

CRISolid

CURRENT RESEARCH INFORMATION SYSTEM

AI for metadata quality improvement in a CRIS: automating data processes for researchers' benefit

David Reymond, IMSIC, Université de Toulon

Joachim Dornbusch, DIREVAL, Université de Paris 1 Panthéon-Sorbonne

euroCRIS Strategic Membership Meeting 2024 SSM2024, nov 2024



Community driven institutional CRIS

- Modular, open source, documented...
- Decentralized way to feed content: researchers in the loop
- SoVisu+ as an interface for researchers to curate and consolidate their own publication data
- Problem addressed here: when gathering multiple databases, duplicates are created. How to maximise deduplicating process through eco-friendly IA?

Reymond, David, et Alaric Tabariès. 2024. « Decentralized research data management: introducing SoVisu+ ». *Procedia Computer Science* 249:216-23.

<https://doi.org/10.1016/j.procs.2024.11.067>.



CRISolid

Update from EuroCRIS 24 in Vienna

L'ÉCOLE
DES HAUTES
ÉTUDES EN
SCIENCES
SOCIALES



Publication side: setting up an institutional CRIS

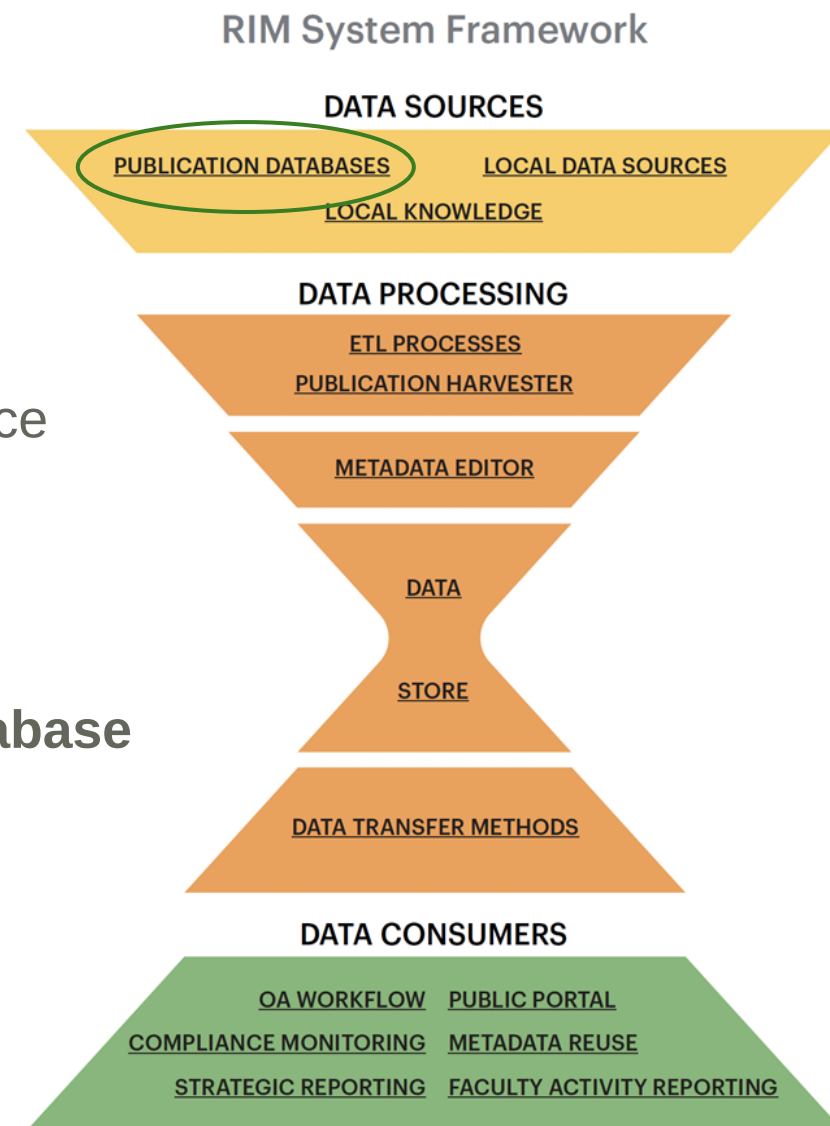
Various ways:

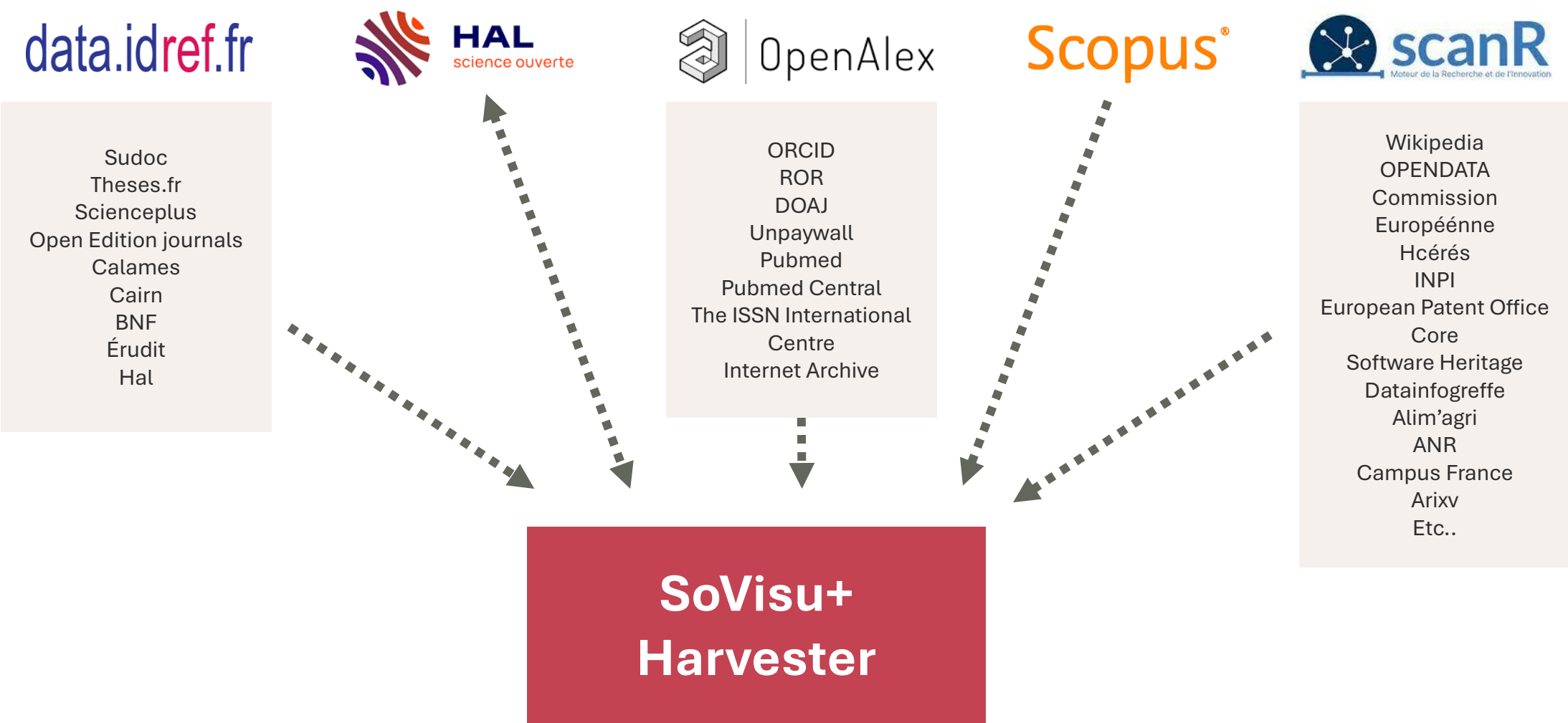
Option 1 – Use a Single Platform as database reference

Option 2 – Restrict to DOI-Cited Publications

Option 3 – Aggregate from several agregators/database

Bryant, Rebecca, Janet Fransen, Pablo de Castro, Brenna Helmstutler, et David Scherer. 2021. « Research Information Management in the United States: Part 1—Findings and Recommendations ». <https://doi.org/10.25333/8HGY-S428>.





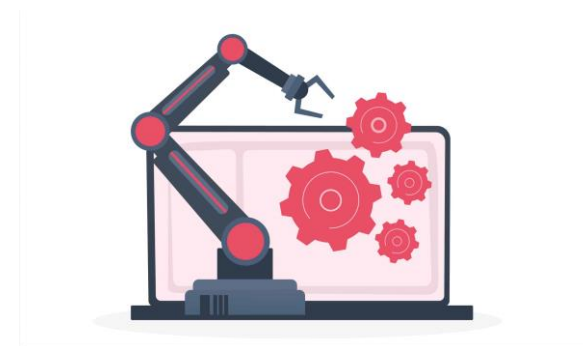


How to consolidate bibliographic data?

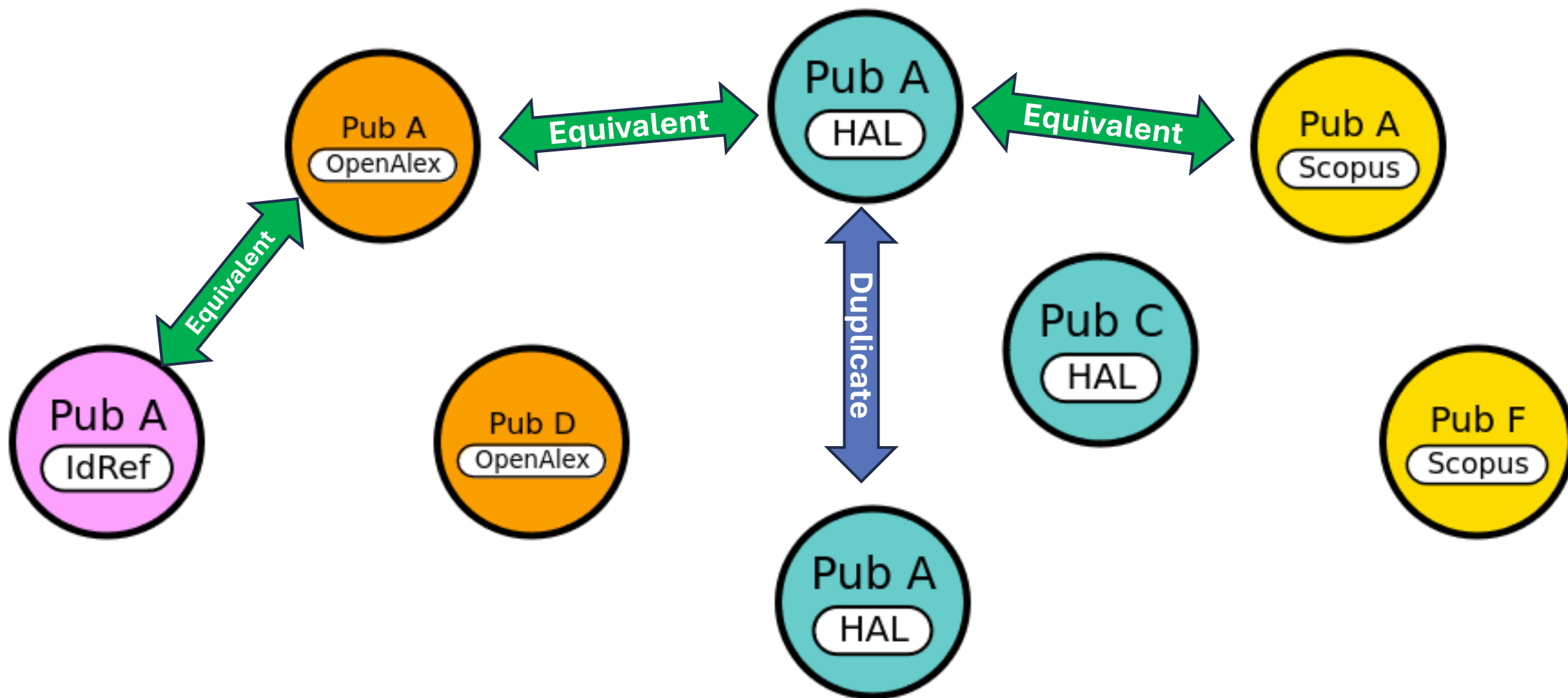
When gathering researchers Id's, the collected data includes:

- Duplicate References
- Equivalent References

Can we process them automatically and propose an acceptable result?

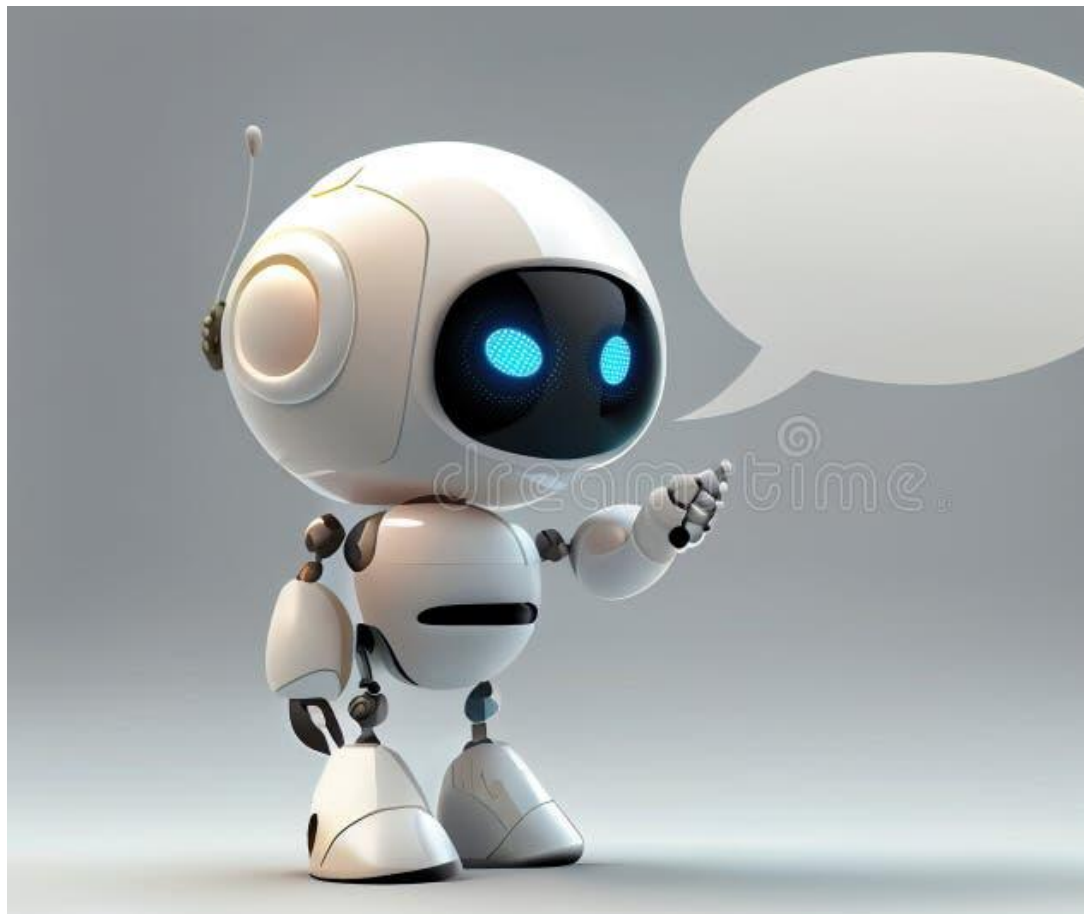


How to consolidate bibliographic data?





Of course, LLMs can do it



- Write a prompt with data with two records

<https://chatgpt.com/share/4d91465a-29f9-4e6f-9b3f-3060927e24d8>

- Yes... But no!
- Huge problem of complexity



Two kinds of problems

Problem of scalability

1. Brute force option
 - Using the API (\$\$)
2. Combinatory considerations

100 000 references = nearly 5 billions calls:

$$\text{Comparaisons totales} = \frac{N(N - 1)}{2}$$

Problem of alignment with business criteria

Concept of Work or Distinct Tasks

- Zero shot → Few shots

Pb. Token cost

- Fine tuning

Pb. Cost of fine tuned model

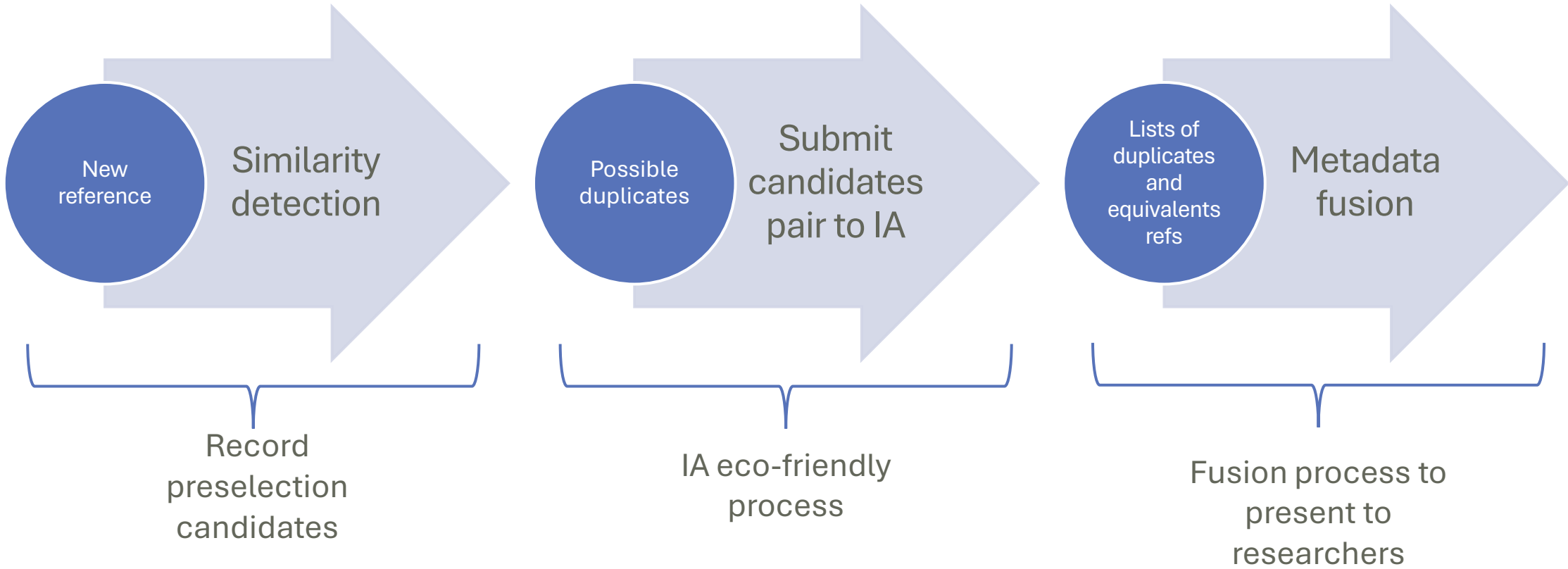
- LLM open-source + fine tuning

Infrastructure

ecoresponsability problem !

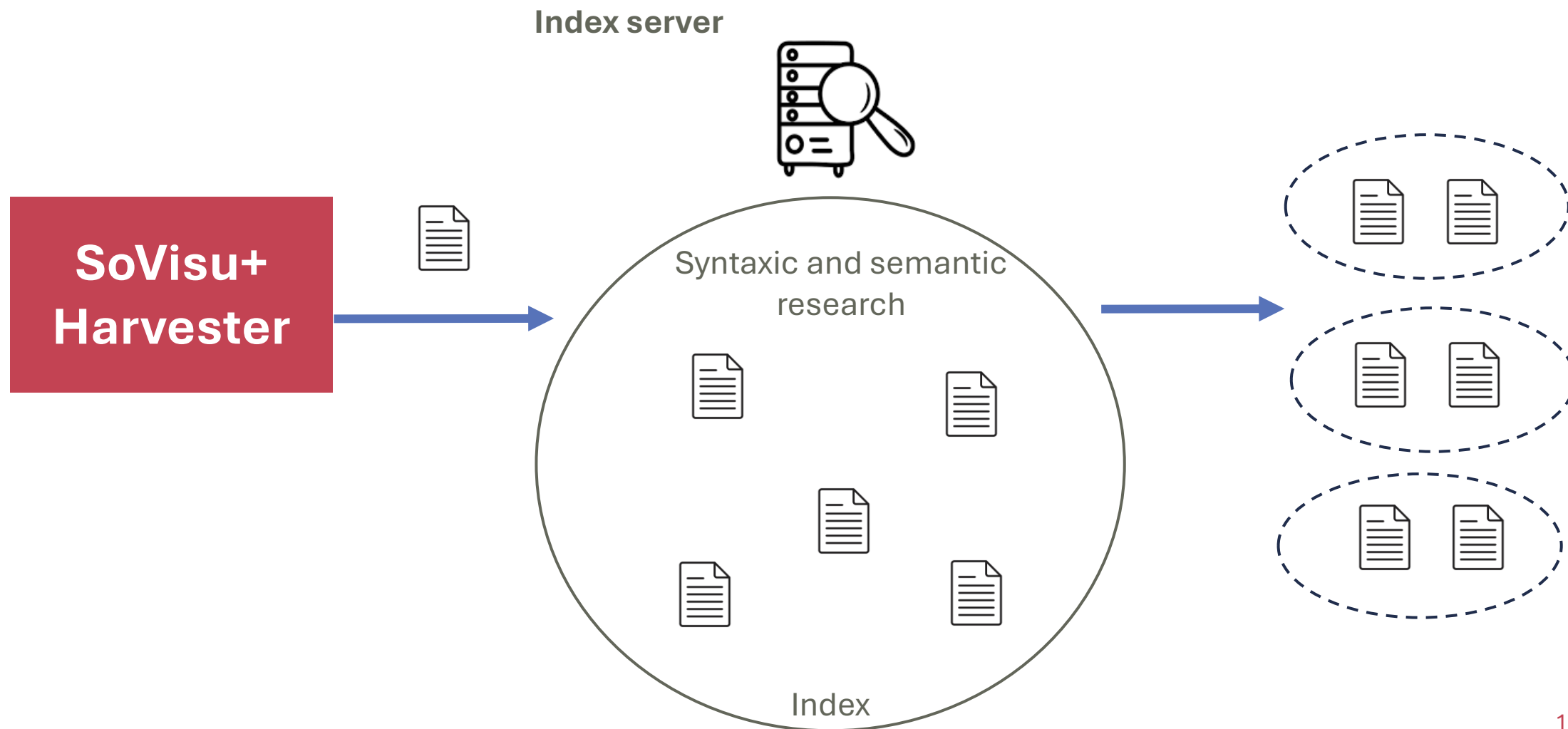


Solving the Problem in three steps

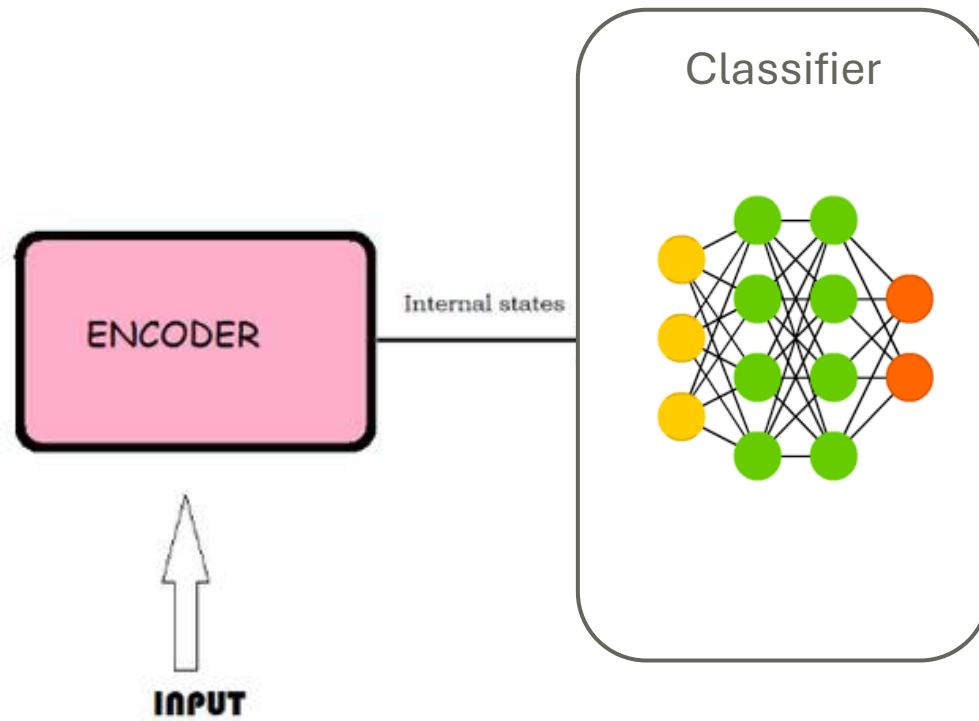




Preselection of duplicates records candidates

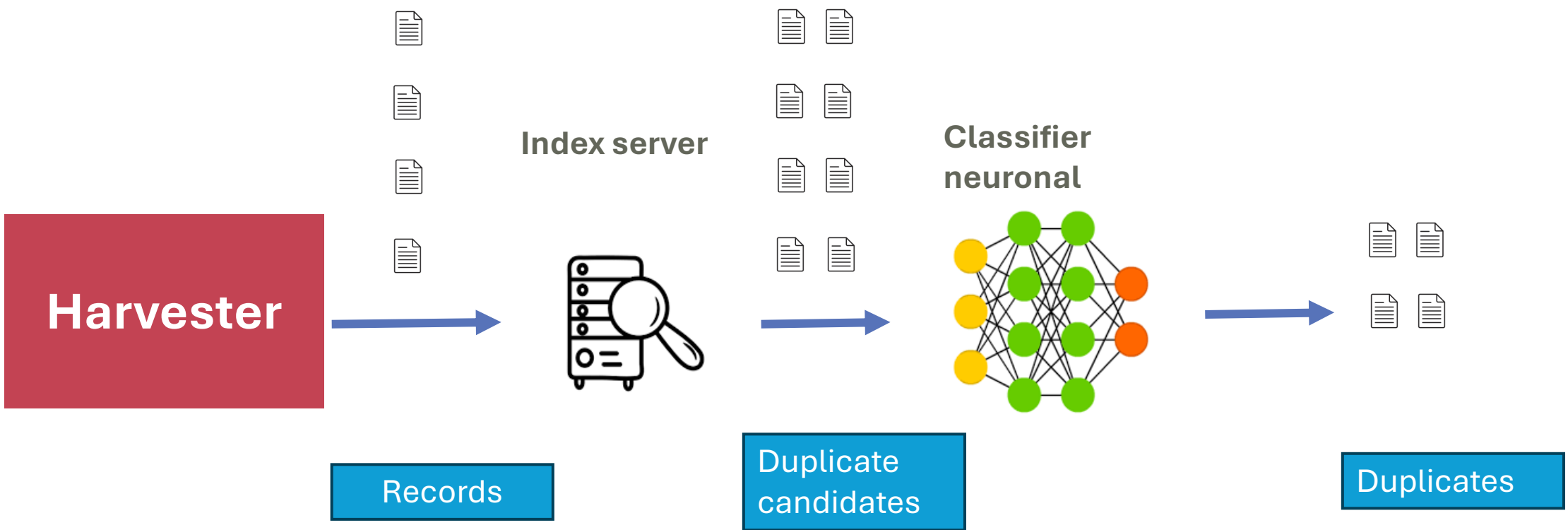


Preliminary step: train a model



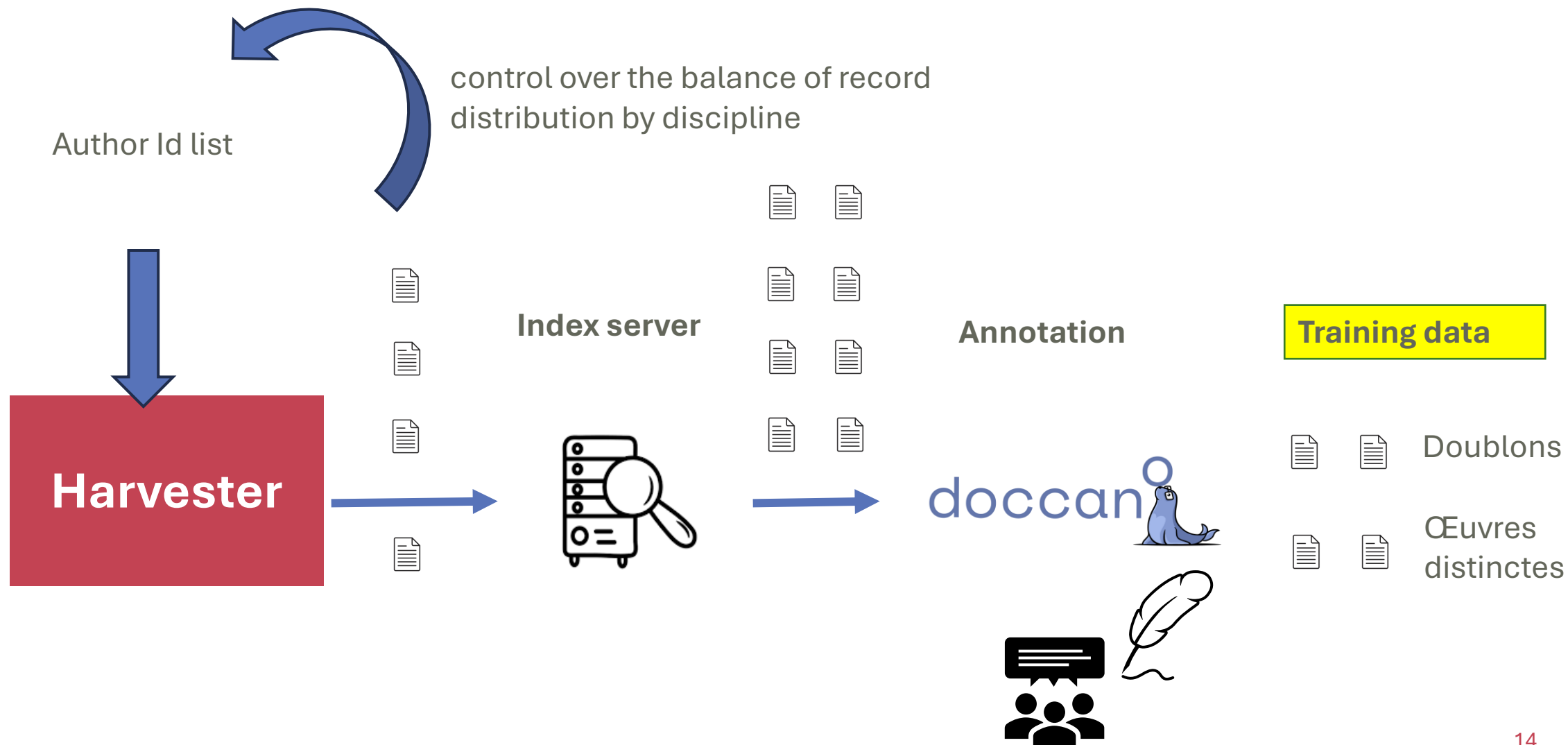


Target pipeline





Creating the corpora for training: the pipeline





Involving the community

- Target : librarians as business' experts
 - Email on dedicated national lists: call for participation
 - Workshop (80 participants)
 - 25 volunteers
- First preparation stage:
- doccano usage,
 - how to decide (no external information check),
 - tree labels (duplicate, not duplicate, unknown)
- Loopback checkpoint and generation of a new corpora (starting now)



The interface for annotators and admin

☰

Test - Données d'entraînement pour la détection de doublons

CRISolid training data

EN ▾

Projects

⋮

Actions ▾

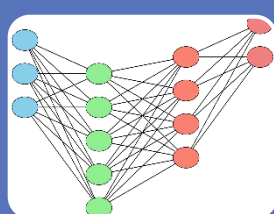
Delete

Delete All

🔍 Search (e.g. label:positive)

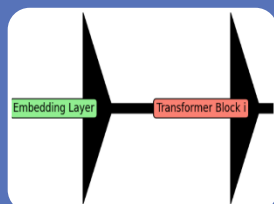
☐	Status	Text	Metadata	Assignee	Action
☐	Finished	Modes d'ajustement à la crise : des relations professionnelles sous tensions	Effectuée dans le cadre des post-enquêtes à l'enquête REPONSE (2010-2011), cette recherche a pour objet de mettre au jour les processus d'ajustement des établissements face à la crise, les caractériser...	Idessaivre ✕ ▾	Edit Annotate
☐	Finished	Metals and Civilisations	No abstract available	anoel ✕ ▾	Edit Annotate
☐	Finished	The Development of Coastal Polities in the Upper Thai-Malay Peninsula in the Late First Millennium BCE	No abstract available	mbereau ✕ ▾	Edit Annotate
☐	Finished	The Development of Coastal Polities in the Upper Thai-Malay Peninsula in the Late First Millennium BCE	No abstract available	anoel ✕ ▾	Edit Annotate
☐	Finished	The Development of Coastal Polities in the Upper Thai-Malay Peninsula in the Late First Millennium BCE	No abstract available	ymosset ✕ ▾	Edit Annotate
☐	Finished	The Inception of the Trans-National Processes between the Indian Ocean and the South China Sea from an Early City-State on the Thai-Malay Peninsula (Fourth-	No abstract available	lpillet ✕ ▾	Edit Annotate

Next steps: compare, implement, open for improvements



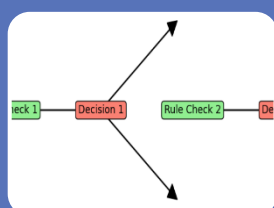
Training a neuronal classifier

- Performance measurement
- costs



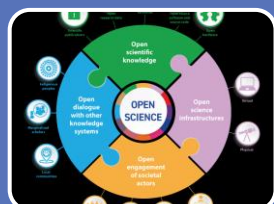
Training a LLM

- Performance measurement
- costs



Rules based process

- Performance measurement
- Costs (?)

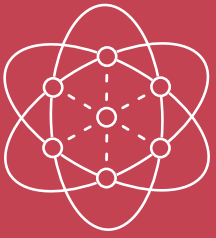


Open the whole process

- Publication of corpus and its operational creation
- Hackathon?
- Other languages as records are distributed in FR an EN?

Compare, implement
best choice

Share, improve,
internationalization, ...



CRISolid

CURRENT RESEARCH INFORMATION SYSTEM

Questions?

Thank you for your attention

david.reymond@univ-tln.fr

Joachim.Dornbusch@univ-paris1.fr