HEG in SONAR

Bern, April the 3rd

WP5 - Recovering of full-text from 3rd-party OA (Feasibility study)

Dec 2018 - April 2020

Tasks overview

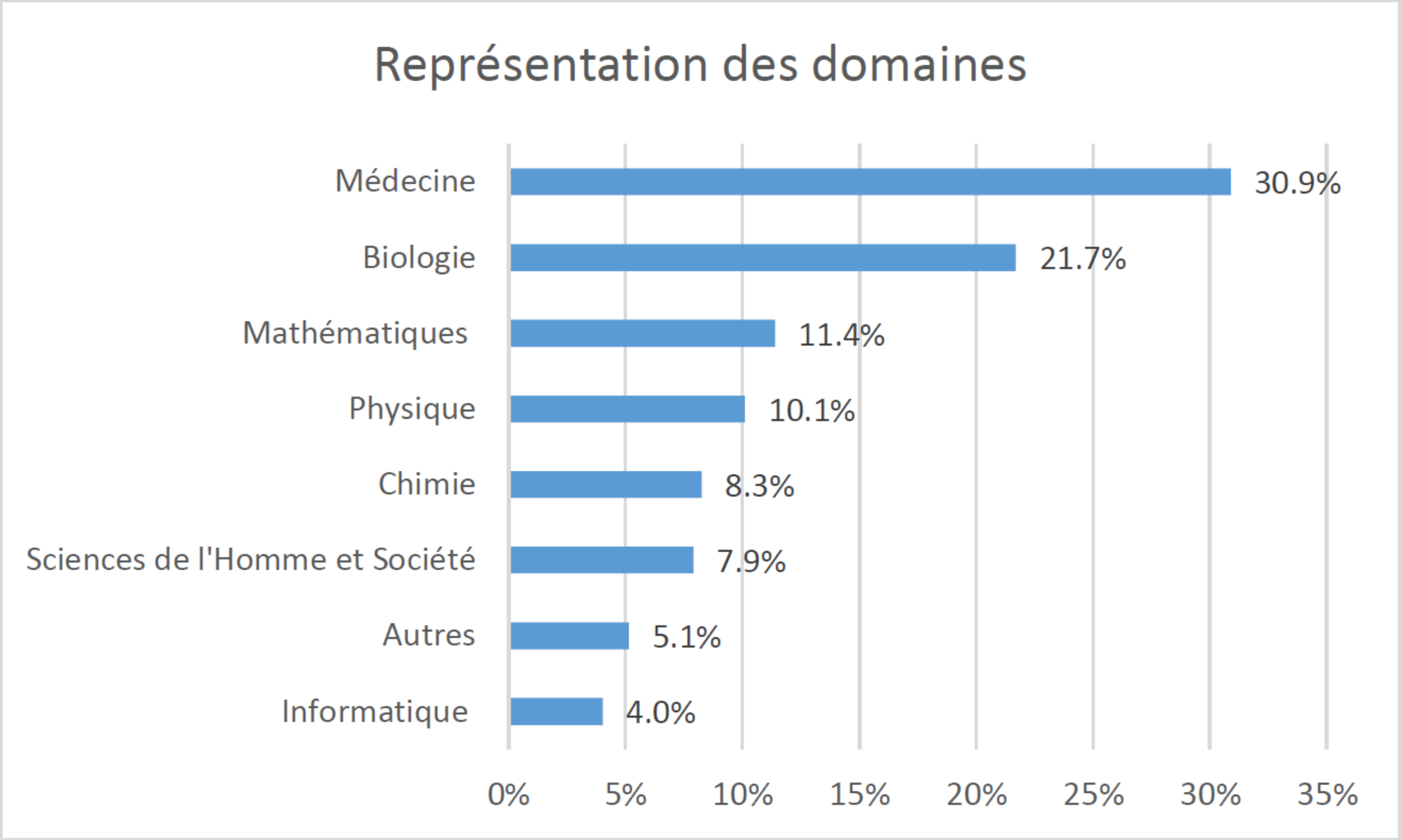
- Define a pilot subset of institutional archives
    - Content provider: PMC, CrossRef/OA, Unpaywall (Linksolver)
    - Completion: 60%
- Identify articles authored by swiss affiliated researchers
    - 40%
- Synchronization and evaluation
    - 10% (PMC)
- Personnel Define API for the harvesting of the full-text
    - 5% (PMC)

Investigated databases so far

- Crossref
  - 24 000 publications with normalized SNF funding metadata
  - 33% with affiliations
- Web of Science
  - **Fees**
  - 66 000 publications with «SNF» in funding metadata
- Unpaywall
  - 100M publications (worldwide)
  - 4% with affiliations
- MEDLINE
  - 28M publications (worldwide, biomedicine)
  - 100% with affiliations

Domain distribution CH [Schwob 2017]: PMC > 50%

Figure 6 : Répartition des domaines d'étude des articles



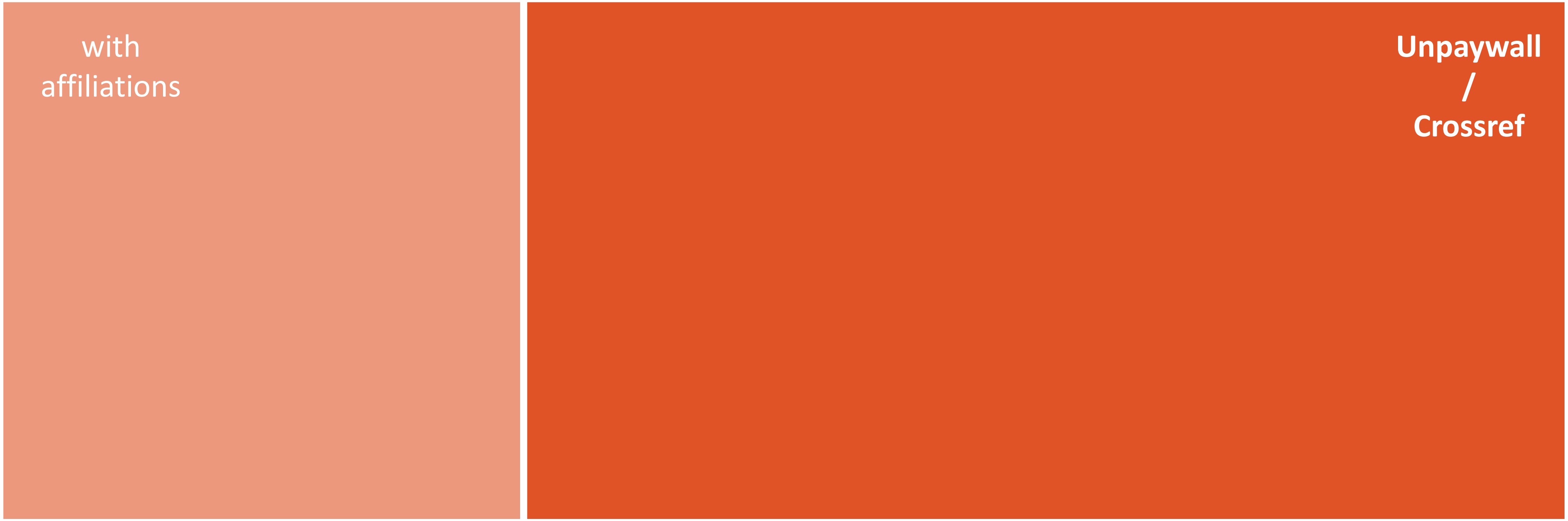
Current landscape

- Unpaywall / crossref : 100M [90% of overall publication ?]

Unpaywall
/
Crossref

Current landscape

- Unpaywall / crossref with affiliations : ~ 30% ?

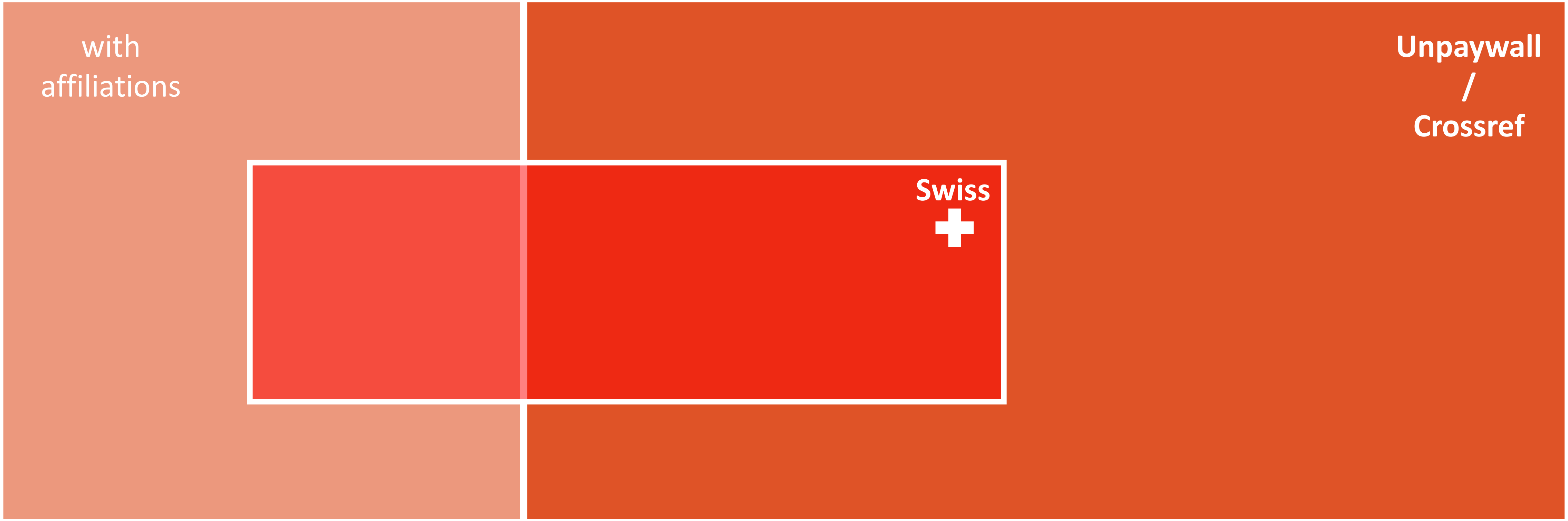


with affiliations

Unpaywall / Crossref

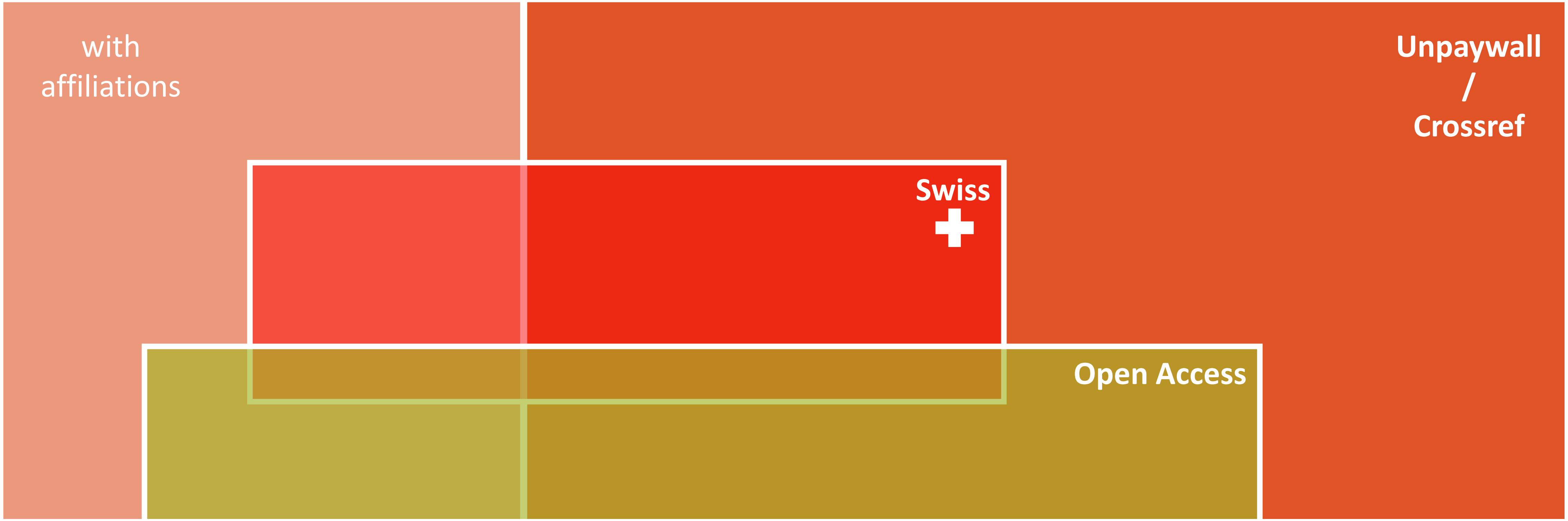
Current landscape

- Swiss research : ?? % (SNF funded in Crossref: 24 000 / in WoS: 66 000)



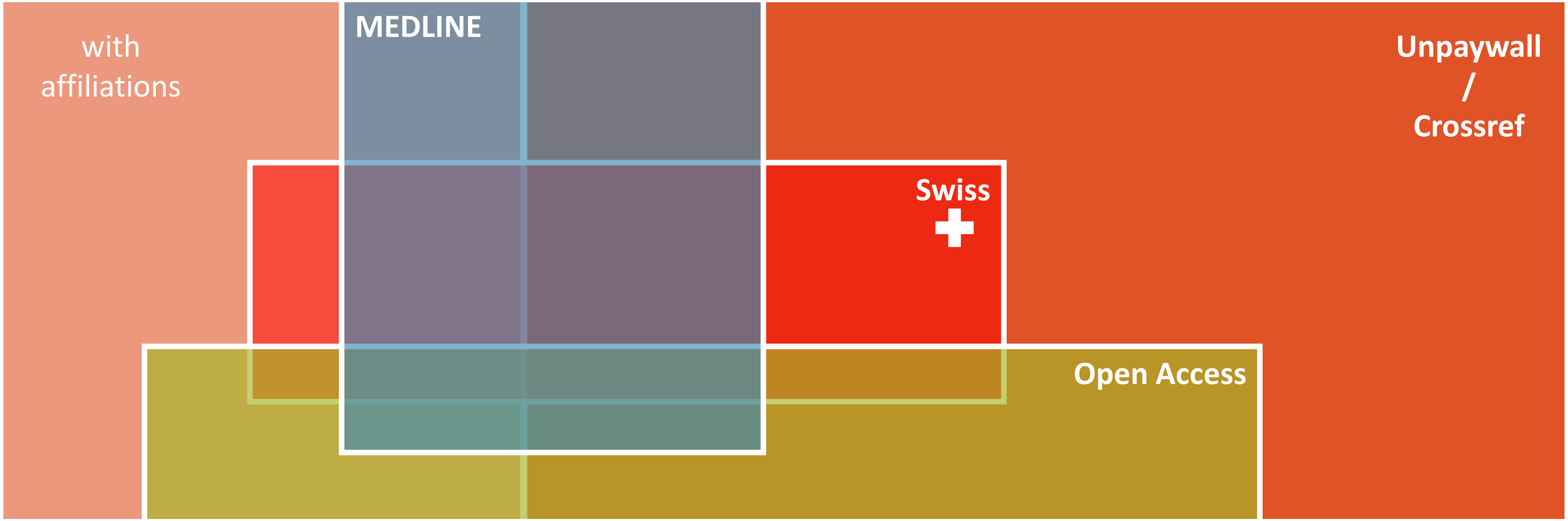
Publications landscape

- Open Access (pdf available) : ~ 10/20 % ?



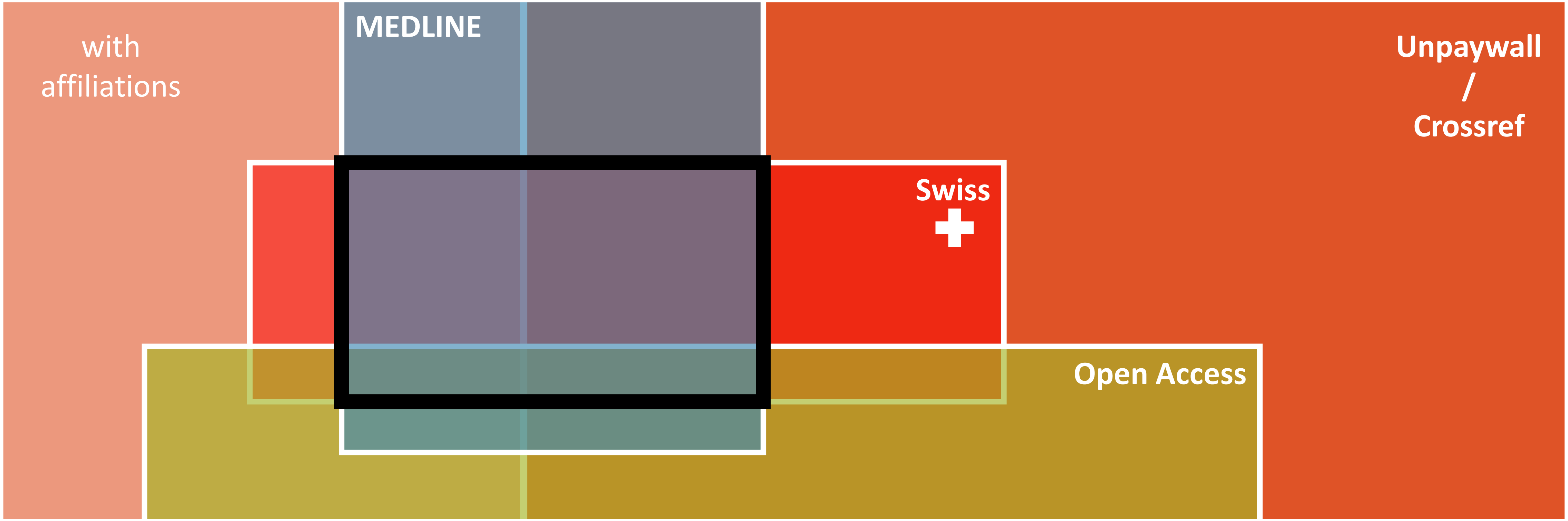
Publications landscape

- MEDLINE (affiliations 100%) : ~ 50% of Swiss publications ?



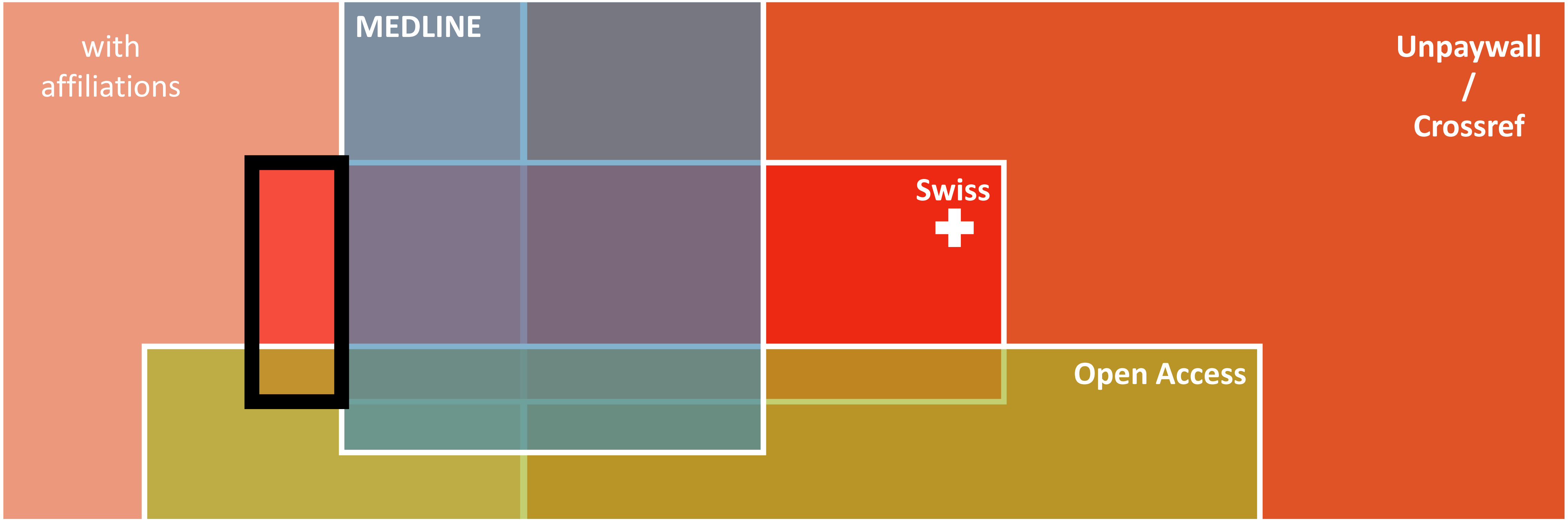
CH and MEDLINE

- ~ 50% of Swiss publications ? [Schwob]
- Records : reachable (100%) via MEDLINE
- Full texts : reachable (~ 10/20% ?) via DOI links (MEDLINE / Unpaywall)



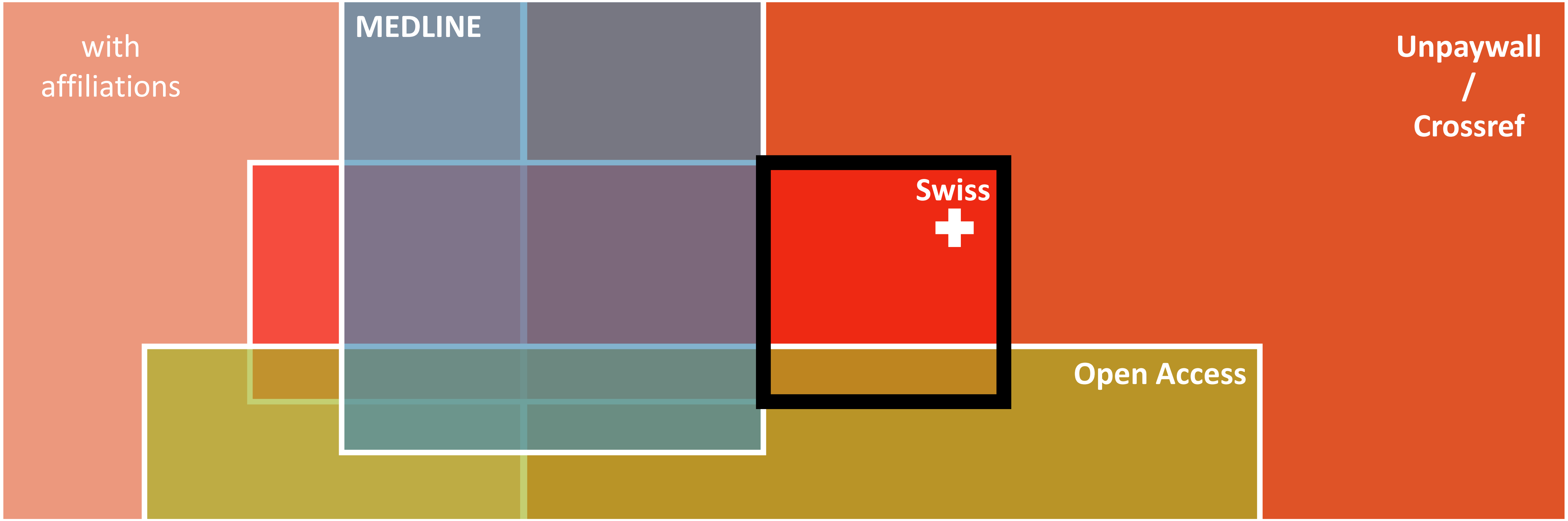
CH not in MEDLINE with known affiliations

- ~ 17% of Swiss publications ?
- Records : reachable (~ 100%) via Unpaywall / Crossref
- Full texts : reachable (~ 10/20%) via DOI links (Unpaywall / Crossref)

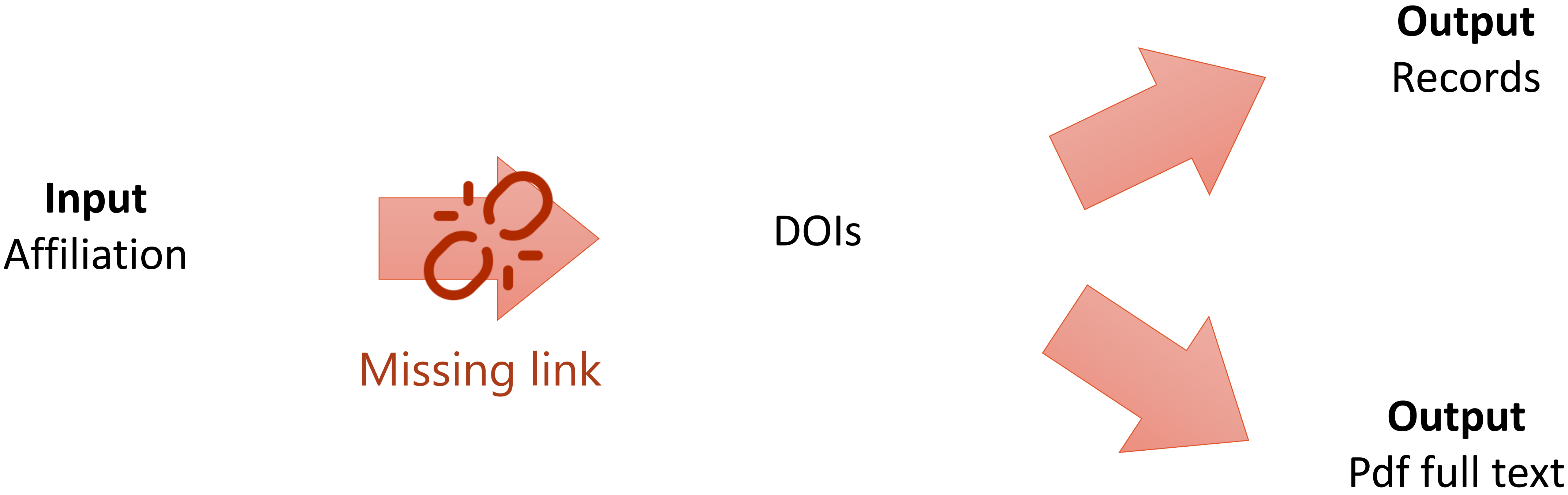


CH not MEDLINE with unknown affiliations

- ~ 33% of Swiss publications ?
- Records : reachable (~ ?? %) via SNF funding or affiliation extraction (Grobid)
- Full texts : reachable (~ 10/20 % ?) once the DOI known



Current pipeline



Input
Affiliation

Missing link

DOIs

Output
Records

Output
Pdf full text

Additional KB

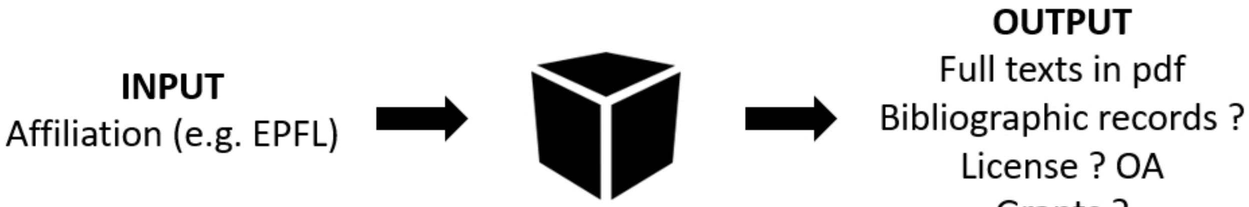
- Semanticscholar.org
  - Promising for affiliations
  - Open science

- Publications without DOI → suggest to leave it for now !

- Affiliation and grants extraction from pdf via grobid
  - Current status

Necessity of extracting details

Final workflow



INPUT
Affiliation (e.g. EPFL)

OUTPUT
Full texts in pdf
Bibliographic records ?
License ? OA
Grants ?

Pic. Courtesy: Julien Gobeill

Status

- **Started with crossref**
  - N=100 → 10000
  - 32% provided with affiliation field
  - Grant field not available most of the time

PDF distribution

- DOIs driven harvesting of PDF

- 72% are not available as full-text
- 28% are available as full-text (licenses to be checked ?)

- From the full-text, 17% can be harvested straightaway
- 11% require scrapping
  - located as “links” within the original page
  - Patterns is journal-specific…

- From the full-text, 17% can be harvested straightaway

Harvesting PDFs from Unpaywall

- Bulk the Unpaywall database (20 M articles, 95 GB)

- **17-28% → 74%**

- Next steps
  - Check affiliation if available in CrossRef
  - Affiliation information extraction from PDFs
  - GROBID

Information Extraction

- Grobid was deployed as a service in a local machine.
- Sample affiliation extraction

```
<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>
```

- Out of the harvested pdfs, **Grobid succeeded in extracting affiliations of 85% of the documents**
- Accuracy to be estimated

Grant information extraction

- **Information: e.g. Funding body + grant number, see AKOA KPI**
- **Found in several sections: footnotes, acknowledgements, …**
- **Combine GROBID with PDF2Txt + RegEx/Dictionaries**

- **<div xmlns="http://www.tei-c.org/ns/1.0"><head>Acknowledgments</head><p>We acknowledge Prof. Yi-Cheng Zhang and Yi-Xiu Kong for fruitful discussion and suggestions. This work is partially supported by the EU FP7 Grant 611272 (Project GROWTHCOM), the Swiss National Science Foundation (Grant Nos. </p></div>**

MEDLINE thread

- 28 M MEDLINE records locally indexed + MongoDB storage

List of authorities for affiliations search

1. Search in Crossref and Medline via API in the affiliation field

   Search in previous work of the Consortium : lists of some publications of swiss institutions with affiliations : https://consortium.ch/open-access/?lang=en

2. Try to figure out the right combination of words to find the maximum of publications by affiliations, e.g. :

"Lausanne university" allows to find →

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

3. Compile a list of all the potential combination for all the affiliations of the institutions listed in SNF-P5 : http://p3.snf.ch/ in order to find the most papers for SONAR (exhaustiveness was not intended, of course…)

T2.4 – Data model (analysis phase)

Current considerations for authors

- Identified (ORCID) and not identified authors
  - two entities in DB ?

Issues

  - Consistency across publishers (Ruch Patrick vs Ruch P)
  - Homonymy (Müller F)

- Affiliations
  - Linked to articles + temporal drift
  - Provided affiliation vs normalized affiliation
  - Multilingual and diacritics issues (Geneva vs Genf, Geneve vs Genève)

Additional questions…

- has_pdf tag ?
- Embargo date ?
- peer-reviewed tag ? MEDLINE tag ?
- Document type ?

- Grants : only SNF ? Other ?

- Publications uploaded by authors :
  - Metadata edited / imported by authors ?
  - Pre-print vs OA published paper ?

AKOA KPI

Document type (Proceedings, Book, Article, Book chapter, …)

Funding source

OA types

Embargo period

APC

Harvesting frequency