

CRISolid

CURRENT RESEARCH INFORMATION SYSTEM

AI for Duplicate Record Detection in Current Research Information System

AI4CRIS - Thursday, April 24th 2025

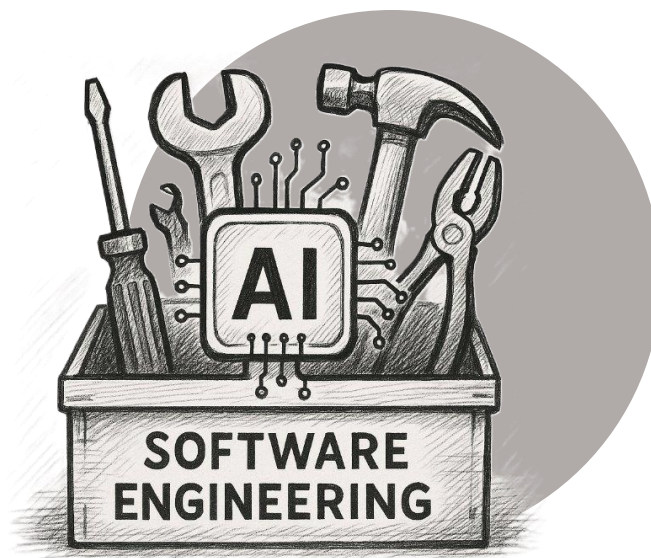
Joachim Dornbusch. Université Paris 1 Panthéon-Sorbonne

Service d'ingénierie numérique pour la recherche



Modern technologies for a brand new consortium

- Software development frameworks / ETL
 - o Python FastAPI
 - o Next.js
 - o Apache Airflow
 - Micro-services architecture :
 - o Service oriented
 - o Message oriented
 - Cloud native technologies :
 - o Docker
 - o K8S
 - o CI/CD pipelines
 - **Transformers**
 - o **SLM (BERT or derivated)**
 - o **LLM**
- We just see AI as a new element in the software engineering toolkit.



SLM/LLM are no longer just research, they are the state of the art in IT. A today's project that doesn't use transformers to solve certain problems will quickly become outdated.

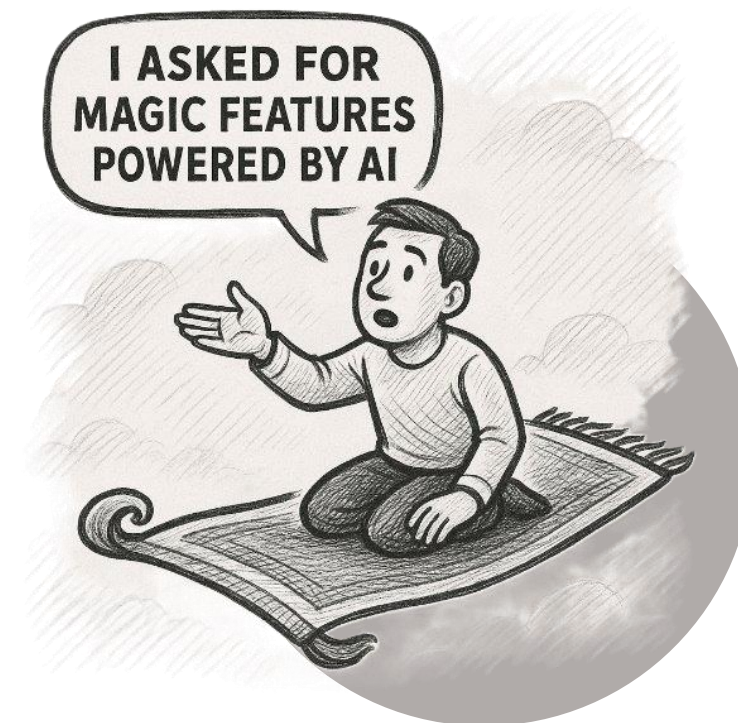


Example of Problems AI Can Solve in CRIS

- Extract missing publications from researchers' CVs, parse metadata
- Identify potential collaborations on projects
- Suggest topics from large controlled vocabularies
- Detect duplicates without shared identifiers — *today's focus*

≠

- *Build a chatbot (Also part of our work)*





The importance of duplicate record detection

The problem to solve : to help research institutions overcome the choice between commercial turnkey tools and the fragmented ecosystem of open platforms

Commercial publication indexes :

- Expensive
- Bad coverage of certain domains / countries

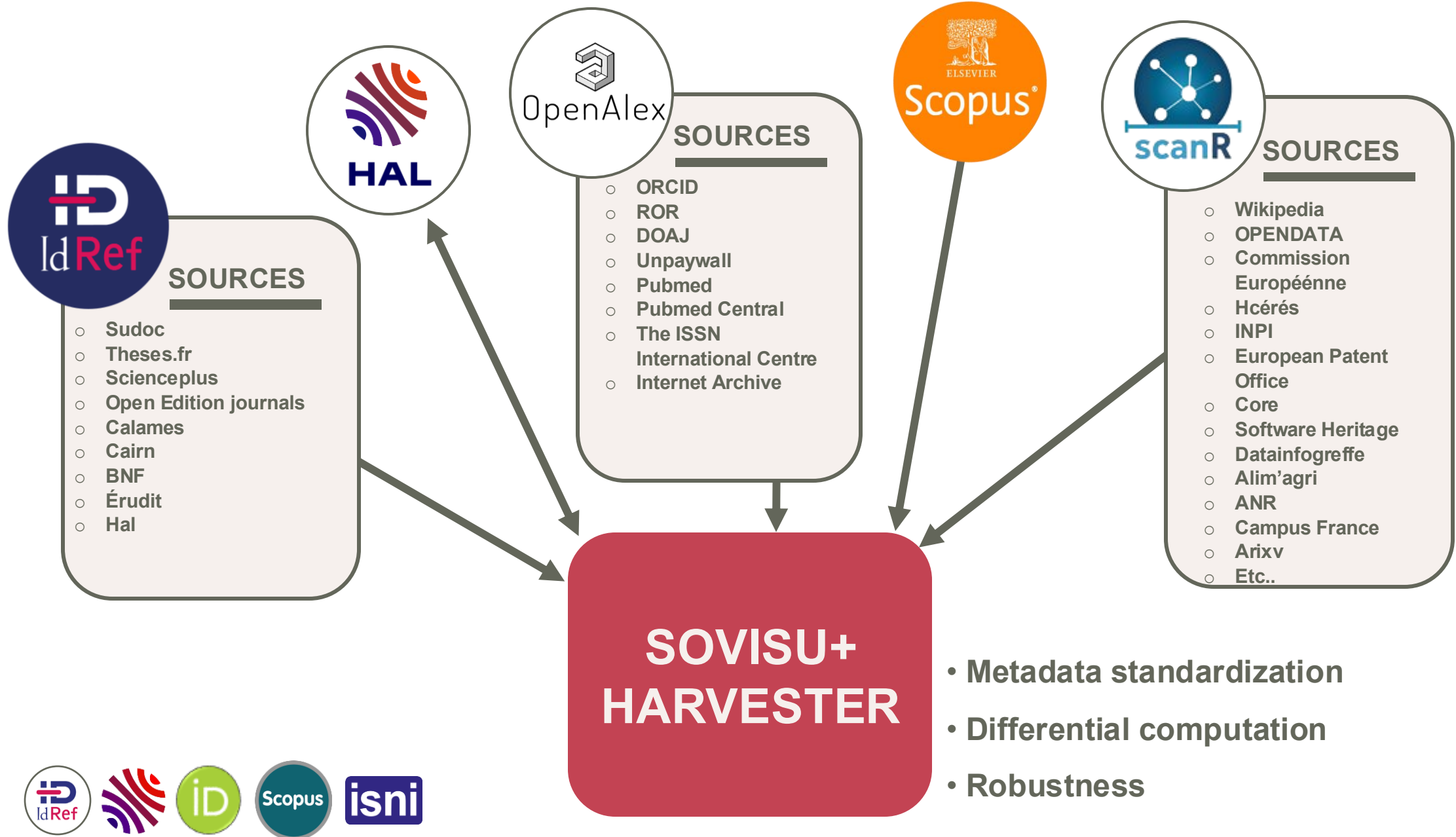


Non commercial publication indexes :

- Hal
- Data.idref.fr
- ScanR
- OpenAlex
- ...

Challenges:

- Representation formats (almost no standard)
- Researcher identifiers used
- Data stability and quality



Multi-platform harvesting generates non-trivial duplicates

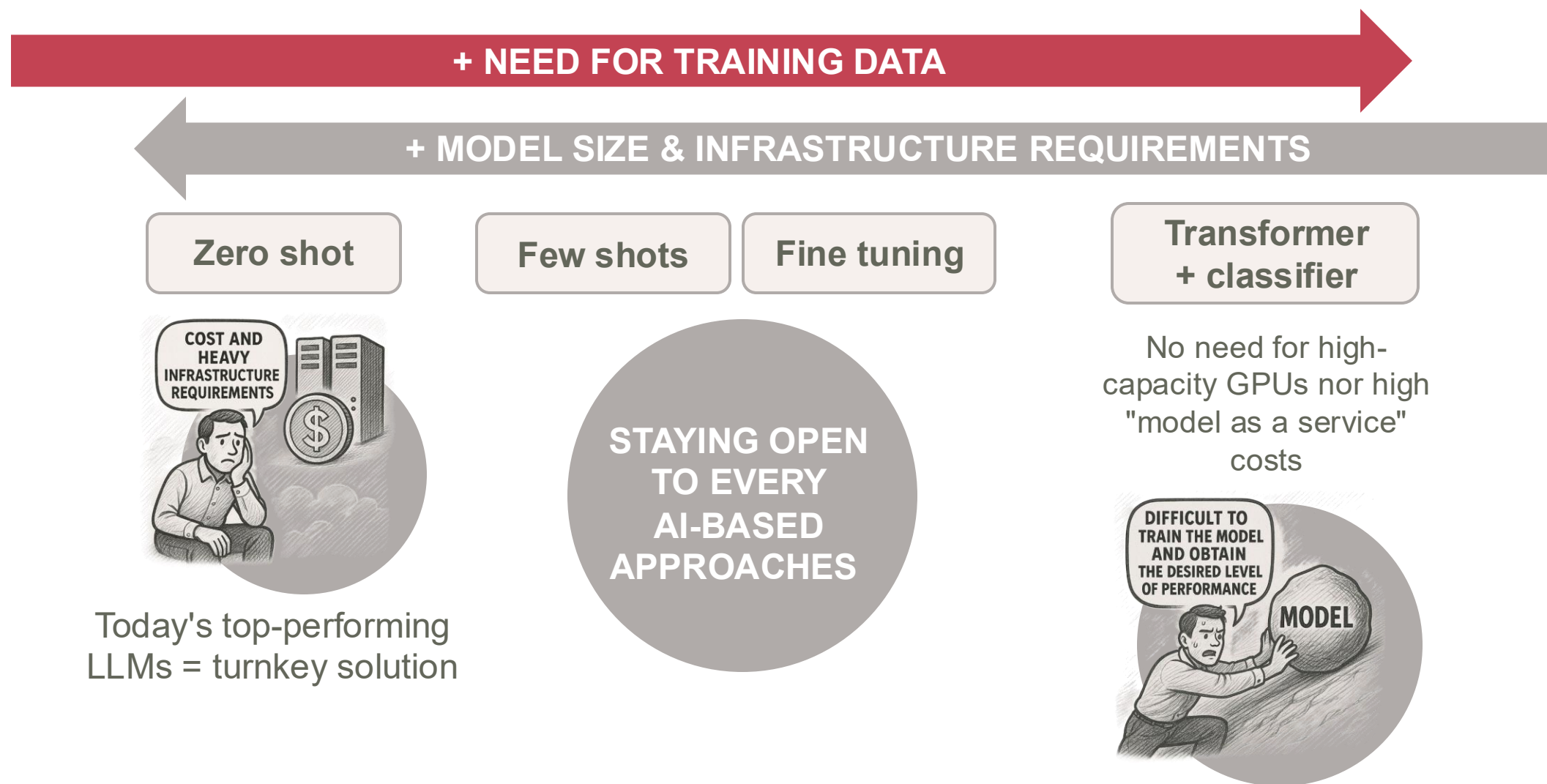
type	titre	auteurs	Da...	paru dans	Statut hal	Sources	Actions
Filtrer par type	Filtrer par titre	Filtrer par auteurs	Min	Filtrer par paru ...	Filtrer par Statut...	Filtrer par Sour	
	Qu'est-ce qu'une démonstration ? fr	Alberto Naibo, Jean Fichot	05-27-2021			 	...
	Qu'est-ce qu'une démonstration ? fr	Alberto Naibo, Jean Fichot	05-27-2021			  	...

In this case, it is possible that the publication was filed twice on Hal and that OpenAlex's harvesting amplified the problem.

	Economic thought in scholasticism en	Irina Chaplygina, André Lapidus	01-01-2016				...
	Economic thought in scholasticism en	Irina Chaplygina, André Lapidus	07-19-2016			 	...

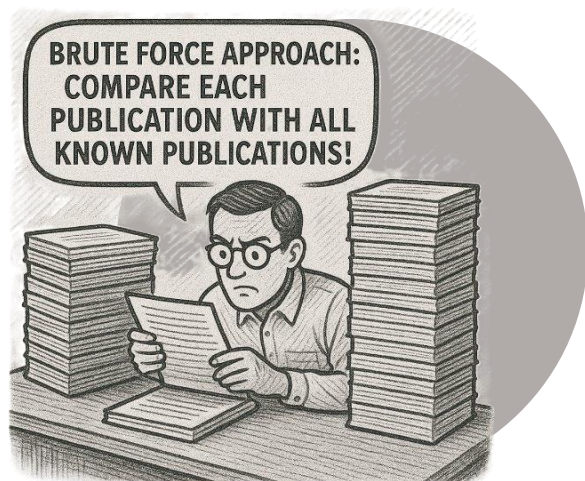
In this case, the same publication has been harvested by OpenAlex from both Hal and IDEAS/Repec.

Question 1. The trade-off between model capacity and training needs





Question 2. The combinatorial issue



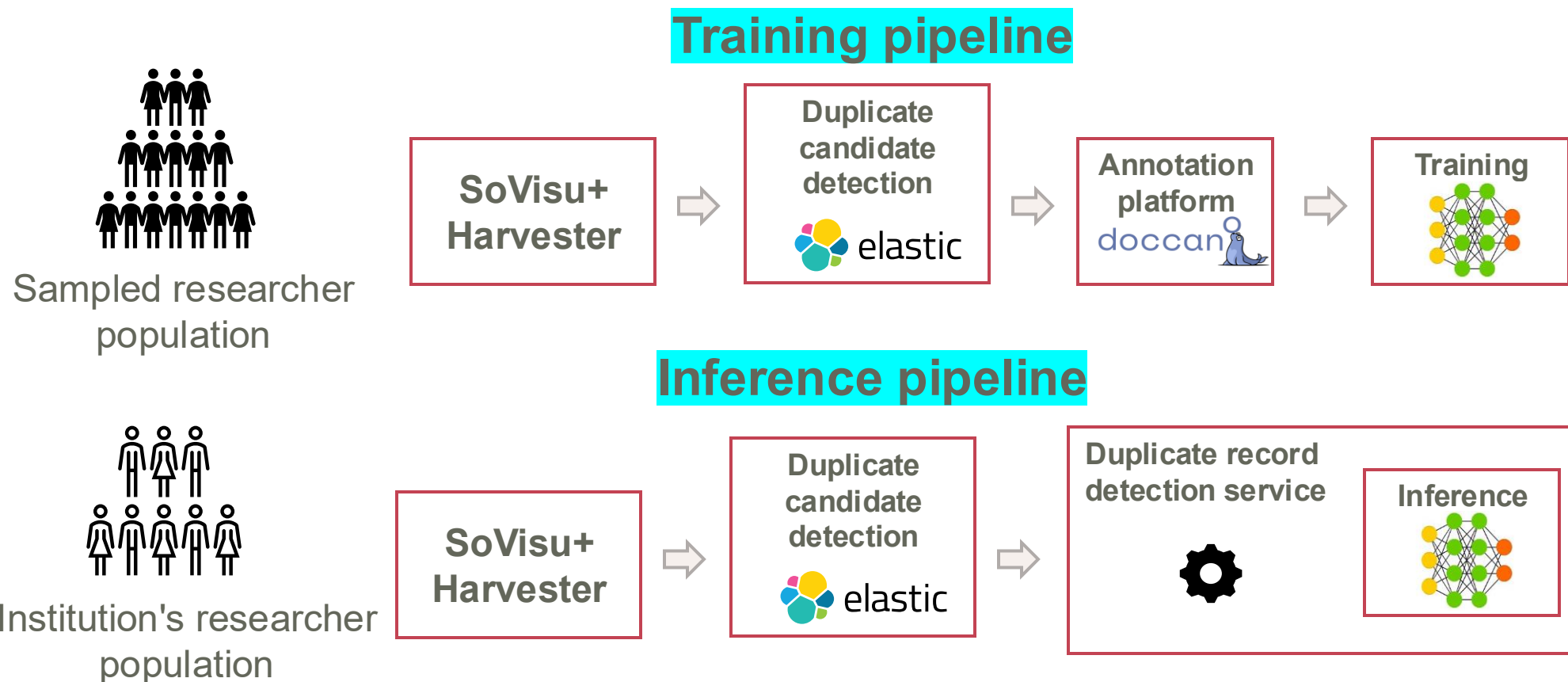
- $O(n^2)$ complexity
- Increasing costs over time
- Especially if combined with a heavy detection algorithm

Need for prefiltering mechanism

- Lightweight "candidate pair selector" (or "potential duplicate selector")
- Part of **SovisuPlus training data**
Modular, parallel, multi-strategy duplicate detection engine
- Each strategy searches for "similar" records using Elastic's built-in mechanisms
- Empirically determined threshold



Question 3. How do we handle the «feature training-inference» issue ?



To avoid a gap between inference features and training features, both the training pipeline and the inference pipeline are built with the same components, around the same data bus



CRISolid

Question 4. How to support and lead the annotator community

Collaborative
annotation platform



<https://training-data.crisolid.org>

✓

🔍

↕

📄

💬

✎

🗑

📱

☰

☰

Doublons ou équivalents0

Œuvres distinctes1

✓ Incertain2

Field	Reference n°1	Reference n°2
	OpenAlex https://openalex.org/W39138945	ScanR halhal-01538901
Title(s)	The Development of Coastal Polities in the Upper Thai-Malay Peninsula in the Late First Millennium BCE	The Development of Coastal Polities in the Upper Thai-Malay Peninsula
Subtitle(s)		
Abstract(s)		
Subjects	archéologie, geography, linguistics, philosophy, history, Antiquité, péninsule, Malay, péninsule Malaise	péninsule Malaise, polity
Document Type(s)	Chapter	Chapter
Contributions	T. Davies Pryce, role: https://id.loc.gov/vocabulary/relators/aut.html Bernard Bernard, role: https://id.loc.gov/vocabulary/relators/aut.html Cristina Castillo, role: https://id.loc.gov/vocabulary/relators/aut.html Boonyarit Chaisuwan, role: https://id.loc.gov/vocabulary/relators/aut.html Sophie Peronnet, role: https://id.loc.gov/vocabulary/relators/aut.html Sachipan Srikanlaya, role: https://id.loc.gov/vocabulary/relators/aut.html Brigitte Borell, role: https://id.loc.gov/vocabulary/relators/aut.html Phaedra Bouvet, role: https://id.loc.gov/vocabulary/relators/aut.html J Malakie LaClair, role: https://id.loc.gov/vocabulary/relators/aut.html Praon Silapanth, role: https://id.loc.gov/vocabulary/relators/aut.html James D. Allen, role: https://id.loc.gov/vocabulary/relators/aut.html Cholawit Thongcharoenchaikit, role: https://id.loc.gov/vocabulary/relators/aut.html Laure Dussubieux, role: https://id.loc.gov/vocabulary/relators/aut.html Bérénice Bellina, role: https://id.loc.gov/vocabulary/relators/aut.html	Boonyarit Chaisuwan, role: https://id.loc.gov/vocabulary/relators/aut.html Cholawit Thongcharoenchaikit, role: https://id.loc.gov/vocabulary/relators/aut.html Phaedra Bouvet, role: https://id.loc.gov/vocabulary/relators/aut.html Vincent Bernard, role: https://id.loc.gov/vocabulary/relators/aut.html Laure Dussubieux, role: https://id.loc.gov/vocabulary/relators/aut.html Bérénice Bellina, role: https://id.loc.gov/vocabulary/relators/aut.html Julie Laclair Malakie, role: https://id.loc.gov/vocabulary/relators/aut.html Sachipan Srikanlaya, role: https://id.loc.gov/vocabulary/relators/aut.html Praon Silapanth, role: https://id.loc.gov/vocabulary/relators/aut.html Brigitte Borell, role: https://id.loc.gov/vocabulary/relators/aut.html Cristina Castillo, role: https://id.loc.gov/vocabulary/relators/aut.html Sophie Peronnet, role: https://id.loc.gov/vocabulary/relators/aut.html Jane Allen, role: https://id.loc.gov/vocabulary/relators/aut.html
Origin	OpenAlex / https://openalex.org/W39138945	ScanR / halhal-01538901
Publication	01-01-2014	14-12-2014



"Sameness" is a social construct !

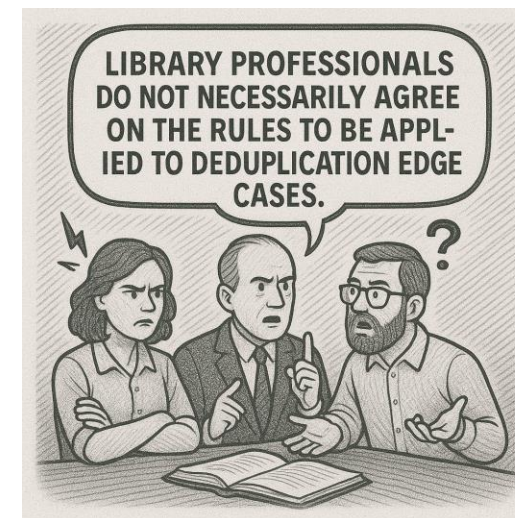
Complex relationships between scholarly publications :

- Preprint / peer-reviewed article
- Translations
- Successive editions
- Conference paper / supplementary material
- Data paper / methodology paper / dataset
- Errata, corrigenda, and retractions

Depends of the use case :

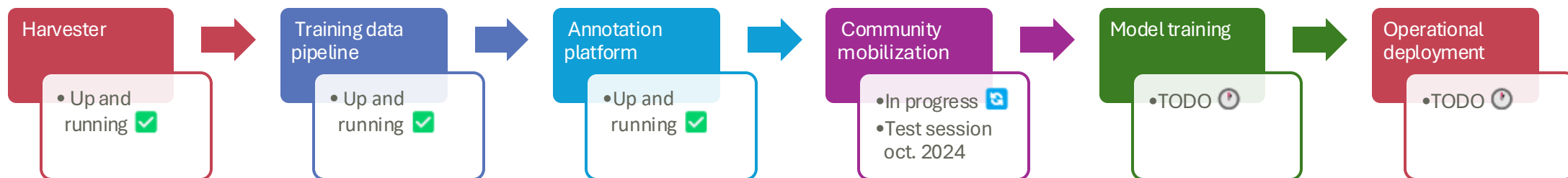
- A researcher managing their own publication list
- An institution that generates analytics
- A system automating deposit to another platform

The challenge is to make annotation practices converge towards a sufficient level of homogeneity to allow the model to generalize.





What stage are we at today?



- **End of 2024** : 70 people from 20 institutions took part in annotation tests
- Many issues in the harvester and training data pipeline have been fixed, e.g. :
 - Trivial duplicates by ISBN
 - Identifiers inferred from URI
 - ...
- **Spring 2024**: re-mobilizing the community !

Roadmap for 2025:

- Experiment with at least 2 approaches
 - BERT Siamese network
 - Llama fine-tuning
- Deploy to production (as knowledge graph helper)
- Share datasets
 - Kaggle
 - Hugging Face
- Why not challenges?