uni Zurich	23,418	60%	32%	6,984	+ 30%	50%
Uni Geneva	9,596	70%	32%	5,190	+ 54%	51%
HES-SO	639	23%	18%	567	+ 89%	66%
Uni Bern	14,515	63%	36%	5,383	+ 37%	45%

Tab. 2: evaluation for Swiss institutional repositories

HES-SO in an outlier in these results, possibly due to absence of DOIs and/or absence of publication types. For other institutions, results are significantly better than those obtained by the SNF benchmarks: between 60 and 63% of records retrieved, and between 32 and 36% for full-texts retrieved. The last two columns show that, if these institutions enriched their repositories with our retrieved dataset, they could potentially gain between +30 and +54% of bibliographic records, including numerous full-texts (for ~ 50% of these records).

6. Error analysis: the unreachable publications

In this section, we focus on the publications contained in the benchmarks which were not retrieved by our pipeline. We aim at understanding why the pipeline has failed. In particular, we manually analyzed a sample of 100 missed publications.

Affiliation metadata not available. As expected with the landscape (fig. 1), some missed publications seem to be unreachable for our pipeline, because the affiliation metadata is not available in our investigated databases: the publication is not in MEDLINE, and its affiliation metadata is not completed in CrossRef. Physics journals are the biggest contributors to these missing affiliations in CrossRef. Often, such as this publication (DOI 10.1002/hlca.201600283), the affiliation is even not available in the publisher's website, nor in any commercial database (Scholar, Scopus, or Researchgate). The only way to obtain the affiliation is to read the full-text. Thus, for Open Access publications, one possible automatic pipeline is to download all available full-texts, and to extract the affiliation thanks to text mining. This is the additional brute force pipeline described on section 7. But for non-Open Access publications, such as our example, the only way to extract the affiliation would be to acquire pirate version of the PDF full-text (e.g. via SciHub). Thus these non-OA publications really are unreachable if we base our retrieval on the affiliation metadata. Yet, it has to be noted that this article also is available (without affiliation metadata) in the Uni Basel repository.

Publications outside the perimeter of this study. Surprisingly, some missed publications are easily reachable in MEDLINE or CrossRef, but the affiliation metadata does not contain what we target. First, this is the case for Swiss but not targeted institutions, such as the CERN. But, especially for the SNF related benchmarks, many publications have an SNF mention for granting, but with affiliations from outside Switzerland. This is the case for mobility grants: for instance DOI 10.1038/s41467-018-05968-x, where all affiliations are american, but the SNF is acknowledged for an early postdoc mobility fellowship. Other publications acknowledge the SNF for granting, without detailed reasons (such as the DOI 10.1037/a0015922); it is possible that researchers upload publications prior to their involvement in a Swiss institution. Moreover, proceedings from Swiss conferences funded by the SNF are likely to appear with an SNF funding in CrossRef or in P3. One example is the 2018 International Conference on Optical MEMS and Nanophotonics (OMN) in Lausanne: 133 proceedings are in P3 and are missed by our pipeline, but all have foreign affiliations. Thanks to details in the P3 data, we can see that 5% of the 2018 publications in the SNF_P3 benchmarks have a “Proceedings” publication type. For this type, the performances of our pipeline are very low (less than 10% of retrieved records, instead of 51% on average). Thus, all errors described in this paragraph are “false errors”, and the performances of our pipeline for retrieving Swiss publications is probably better than what show the both SNF related benchmarks. This also is a limit of SNF related benchmarks.

7. Additional brute force pipeline

Missed publications have no available affiliation metadata in our investigated resources – and sometimes nowhere –, but many of them are Open Access. Consequently, the affiliation could be automatically extracted from their PDF files. We explored an additional brute force pipeline. This approach consists in downloading all available full-texts (possibly 4.5M per year, according to Jeangirard 2019), and to extract the affiliation thanks to a Text Mining solution.

7.1. Grobid

Grobid (https://github.com/kermitt2/grobid),(a machine learning library for extracting, parsing and re-structuring raw documents such as PDF into structured XML/TEI encoded documents with a particular focus on technical and scientific publications) was used to check the possibility of extracting the affiliation information from the pdf file. If the affiliation data were properly extracted, they were matched against our authority list.

An example Grobid output looks like below:

```
<affiliation key="aff0">
<orgName type="department" key="dep1">Département de Physique Théorique and
Center for Astroparticle Physics</orgName>
<orgName type="department" key="dep2">Departamento de Física</orgName>
<orgName type="institution">Université de Genève</orgName>
<address>
<addrLine>24 quai Ansermet, CH1211 Genéve 4</addrLine>
<country key="CH">Switzerland</country>
</address>
</affiliation>
```

For assessing the gain and the effort of this additional pipeline, we designed a sample of 1,367 DOIs which were retrieved by none of our resources. 821 (60%) were perceived as OA in Unpaywall. However, only 787 entries (57%) provided a location that could be used to download the relevant PDF file. For the remaining 34 entries, there were no indication of PDF link location, even though they were mentioned as open access. An example DOI which is open access, but has no PDF location can be checked using the URL below:
https://api.unpaywall.org/v2/10.1016/j.alit.2015.11.004?email=abc@gmail.com

The 787 entries with links for the location of the PDF file were used to download the PDF files. Downloading the file was done using wget command, with maximum 5 attempts.
Only 655 DOIs (48%) were properly downloaded. The remaining 132 were not downloaded, due to moved locations of the file, incorrect links and particular sites which did not allow automatic crawling. The retrieved 655 pdfs were parsed through Grobid, from which 614 entries were able to get accurate affiliation extractions.

Step	Number of resulting entries	Percentage
Initial entries	1,367	
mentioned as Open Access in unpaywall	821	60%
Entries providing a pdf location in unpaywall	787	57.5%
Entries downloadable via wget	655	47.9%
Entries parsed properly by Grobid	614	44.9%

Tab. 3: global results for Grobid parsing, on our exploratory sample of 1,367 entries.

Upto this step, 44.9% of the initial entries were preserved, through the pipeline. The resulting parsed output 614 entries were matched against the list of affiliations mentioned in section 4.a, to check whether at least one affiliation of the authority list can be found within the Grobid extraction. Upon observation, the affiliations extracted via Grobid could be categorized into 4 forms:

- No result: Grobid could not extract affiliations at all.
- NSA: Grobid succeeded in extracting affiliations, but there are no Swiss Affiliations in the PDF file.
- NL: Swiss affiliation but not in the authority list.
- GE: Grobid did extract the affiliations, but incorrectly (e.g encoding errors in text)
- Correct: Grobid succeeded in extracting affiliations and at least one of the affiliations in the authority list.
- NoResultsGrobid: Grobid did not extract affiliations at all.

The table below shows the results obtained for affiliation extraction for 614 entries.

Ouput	Entries with this output	Percentage
No result	94	15.3%
NSA	98	15.9%
NL	16	2.6%
GE	3	0.4%
Correct	403	65.6%

Tab. 4: detailed results for Grobid parsing.

As can be seen, 15.9% of accurately extracted entries by grobid contained affiliations that did not include any swiss affiliation, and 65.6% of the accurately extracted entries by grobid contained swiss affiliations mentioned in the initial list.

7.2. Extrapolated gains and efforts

Extrapolated gains. The following figures shows the proportion of retrieved records and full-texts by each resource for the SNF_Xref and UNIGE institutional repository. This is the same statistics than in the evaluation section, but with the extrapolated supply of the additional brute force pipeline with Grobid was added. The supply is estimated because it was measured on 5% random samples.

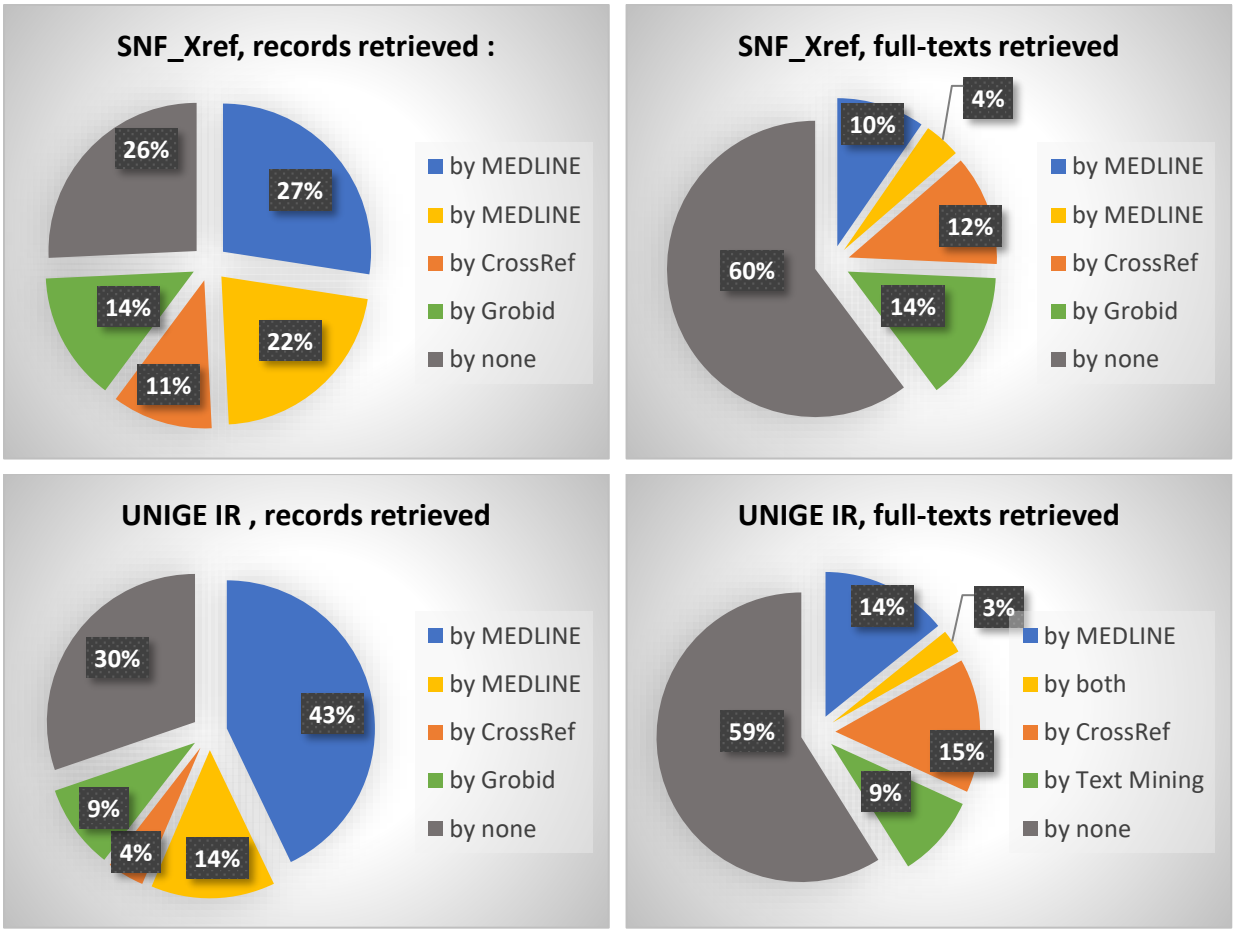

Fig. 15: proportions of recent (published since 2015) records and full-texts, retrieved for the benchmark SNF_Xref (top) and the UNIGE repository (bottom).

There is a clear way for improvement thanks to this additional brute force pipeline. It could improve the records retrieval in both benchmarks, with +14% for SNF_Xref and +9% for UNIGE IR (green parts on left pie charts). As Grobid only extracts affiliation from Open Access files, the same publications also improve the full-texts retrieval.

Efforts. The previous extrapolated gain must be counterbalanced with the processing effort. Indeed, the previous experiment consisted in downloading and apply Grobid to DOIs contained in the benchmarks and not retrieved. It was quite cheating, as we knew what DOIs were relevant.

In a real-case scenario, we need to download all available Open Access publications (PDF files) in Unpaywall, then to apply Grobid. For a sample of 100 DOIs, the processing time on a data server was around 15 minutes (12 minutes for downloading, 3 minutes for applying Grobid). According to (Jeangirard, 2019), there are approximately 4.5 millions of DOIs published each year. As 1 million per year should be found in MEDLINE, and approximately 37% of records have a completed affiliation in MEDLINE, we can hypothesize that between 2 and 3 millions of bibliographic records should be investigated each year (downloading the PDF when OA, and treating by Grobid). This represents a computing time of between 208 and 312 consecutive days per year for a single core, or between 13 and 20 hours each day. This computation time is a serious hurdle, but could be parallelized in order to be shortened.

8. References

Jeangirard, E. (2019, June). Monitoring Open Access at a national level: French case study.

Piwowar, H., Priem, J., Larivière, V., Alperin, J. P., Matthias, L., Norlander, B., ... & Haustein, S. (2018). The state of OA: a large-scale analysis of the prevalence and impact of Open Access articles. PeerJ, 6, e4375.

Putallaz, M., Schwob, E., & Ruch, P. (2018). Enrichissement des dépôts institutionnels suisses (No. TRMASID 16). Haute école de gestion de Genève.